a cura di / edited by
Elena Marangoni
Camillo Carlo Pellizzari di San Girolamo

FIRENZE UNIVERSITY PRESS

Wikidata e la ricerca. Condividere esperienze / Wikidata and Research. Sharing Experiences

Atti del Convegno, Firenze, 5-6 giugno 2025 /
Conference Proceedings, Florence, 5-6 giugno 2025

WIKIDATA & Research
FLORENCE 2025

STUDI E SAGGI
ISSN 2704-6478 (PRINT) - ISSN 2704-5919 (ONLINE)
– 274 –

CONVEGNO “WIKIDATA E LA RICERCA”

Comitato scientifico

Franco Bagnoli, Università di Firenze
Monica Berti, Universität Leipzig
Carlo Bianchini, Università di Pavia
Alessandra Boccone, Università degli Studi di Salerno
Silvia Bruni, Università di Firenze
Antonella Buccianti, Università di Firenze
Dimitri D’Andrea, Università di Firenze
Annick Farina, Università di Firenze
Fulvio Guatelli, Università di Firenze
Elena Marangoni, Wikimedia Italia
Luca Martinelli, Wikimedia Foundation
Alessio Melandri, Wikimedia Italia
Daniel Mietchen, FIZ Karlsruhe
Rossana Morriello, Università di Firenze
Federico Morando, Centro Nexa su Internet e società del Politecnico di Torino
Camillo Carlo Pellizzari di San Girolamo, Scuola Normale Superiore, Pisa
Iolanda Pensa, SUPSI University of Applied Sciences and Arts of Southern Switzerland
Maria Ranieri, Università di Firenze
Miriam Redi, Wikimedia Foundation
Lucia Sardo, Università di Bologna

Wikidata e la ricerca.
Condividere esperienze

Atti del Convegno, Firenze, 5-6 giugno 2025

Wikidata and Research.
Sharing Experiences

Conference Proceedings, Florence, 5-6 giugno 2025

a cura di / edited by
Elena Marangoni
Camillo Carlo Pellizzari di San Girolamo

FIRENZE UNIVERSITY PRESS
2026

Wikidata e la ricerca. Condividere esperienze = Wikidata and Research. Sharing Experiences : atti del Convegno, Firenze, 5-6 giugno 2025 = Conference proceedings, Florence, 5-6 June 2025 / a cura di = edited by Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo. – Firenze : Firenze University Press, 2026.
(Studi e saggi ; 274)

https://books.fupress.com/isbn/9791221510102

ISSN 2704-6478 (print)
ISSN 2704-5919 (online)
ISBN 979-12-215-1009-6 (Print)
ISBN 979-12-215-1010-2 (PDF)
ISBN 979-12-215-1011-9 (XML)
DOI 10.36253/979-12-215-1010-2

Graphic design: Alberto Pizarro Fernández, Lettera Meccanica SRLs
Front cover image: Alexmar983 e Mascha 2013, CC0 via Wikimedia Commons

Con la collaborazione e il sostegno di Wikimedia Italia / With the collaboration and support of Wikimedia Italia.

UNIVERSITÀ
DEGLI STUDI
FIRENZE

Peer Review Policy
Peer-review is the cornerstone of the scientific evaluation of a book. All FUP's publications undergo a peer-review process by external experts under the responsibility of the Editorial Board and the Scientific Boards of each series (DOI 10.36253/fup_best_practice.3).

Referee List
In order to strengthen the network of researchers supporting FUP's evaluation process, and to recognise the valuable contribution of referees, a Referee List is published and constantly updated on FUP's website (DOI 10.36253/fup_referee_list).

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

The online digital edition is published in Open Access on www.fupress.com.

Published by Firenze University Press
Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

This book is printed on acid-free paper
Printed in Italy

Sommario / Table of Contents

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, © 2026 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

POSTER

Premessa

Debora Berti

La pubblicazione degli Atti di questo Convegno rappresenta un traguardo di rilievo per l'Università degli Studi di Firenze, non solo per il valore scientifico dei contributi raccolti, ma anche perché testimonia la vitalità di una visione strategica che il nostro Ateneo persegue con convinzione: l'apertura, la condivisione e la restituzione sociale della conoscenza.

Il volume si colloca programmaticamente nel solco dell'Open Access e dell'Open Science. La transizione verso una scienza aperta non è una mera declinazione tecnica o un adempimento formale, bensì un dovere etico e culturale. In quest'ottica, la scelta di dedicare una riflessione accademica strutturata a una piattaforma come Wikidata — qui analizzata per la prima volta in Italia attraverso la lente della ricerca scientifica — assume un valore paradigmatico. Wikidata si offre alla comunità scientifica come un formidabile *knowledge graph globale*, multilingue e basato su dati aperti; uno strumento capace di superare le barriere linguistiche e di far dialogare le macchine e gli umani, potenziando la visibilità, l'interoperabilità e l'usabilità dei dati prodotti dalla ricerca accademica.

Questo strumento ha una natura interdisciplinare. Come emerge con chiarezza dalla struttura stessa del volume — articolata nelle tre macro-sezioni metodologiche dedicate al *Modellare*, all'*Interrogare* e al *Visualizzare* i dati — Wikidata agisce come un tessuto connettivo trasversale. Esso accoglie e fa dialogare ambiti del sapere tradizionalmente distanti: dai cataloghi delle biblioteche agli studi classici, dalla storia dell'arte all'architettura, fino ad arrivare alle scienze naturali e alla *citizen science*. Consente, dunque, a ricercatori di settori diversi di confrontare ontologie, metodologie e linguaggi, offrendo risposte a domande di ricerca complesse che nessuna singola disciplina, isolatamente, potrebbe oggi formulare.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, © 2026 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

È tuttavia sul piano della Terza Missione che questa collaborazione rivela la sua componente più innovativa e feconda. L'Ateneo fiorentino non intende la ricerca come un processo autoreferenziale, ma come un bene comune che deve generare un impatto concreto sul territorio e sulla società civile. L'esperienza pionieristica della nostra *Wikistazione*, attiva presso la Biblioteca di Scienze Sociali, rappresenta questo modello organizzativo paritario e permanente, capace di far collaborare studenti, docenti, bibliotecari e cittadini in una comunità di pratica scientifica.

Questa sinergia tra professionisti della ricerca e comunità di pratica online dei progetti di Wikimedia Italia, non solo convalida e arricchisce la qualità dell'ecosistema della conoscenza libera, ma ridefinisce il ruolo stesso dell'università. L'Ateneo si fa mediatore culturale, uscendo dalle proprie mura per co-progettare il sapere insieme alla società civile.

Un ringraziamento sincero va a Wikimedia Italia per aver condiviso questa sfida, ai membri del Comitato scientifico, alle strutture bibliotecarie del nostro Ateneo e a tutti gli autori che hanno arricchito questo volume. Un plauso particolare, infine, va alla Firenze University Press, e al suo impegno istituzionale nell'editoria ad accesso aperto.

Prorettrice alla ricerca
dell'Università degli Studi di Firenze

PREFAZIONE

Un convegno su Wikidata con l'Università di Firenze

Ferdinando Traversa

L'organizzazione del convegno "Wikidata e la ricerca" con l'Università di Firenze ha rappresentato allo stesso tempo una conferma e una sfida. Una conferma perché Wikimedia Italia, associazione riconosciuta come capitolo italiano del movimento Wikimedia, ha tra i propri obiettivi l'organizzazione e il supporto a eventi che creino occasioni di rafforzamento e crescita della comunità dei contributori, favorendo la diffusione della conoscenza dei progetti Wikimedia. Sebbene le comunità dei volontari che contribuiscono a Wikipedia e ai progetti fratelli siano comunità online per eccellenza, i momenti di ritrovo in presenza rimangono fondamentali per il confronto, lo scambio di idee ed esperienze e la nascita di nuovi progetti e iniziative. Le collaborazioni con il mondo accademico, così come quelle con gli enti culturali, sono da sempre sostenute dall'associazione perché permettono di condividere con licenza libera contenuti e contributi di grande qualità[1]. Tuttavia, tali collaborazioni vedono più spesso al centro Wikipedia, quale strumento didattico o divulgativo, poiché è il progetto più conosciuto e di largo impatto.

[1] Tra le collaborazioni realizzate con il sostegno di Wikimedia Italia vi sono, ad esempio, quella pluriennale con l'Università di Padova, che ha portato, tra l'altro, al convegno "Wikipedia in Academia" (2019) <http://www.maldura.unipd.it/wikipedia-in-academia/#service> e alla realizzazione di tre corsi online liberamente fruibili: <https://learn.eduopen.org/course/index.php?mycourses=0&search=wiki>, e quella con il Politecnico di Milano che dal 2016 offre un corso per dottorandi sulla scrittura di voci su Wikipedia: <https://it.wikipedia.org/wiki/Progetto:Coordinamento/Università/PoliMi/Dottorandi/2025>.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, © 2026 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

Questa è stata invece la prima volta, almeno per l'Italia, in cui si è tenuto un convegno accademico su Wikidata. Tra i progetti della galassia Wikimedia, Wikidata è quello più vicino al mondo della ricerca per la sua stessa natura di repository di dati strutturati, aperto e interdisciplinare, in linea con i principi dell'Open science e dalla condivisione dei dati della ricerca secondo i principi FAIR[2]. Gli usi nell'ambito della ricerca sono molteplici ma molto specifici, e in questo senso la sfida consisteva nel coinvolgere e far dialogare interessi disciplinari, linguaggi (ontologie) e metodologie diverse.

Vi era alle spalle l'esperienza positiva dei Wikidata Days, svolti nel 2024 presso la biblioteca del CNR di Bologna[3]: due giorni di incontri della comunità italofona di Wikidata con un forte focus sulle biblioteche e contributi anche da parte di volontari di OpenStreetMap. Grazie all'entusiasmo dei volontari, dei bibliotecari attivi nel GWMAB (Gruppo Wikidata per Musei, Archivi e Biblioteche)[4] e alla presenza della Wikistazione presso la Biblioteca di Scienze sociali dell'ateneo fiorentino abbiamo quindi accolto la sfida di organizzare un altro appuntamento dedicato a Wikidata, dando continuità all'impegno verso tale comunità. Speriamo di aver tutti insieme offerto un'occasione di dialogo a livello internazionale e interdisciplinare a ricercatori, bibliotecari e volontari che già usano Wikidata e Wikibase nei propri progetti, ma anche un'opportunità di conoscere tale piattaforma per coloro che ancora non la utilizzano.

Un ringraziamento sincero va a tutti i docenti, i ricercatori, i bibliotecari, il personale e i volontari impegnati nell'organizzazione e nel comitato scientifico, nonché alla Firenze University Press per la pubblicazione del presente volume.

Presidente di Wikimedia Italia

[2] Si veda: <https://en.wikipedia.org/wiki/FAIR_data>.

[3] Si veda la pagina: <https://www.wikidata.org/wiki/Wikidata:Events/Wikidata_Days_Bologna_2024>.

[4] Su scopi e attività del Gruppo si veda: <https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche>.

Introduzione

Silvia Bruni, Camillo Carlo Pellizzari di San Girolamo, Elena Marangoni

1. I progetti Wikimedia all'Università di Firenze

Da molti anni le biblioteche, anche quelle universitarie, collaborano con le piattaforme Wikimedia. Questo legame nasce dalla consapevolezza che tali spazi collaborativi permettono sia di valorizzare le collezioni conservate, sia di stimolare il coinvolgimento attivo degli utenti (Stinson et al. 2018).

Queste esperienze sono piuttosto frammentate ed è difficile delineare modelli di organizzazione prevalenti: in certi casi sono singoli bibliotecari che lavorano sui progetti Wikimedia, in altri sono le biblioteche che si attivano, magari a partire da bandi promossi da Wikimedia Italia[1]. A volte si sviluppano attività su una piattaforma in particolare partendo dalle necessità di una biblioteca o dell'università stessa.

L'esperienza della Wikistazione della Biblioteca di Scienze sociali dell'Università di Firenze è probabilmente unica a livello italiano (Bruni 2025). L'idea è stata quella di creare all'interno dell'organigramma della biblioteca un ruolo specifico per portare avanti con continuità progetti Wiki inseriti negli obiettivi di lavoro della biblioteca. Una sorta di laboratorio, in cui la biblioteca diventa

[1] Come il bando piccoli musei e poi Bando MAB che dal 2020 sostiene i progetti degli enti culturali: <https://wiki.wikimedia.it/wiki/Bando_musei_archivi_biblioteche>.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, © 2026 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

un luogo di mediazione e diffusione dell'uso delle piattaforme di Wikimedia all'interno dell'università[2].

Questo modello presenta diversi vantaggi:

- le attività vengono mantenute nel tempo (continuità);
- vengono utilizzate diverse piattaforme (Wikibooks, Wikidata, Wikipedia, ecc.) e si possono gestire più progetti;
- c'è un allineamento con gli obiettivi della biblioteca e dell'università (coerenza con la missione);
- si creano collaborazioni con altre biblioteche e la comunità Wikimedia (rete);
- la collaborazione stimola nuove idee (creatività);
- si sperimenta un'organizzazione basata sulla collaborazione paritaria tra studenti, docenti e bibliotecari.

Soffermarsi a riflettere sul modello di collaborazione fra istituzioni, in questo caso le università, e Wikimedia è una questione importante, spesso trascurata, perché si dà maggiore visibilità al contenuto dei progetti piuttosto che ai modelli organizzativi grazie ai quali questi sono stati avviati. In questo senso una ricognizione in ambito GLAM potrebbe essere molto interessante per identificare esempi di strutturazione del lavoro nelle istituzioni con le piattaforme Wikimedia.

Anche per progettare il convegno su "Wikidata e la ricerca" era necessario immaginare un modello di lavoro. Il gruppo organizzatore di Wikimedia Italia ha contattato la Wikistazione e sono stati ipotizzati alcuni risultati che l'evento poteva portare per le due comunità, quella universitaria e quella wikidatiana.

Per la comunità accademica, il convegno avrebbe messo in luce come i dati prodotti dalla ricerca accademica possono essere integrati in un database globale e multilingue come Wikidata, aumentando la visibilità della ricerca e l'usabilità dei dati prodotti. Inoltre, avrebbe offerto a ricercatori, bibliotecari e studenti l'opportunità di acquisire competenze pratiche su come usare Wikidata come strumento per strutturare e collegare i dati (Boccone 2022; Evenstein Sigalov et al. 2024). L'evento avrebbe infine promosso la riflessione su come rendere gli archivi istituzionali più interoperabili (Bargioni et al. 2023).

Per la comunità di Wikidata, il convegno avrebbe rappresentato un'occasione per costruire e rafforzare i legami tra gli utenti volontari di Wikidata e i professionisti del mondo accademico. Questo scambio di conoscenze avrebbe permesso di validare e migliorare la qualità delle informazioni su Wikidata, arricchendo l'intero ecosistema della conoscenza libera (Shenoy et al. 2022; Waagmeester et al. 2020; Zhao 2022).

Le aspettative erano alte e riguardavano l'affinamento dell'integrazione fra mondo accademico e Wikidata.

[2] Nel 2024 è nato un gruppo di lavoro su Wikidata di bibliotecari dell'Università di Firenze e della Biblioteca nazionale centrale di Firenze. Una delle sperimentazioni realizzate è descritta in questo volume.

Tra le criticità che l'organizzazione del convegno ha presentato vi è stata quella di proporre un tema piuttosto nuovo (Casalini e D'Addario 2019). Infatti, a differenza di Wikipedia, ormai nota a tutti, Wikidata è ancora un oggetto misterioso. Il suo funzionamento, legato ai dati e ai metadati, richiede conoscenze tecniche e interessi specifici.

Nonostante questo, siamo riusciti a creare un comitato scientifico formato da wikidatiani esperti, professionisti e accademici e i contributi presentati sono stati molti, da parte di autori provenienti da università e istituti di ricerca di diversi paesi. Avremmo voluto coinvolgere maggiormente la comunità studentesca per avvicinare nuovi potenziali contributori e dare più spazio agli aspetti divulgativi, ma la 'porta principale' dell'università, pur con qualche scricchiolio, si è aperta alla comunità di Wikidata. La stessa pubblicazione degli atti da parte della Firenze University Press è un segno di interesse significativo che fa ben sperare nella prosecuzione di laboratori permanenti come la Wikistazione, che possono svolgere un ruolo cruciale nel mantenere viva una progettualità di lungo periodo per costruire comunità di pratica e reti di lavoro[3].

2. Cos'è Wikidata?

Wikidata[4] è un *knowledge graph* contenente dati strutturati secondo lo standard RDF e pubblicati con licenza CC0, nato nel 2012; usa come software MediaWiki con la fondamentale aggiunta dell'estensione Wikibase, che rende possibile immagazzinare dati strutturati. Strutturare i dati significa codificare le informazioni tramite relazioni tra entità, che prescindono dalla lingua e sono leggibili anche dalle macchine. Alcuni dei vantaggi dei dati strutturati sono:

- risparmiare tempo e raggiungere potenzialmente tutti gli utenti (anche le macchine) superando le barriere linguistiche;
- fare traduzioni automatiche;
- interrogare i dati in modo preciso e in profondità, grazie alla granularità dell'informazione e al legame tra dati e metadati;
- rispondere a domande complesse;
- creare visualizzazioni nuove (mappe, timeline, grafici ecc.);
- confrontare più database tra loro.

L'esigenza iniziale è stata quella di centralizzare i legami tra le voci nelle varie edizioni linguistiche di Wikipedia e i dati usati negli *infobox* di queste ultime; le diverse Wikipedie, infatti, sono scritte dai volontari in modo indipendente. Wikidata ha permesso quindi di strutturare e aggiornare centralmente, rendendole disponibili per

[3] Un primo risultato della rete di collaborazioni creata attorno al Convegno è l'uscita di un numero monografico del *Journal of Open Humanities Data (JOHD)* dal titolo *Wikidata Across the Humanities: Datasets, Methodologies, Reuse*, che vede coinvolti nel comitato scientifico alcuni degli organizzatori e autori del presente volume (cfr. <https://openhumanitiesdata.metajnl.com/collections/wikidata_across_the_humanities>).

[4] <https://www.wikidata.org>.

tutte le versioni linguistiche di Wikipedia, una serie di informazioni, corredate da riferimenti alle fonti: anni di nascita o morte di persone, coordinate geografiche di luoghi, numero di abitanti di paesi o città nelle diverse rilevazioni statistiche, e qualunque altro dato possa essere registrato in forma di relazione soggetto-predicato-oggetto.

Confrontando numericamente i contenuti, la Wikipedia in lingua inglese (la più grande) contiene oltre 7 milioni di voci, le pagine presenti in almeno una Wikipedia sono oltre 28 milioni[5], e il totale delle voci di Wikipedia è di oltre 65 milioni[6]. Wikidata ha elementi riguardo agli argomenti di tutte queste voci, ma anche riguardo a decine di milioni di altri temi: i suoi elementi, infatti, sono oggi oltre 117 milioni[7], contenenti oltre 1,6 miliardi di dichiarazioni[8], ovvero relazioni soggetto-predicato-oggetto, il cui soggetto sono i suddetti elementi e il cui predicato sono le proprietà, oltre 12.000[9].

Un'attenzione speciale meritano, tra queste proprietà, gli oltre 9.000 identificativi esterni[10]: ogni identificativo esterno presente in un elemento indica una relazione di identità tra esso e un'entità in un altro sito web. Ciò significa stabilire un collegamento tra Wikidata e altri siti web ma anche, grazie a Wikidata come nodo comune, collegare gli altri siti web tra di loro. Ad esempio, l'elemento corrispondente allo scrittore Luigi Pirandello (Q1403[11]) possiede oltre 230 relazioni con identificativi esterni (elencati nella sezione "Identificativi")[12]: ciascuna di esse è un link che, a partire da Wikidata, conduce ad altrettanti siti web che contengono informazioni sullo scrittore, come cataloghi di biblioteche che possiedono i suoi libri o libri su di lui, archivi che conservano fonti primarie su tale autore, biblioteche digitali che offrono la possibilità di consultare le sue opere digitalizzate in pubblico dominio, enciclopedie online che ne contengono la biografia, banche dati specifiche di critica letteraria, filologia o didattica, banche dati cinematografiche che descrivono i film tratti dalle sue opere, o raccolte di audiovisivi con registrazioni di messe in scena delle sue opere teatrali, solo per fare alcuni esempi di una rete di conoscenze collegate, a disposizione di ognuno e interrogabili in modo puntuale. In un unico luogo sono raccolti i link a un amplissimo e variegato insieme di fonti su un argomento: per questo Wikidata può essere sinteticamente descritta come «hub degli hub» della Linked Open Data Cloud[13].

Cosa può essere aggiunto in Wikidata? Questa soglia è espressa con il concetto di «notabilità» (*notability*)[14] e attualmente prevede che in Wikidata possa

5 <https://qlever.dev/wikidata/dxVruA>.
6 <https://meta.wikimedia.org/wiki/List_of_Wikipedias>.
7 <https://qlever.dev/wikidata/TWEPBF>.
8 <https://qlever.dev/wikidata/OMvoUO>.
9 <https://qlever.dev/wikidata/AdPf5q>.
10 <https://qlever.dev/wikidata/31HQ0i>.
11 <https://www.wikidata.org/wiki/Q1403>.
12 <https://qlever.dev/wikidata/D8FLpa>.
13 <https://lod-cloud.net/>.
14 <https://www.wikidata.org/wiki/Wikidata:Notability>.

essere inserita «qualunque entità concettuale o materiale chiaramente identificabile che possa essere descritta utilizzando fonti serie e disponibili al pubblico». Un criterio molto più ampio dei criteri di enciclopedicità previsti dalle varie edizioni linguistiche di Wikipedia, poiché Wikidata deve contenere potenzialmente ogni elemento e relazione utile a descrivere la realtà e la conoscenza umana; è tuttavia necessario l'aggancio al pilastro fondamentale dei progetti Wikimedia, ovvero quello della presenza di fonti attendibili e verificabili, che salvaguardano la qualità e il valore dei dati raccolti.

3. Wikibase

Un'altra scelta possibile per strutturare un insieme di dati, oltre all'inserimento in Wikidata, è quella di creare un'istanza indipendente di Wikibase, usando Wikibase Cloud[15] o Wikibase Suite[16].

Wikibase Cloud è un servizio ospitato sui server di Wikimedia Deutschland e quindi non comporta per l'utente alcun costo di manutenzione, né richiede che sia in possesso di particolari competenze tecniche; tuttavia, le possibilità di personalizzare l'interfaccia sono limitate. Usare Wikibase Suite significa invece creare un proprio sito web usando Wikibase, quindi con piena facoltà di personalizzarlo, ma facendosi carico sia dell'installazione, della configurazione e del costante aggiornamento del software, sia dei costi di manutenzione.

Un'istanza indipendente di Wikibase dà piena libertà nello scegliere:

- chi può modificare (gli utenti);
- come strutturare i dati (le proprietà);
- quali contenuti modificare (gli elementi);
- il grado di riutilizzabilità dei dati (la licenza).

Su Wikidata invece:

- chiunque può modificare;
- la struttura dei dati è approvata comunitariamente;
- gli elementi devono avere una rilevanza generale (ossia devono rispettare la già citata soglia di «notabilità»);
- i dati strutturati sono pubblicati in licenza CC0.

Wikidata è un punto di snodo fondamentale del web semantico, quindi garantisce ai dati inseriti un'amplissima visibilità e un amplissimo riuso, in molteplici ambiti e nel tempo. Inoltre, il contributo della comunità può portare a migliorare i dati. Il modello dei dati per i diversi settori di conoscenza è frutto di lunghe discussioni tra persone provenienti da ambiti molto diversi e quindi è spesso un buon compromesso tra esigenze diverse, dovendo Wikidata raccogliere dati su qualunque cosa e non esclusivamente su ambiti di

[15] <https://www.wikibase.cloud/>.

[16] <https://wikiba.se/>.

studio molto specifici; inoltre, se non è soddisfacente si possono proporre modifiche, in particolare aggiungendo nuove proprietà. Il confronto con i dati già presenti, infine, può essere utile per consolidare i propri o per vederli da una prospettiva diversa. Wikidata permette insomma un'interessante combinazione tra confronto interdisciplinare e logica collaborativa da un lato e specificità settoriale dall'altro.

D'altra parte, Wikidata ha dei limiti, sia da un punto di vista tecnico sia dal punto di vista della comunità di volontari che ne costruiscono i contenuti[17]: la scalabilità del progetto è un problema molto sentito negli ultimi anni e sono in corso importanti discussioni in merito[18]. Stabilire quali contenuti debbano essere inseriti in Wikidata e quali invece possano essere inseriti in altre istanze di Wikibase, connesse con Wikidata e quindi interrogabili congiuntamente tramite query federate, è ancora una questione aperta[19], e speriamo che anche questo volume possa fornire un contributo utile al dibattito in corso.

4. Contenuti

I contributi presentati al convegno sono stati raggruppati secondo una logica di metodo nelle tre parti *1) Modellare*, *2) Interrogare* e *3) Visualizzare i dati* e, secondariamente, per quanto riguarda la prima e più corposa parte, secondo l'area disciplinare sebbene, per la prerogativa stessa dello strumento, l'attribuzione a uno specifico settore sia talvolta non scontata e le metodologie, gli approcci e i quesiti di ricerca siano quanto mai trasversali. Vi si troveranno inoltre, per specifica scelta in fase di definizione dei criteri di ammissibilità dei contributi, sia progetti compiuti o avviati sia progetti in fase preparatoria, che possono comunque costituire spunti per nuove progettualità. Del convegno ha fatto parte anche una partecipata sezione dedicata ai poster, accompagnati in questa sede da un testo descrittivo per facilitarne la lettura.

Tutti i contributi proposti sono stati selezionati e poi revisionati tramite un processo di open peer review, con trasparenza reciproca di autori, revisori e revisioni, da almeno due revisori tra i membri del comitato scientifico.

Si ringraziano tutti gli autori che hanno voluto partecipare all'evento e condividere le loro esperienze e i membri del comitato scientifico che, su base volontaria e trasparente, hanno contribuito coi loro consigli al miglioramento dei saggi qui presentati. Un ringraziamento particolare va a Iolanda Pensa per il prezioso ruolo di coordinamento del comitato scientifico e a Marta Arosio per l'indispensabile supporto nell'organizzazione del convegno.

[17] <https://www.wikidata.org/wiki/Wikidata:WikiProject_Limits_of_Wikidata>.

[18] <https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Mass-editing_policy> e soprattutto <https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Notability_policy_reform>.

[19] Si veda ad esempio <https://commons.wikimedia.org/wiki/File:Community_report_Data_Governance_Research_(Q2_2025).pdf>.

Riferimenti bibliografici

Bargioni, Stefano, Carlo Bianchini, e Camillo Carlo Pellizzari di San Girolamo. 2023. "IRIS e Wikidata: un progetto per una migliore valorizzazione e fruizione degli archivi della ricerca scientifica italiana." In *Guardando oltre i confini : partire dalla tradizione per costruire il futuro delle biblioteche : studi e testimonianze per i 70 anni di Mauro Guerrini*, 31-45. Roma: Associazione italiana biblioteche.

Boccone, Alessandra. 2022. "Role of the Wikidata librarian in a renewed Bibliographical Universe: "next generation metadata", next generation librarians." *JLIS.it* 13, 2: 45-57. https://doi.org/10.36253/jlis.it-460

Casalini, Brunella, e Silvia D'Addario, a cura di. 2019. *Il tempo per pensare: Un bene essenziale per la comunità universitaria*, vol. I. Comitato Unico di Garanzia. Firenze: Firenze University Press. https://doi.org/10.36253/978-88-6453-845-7

Evenstein Sigalov, Shani Anat Cohen, and Rafi Nachmias. 2025. "Transforming Higher Education: A Decade of Integrating Wikipedia and Wikidata for Literacy Enhancement and Social Impact." *Journal of Computers in Education* 12, 3: 953-95. https://doi.org/10.1007/s40692-024-00334-x.

Shenoy, Kartik, Filip Ilievski, Daniel Garijo, Daniel Schwabe, and Pedro Szekely. 2022. "A Study of the Quality of Wikidata." *Journal of Web Semantics* 72: 100679. https://doi.org/10.1016/j.websem.2021.100679

Stinson, Alexander D., Liam Wyatt, and Sandra Fauconnier. 2018. "Stepping Beyond Libraries : the Changing Orientation in Global GLAM-Wiki." *JLIS* 3. https://doi.org/10.4403/jlis.it-12480

Waagmeester, Andra, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M Good, Malachi Griffith, Obi L Griffith, Kristina Hanspers et al. 2020. "Wikidata as a Knowledge Graph for the Life Sciences." *eLife* 9: e52614. https://doi.org/10.7554/eLife.52614

Zhao, Fudie. 2023. "A Systematic Review of Wikidata in Digital Humanities Projects." *Digital Scholarship in the Humanities* 38, 2: 852-74. https://doi.org/10.1093/llc/fqac083

PARTE PRIMA

Modellare i dati / Data modeling

SEZIONE 1

Biblioteche e cataloghi / Libraries and catalogues

Il progetto OpenAcolit, un repertorio delle biblioteche italiane realizzato con Wikibase

Stefano Bargioni

Abstract: OpenAcolit, the successor to Acolit, is an innovative project developed by ABEI, PUSC, BCE-CEI, and DMBC-UniPV. It does not aim to transcribe the paper volumes of the Acolit repertoire but rather to bring its data into the semantic web through Wikibase Cloud. This digital initiative is built on collaboration between librarians and scholars, offering new ways of access and interaction with other software.

Keywords: OpenAcolit, Acolit, Wikibase, authority list, linked open data.

1. Introduzione

OpenAcolit[1] succede ad Acolit, *Autori cattolici ed opere liturgiche*, opera in 4 volumi pubblicata dal 1998 al 2010 da ABEI. OpenAcolit è un progetto di ABEI, PUSC, BCE-CEI, DMBC-UniPV[2].

La natura di Acolit è quella del repertorio, della risorsa di riferimento per la corretta dicitura di nomi di autori, enti, opere legati ad uno dei più ampi ambiti della nostra cultura, quello ecclesiologico, dove storia, diritto, teologia e filosofia si richiamano costantemente[3].

OpenAcolit intende aggiornare Acolit sia in quanto al contenuto, sia in quanto alle funzionalità. Forse gli anni trascorsi dall'inizio della sua stesura non danno l'impressione di grandi cambiamenti nelle corrette forme di citazione delle entità suddette. Tuttavia la sola ipotesi di procedere ad una nuova edizione car-

1 <https://openacolit.wikibase.cloud>.

2 ABEI – Associazione dei Bibliotecari Ecclesiastici Italiani; PUSC – Pontificia Università della Santa Croce; BCE-CEI – Ufficio Nazionale per i beni culturali ecclesiastici e l'edilizia di culto della Conferenza Episcopale Italiana; DMBC-UniPV – Dipartimento di Musicologia e Beni Culturali dell'Università degli Studi di Pavia.

3 Per il posizionamento di Acolit nel lavoro di catalogazione delle biblioteche italiane e il suo rapporto con REICAT e Nuovo Soggettario, si veda Guerrini 2009.

Stefano Bargioni, Pontificia Università della Santa Croce, Italy, bargioni@pusc.it, 0000-0001-7679-2874

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Stefano Bargioni, *Il progetto OpenAcolit, un repertorio delle biblioteche italiane realizzato con Wikibase,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.07, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 25-32, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

tacea per introdurvi gli aggiornamenti ha costituito per diversi anni un blocco psicologico difficilmente sormontabile per ABEI. Anche l'ipotesi di concentrare questo compito su poche persone, pur passando a una base informatica, non permetteva di prendere una decisione.

La soluzione proposta, più avanti illustrata in maggior dettaglio, intende basarsi sul contributo di un gruppo aperto di bibliotecari direttamente interessati all'utilizzo dei dati, di esperti e studiosi, così come intende contribuire al Controllo Bibliografico Universale (UBC)[4]. Wikibase[5]. e più precisamente Wikibase Cloud, si è dimostrato in grado di adempiere ai due compiti fondamentali suddetti[6]. Una volta creata l'istanza OpenAcolit in wikibase.cloud, in data 29 giugno 2023, su di essa sono stati intrapresi lo studio di fattibilità, i primi test reali e le prime pratiche di scrittura concorrenziale dei dati.

Il progetto è stato presentato ed utilizzato in tre diverse giornate di incontri di bibliotecari e bibliotecarie, tenutesi in autunno 2024 a Roma, Andria e Milano, grazie anche alla fattiva cooperazione di Wikimedia Italia[7].

2. Significati di «open» in OpenAcolit

L'aggiunta della parola «open» al nome di Acolit è stata un passo quasi naturale e nel tempo si è dimostrata essere una chiave di lettura trasversale del progetto:

- chiunque lo può consultare e ne può utilizzare i dati, compresi gli URI
- chiunque, con la sufficiente formazione professionale[8]. può concorrere al mantenimento
- non ha carattere commerciale, implicito nella pubblicazione di un'opera cartacea o di un database ad accesso controllato con credenziali
- va oltre il contenuto di Acolit, permettendo aggiunte e correzioni
- va potenzialmente oltre le tipologie di entità trattate nei 4 volumi, per esempio potrebbe includere i santi, risolvendo la problematica della duplicazione causata da papi santi
- oltre alla forma di Acolit, riporta o può riportare le forme costruite in base a normative catalografiche di ambiti diversi, vigenti, in disuso o di prossima introduzione

[4] Per il progressivo sviluppo dell'UBC si può vedere Sardo 2021.

[5] Wikibase è un software libero e open source realizzato e supportato da Wikimedia Deutschland e da una vasta comunità di collaboratori. Wikibase è disponibile in cloud (<https://www.wikibase.cloud>) o installabile in locale (<https://wikiba.se/>). In entrambi i casi è gratuito. Anche Wikidata è realizzato con Wikibase.

[6] Il rapporto tra UBC e Wikidata, e per analogia anche tra UBC e sistemi basati su Wikibase, è esaminato in Bianchini e Sardo 2022.

[7] Alcune statistiche, continuamente aggiornate, si trovano alla pagina <https://openacolit.wikibase.cloud/wiki/Special:Statistics>.

[8] Per diventare utenti e poter editare OpenAcolit, occorre richiedere l'accesso. Non sono ammesse la creazione automatica di un account né le modifiche anonime, come invece accade in Wikidata o in Wikipedia.

- può andare oltre l'ambito cattolico e includere altre confessioni
- può andare oltre l'ambito della lingua italiana, anche nella lingua dell'interfaccia
- è potenziale strumento di studio comparativo ed evolutivo di normative
- è potenziale strumento di migrazione delle forme usate in un catalogo da una normativa all'altra
- è attore sul palco dei LOD: genera URI per ogni entità, è unito a Wikidata tramite apposita proprietà ed è consultabile con API e query SPARQL, anche federate[9].

3. Stato del progetto

Acolit raccoglie le forme di nomi relativi a:

- Bibbia, Chiesa cattolica, Curia romana, Stato pontificio, Vaticano, papi e antipapi;
- ordini religiosi;
- opere liturgiche;
- Padri della Chiesa e scrittori ecclesiastici occidentali fino al secolo XIII.

La duttilità di un database di triple RDF come quello su cui si basa OpenAcolit, consente di definire proprietà a sostegno di tipi diversi di entità. Quella fondamentale è «istanza di», che permette di distinguere i vari tipi o classi. Anche la proprietà che unisce un'istanza di OpenAcolit al corrispondente elemento di Wikidata è comune a tutte le classi. Ogni classe ha poi proprietà specifiche. La query seguente mostra le istanze attualmente definite in OpenAcolit[10].

```
PREFIX oawdt: <https://openacolit.wikibase.cloud/prop/direct/>
PREFIX wikibase: <http://wikiba.se/ontology#>
SELECT ?istanzaLabel (COUNT(?q) AS ?count) WHERE {
 ?q oawdt:P1 ?istanza.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_
LANGUAGE],en". }
}
GROUP BY ?istanza ?istanzaLabel
```

I conteggi risultanti dalla query sono: 2 per res, 118 per opera, 931 per espressione, 165 per lingua, 6 per tipo di fonte. Come si nota, si tratta di istanze relative a opere, e in particolare a opere della Bibbia e a loro espressioni in lingue diverse. E come forse è stato già intuito, lo scopo biblioteconomico e quindi non letterario, né storico o altro, ha portato a limitare la descrizione delle opere e delle espressioni raccogliendo per ognuna la forma prevista da una o più normative catalografiche o fonti.

[9] Le query federate sono un tipo di interrogazione di database RDF in cui possono essere richiesti dati contemporaneamente a più di un server SPARQL. SPARQL è acronimo ricorsivo di «SPARQL Protocol and RDF Query Language» ed è un linguaggio di interrogazione standard per il recupero e la manipolazione dei dati memorizzati in formato RDF.

[10] I risultati sono aggiornati al 2 aprile 2025.

Al momento di scrivere, le 6 fonti sono: ACOLIT, REICAT, AACR2, BAV, URBE, Volpi[11]. Dal lato pratico, quindi, un'agenzia di catalogazione effettuerà una ricerca tra le forme della normativa di proprio riferimento. Probabilmente OpenAcolit verrà adoperata principalmente per le REICAT, ma potrà fare lo stesso esatto servizio per altre normative che nel tempo verranno introdotte in OpenAcolit.

Sarà stato notato l'utilizzo del modello IFLA-LRM[12]. Ogni espressione, come può essere un libro dell'Antico o del Nuovo Testamento in una determinata lingua, realizza l'opera corrispondente, e il legame tra le due è rappresentato da una apposita proprietà[13]. nel verso di «molte a una».

L'esame di un elemento di OpenAcolit, l'opera «Bibbia. A.T. Tobia»[14] consente di comprendere meglio diversi aspetti e potenzialità di OpenAcolit.

istanza di	opera
lingua	ebraico
wikidata	Q131737[15]
è parte di	Bibbia
	Bibbia. A.T.
	Bibbia. A.T. Scritti
	Bibbia. A.T. Libri storici
	Bibbia. A.T. Deuterocanonici
forma autorevole	Bibbia. A.T. Tobia
	fonte ACOLIT
	Bibbia. Antico Testamento. Tobia
fonte	REICAT
	Biblia. V.T. Tobias
fonte	BAV
	Biblia. V.T. Thobias
	fonte Varianti Locali URBE

Ogni occorrenza della proprietà «forma autorevole» (P27) è distinta dalle altre dal qualificatore «fonte». Questo raggruppamento di forme, ognuna autorevole nell'ambito di applicazione identificato dalla fonte stessa, è al momen-

[11] La fonte Volpi si riferisce al repertorio curato e pubblicato da Vittorio Volpi nel 1994, dal titolo *DOC: dizionario delle opere classiche: intestazioni uniformi degli autori, elenco delle opere e delle parti componenti, indici degli autori, dei titoli e delle parole chiave della letteratura classica, medievale e bizantina*. Questa fonte è stata introdotta in OpenAcolit non per riportare i dati del repertorio stesso, ma per poterlo usare come citazione in occasioni particolari. Le altre fonti si riferiscono rispettivamente all'opera Acolit stessa, alle Regole Italiane di Catalogazione, alle Anglo-American Cataloguing Rules, alle Norme vaticane e alle "Varianti locali" di URBE, derivate dalle Norme vaticane e presentate in Leoni 2024.

[12] <https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712-it.pdf>.

[13] <https://openacolit.wikibase.cloud/wiki/Property:P7>.

[14] <https://openacolit.wikibase.cloud/entity/Q67>.

[15] A sua volta, e a suo tempo, Wikidata potrà contenere Q67, e in generale gli identificatori di OpenAcolit, in un'apposita proprietà.

to – per quanto è dato sapere – un unicum nel mondo della biblioteconomia. Come accennato in precedenza, P27 è la base per il confronto tra fonti, anche tra fonti ormai non più vigenti ma eventualmente ancora presenti in cataloghi non convertiti. Va notata anche la possibilità di sviluppare una propria nuova normativa, o di trascrivere in supporto informatico una normativa vigente ma regolamentata ancora su supporto cartaceo, permettendo così che interagisca con gli attuali sistemi di catalogazione e metadatazione[16].

Per gli altri tipi di entità trattati da Acolit, OpenAcolit si svilupperà in modo analogo come fatto con opere ed espressioni della Bibbia: definizione delle proprietà specifiche, conversione dei dati dai file di Acolit, loro importazione automatica, controllo manuale della completezza e della qualità dell'importazione, completamento ed aggiornamento, anche a consultazione in corso[17].

4. Esempi di utilizzo

4.1 In catalogazione

L'utilizzo più frequente che si può ipotizzare per OpenAcolit è costituito dal recupero della corretta forma da impiegare in catalogazione. Nel caso di opere ed espressioni della Bibbia e in ambito MARC21, i tag interessati sono 130, 730, 630, che in ambito Unimarc (SBN) corrispondono ai tag 740, 742, 606.

È possibile ricercare la forma in OpenAcolit, copiarla e riportarla nel proprio catalogo; ma è anche possibile, e molto più produttivo, associare una funzionalità di autocompletamento (o autosuggest) basata sull'interrogazione automatica dell'endpoint SPARQL di OpenAcolit[18].

La variabile «query» impostata con il testo del codice SPARQL seguente:

```
#title: Autosuggest per "apocalisse" (REICAT)
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX oap: <https://openacolit.wikibase.cloud/prop/>
PREFIX oaps: <https://openacolit.wikibase.cloud/prop/statement/>
PREFIX oapq: <https://openacolit.wikibase.cloud/prop/qualifier/>
PREFIX oawd: <https://openacolit.wikibase.cloud/entity/>
PREFIX oawdt: <https://openacolit.wikibase.cloud/prop/direct/>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX bd: <http://www.bigdata.com/rdf#>
SELECT ?q ?forma ?fonteLabel WHERE {
 ?q rdfs:label ?l;
      oawdt:P1 ?istanza;
```

[16] Potrebbe essere il caso delle "Varianti locali" di URBE o della Biblioteca apostolica vaticana, per la quale sono vigenti in cartaceo le "Norme per il catalogo degli stampati" terza ed., 1949.

[17] Il Gruppo Promotore di OpenAcolit non ha ancora definito successione e tempistiche per il trattamento dei dati rimanenti di Acolit.

[18] L'endpoint SPARQL di OpenAcolit si trova all'indirizzo <https://openacolit.wikibase.cloud/query>.

```
        oawdt:P17 ?forma;
        oap:P17 _:b26.
  _:b26 oaps:P17 ?forma;
        oapq:P16 ?fonte.
 values ?fonte {oawd:Q1236} # Q1236 REICAT ; Q1243 URBE ; Q1242 AACR2
  FILTER((LANG(?l)) = "it")
  FILTER(CONTAINS(LCASE(?l), "apocalisse"))
  SERVICE wikibase:label { bd:serviceParam wikibase:language "it". }
}
ORDER BY asc(?forma)
```

può essere adoperata con successo all'interno di un ambiente di catalogazione che interroghi OpenAcolit, per esempio con il seguente codice JavaScript/jQuery:

```
let Q = encodeURIComponent(query);
jQuery.getJSON(`https://openacolit.wikibase.cloud/query/
sparql?query=${Q}`, function(R){
        // utilizzo della risposta R
})
```

In questo caso, insieme alla forma si ha anche il vantaggio della registrazione automatica dell'URI di OpenAcolit, da conservare in un apposito sottocampo del tag da popolare[19].

È decisivo riportare l'URI dell'entità nei propri dati allo scopo di legare questi ultimi ai Linked Open Data a cui OpenAcolit appartiene per costituzione. Mentre nel caso di forma mancante, la registrazione di questa in OpenAcolit prima di procedere al suo utilizzo, permette non solo di fissare la stringa autorevole (almeno quella della normativa di proprio interesse) e possibili varianti, ma anche di assegnarle un URI.

4.2 Per studio e ricerca

Un ulteriore caso è rappresentato dallo studio dei dati di OpenAcolit. In questo esempio si confrontano le forme di alcune espressioni dell'opera Apocalisse presenti in OpenAcolit.

```
#title: Confronto tra le forme REICAT e URBE per opere ed espressioni
dell'Apocalisse
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX oap: <https://openacolit.wikibase.cloud/prop/>
```

[19] Va anche notato che la risposta è di tipo CORS: tutte le basi Wikibase includono nei dati della risposta la dichiarazione HTTP "access-control-allow-origin: *" che permette a un browser web di accettare i dati ricevuti da terze parti come riutilizzabili, mentre ciò ordinariamente non avviene. Per una trattazione esaustiva della problematica, si veda <https://developer.mozilla.org/en-US/docs/Web/HTTP/Guides/CORS/Errors>.

```
PREFIX oaps: <https://openacolit.wikibase.cloud/prop/statement/>
PREFIX oapq: <https://openacolit.wikibase.cloud/prop/qualifier/>
PREFIX oawd: <https://openacolit.wikibase.cloud/entity/>
PREFIX oawdt: <https://openacolit.wikibase.cloud/prop/direct/>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX bd: <http://www.bigdata.com/rdf#>
SELECT ?forma ?fonteLabel WHERE {
 ?q rdfs:label ?l;
       oawdt:P1 ?istanza;
       oawdt:P17 ?forma;
       oap:P17 _:b26.
 _:b26 oaps:P17 ?forma;
       oapq:P16 ?fonte.
 VALUES ?fonte {
       oawd:Q1243
       oawd:Q1236
 }
 FILTER((LANG(?l)) = "it")
 FILTER(CONTAINS(LCASE(?l), "apocalisse"))
 SERVICE wikibase:label {
bd:serviceParam wikibase:language "it". }
 }
ORDER BY (?q) (?fonteLabel)
```

La risposta è:

Bibbia. Nuovo Testamento. Apocalisse	REICAT
Biblia. N.T. Apocalypsis	URBE
Bibbia. Nuovo Testamento. Apocalisse (in francese)	REICAT
Biblia. N.T. Apocalypsis. Francese	URBE
Bibbia. Nuovo Testamento. Apocalisse (in inglese)	REICAT
Biblia. N.T. Apocalypsis. Inglese	URBE
Bibbia. Nuovo Testamento. Apocalisse (in italiano)	REICAT
Biblia. N.T. Apocalypsis. Italiano	URBE
Bibbia. Nuovo Testamento. Apocalisse (in latino)	REICAT
Biblia. N.T. Apocalypsis. Latino	URBE
Bibbia. Nuovo Testamento. Apocalisse (in spagnolo)	REICAT
Biblia. N.T. Apocalypsis. Spagnolo	URBE
Bibbia. Nuovo Testamento. Apocalisse (in tedesco)	REICAT
Biblia. N.T. Apocalypsis. Tedesco	URBE

5. Conclusioni

OpenAcolit possiede diverse caratteristiche della tradizione repertoriale cartacea. Non trattandosi però di una trascrizione di Acolit, è più appropriato dire che si tratta di una trasposizione dei suoi dati. OpenAcolit rende tali dati

fruibili in modo completamente diverso, li integra nel web semantico e ne permette il riuso con altri software. Ancor più, la compartecipazione nella sempre aperta redazione di numerosi attori sufficientemente esperti in ogni settore può assicurare la continuità delle authority list di Acolit.

OpenAcolit quindi rappresenta una nuova frontiera del concetto di repertorio. Senza esserselo proposto direttamente, risponde alle domande che pose Barbara Tillett in una intervista a Mauro Guerrini riportata nella presentazione del Volume 2 di Acolit, in particolare alle domande su forme in altre lingue, sulla possibilità e modalità di aggiornamento dell'opera, sullo sfruttamento informatico dei dati, e soprattutto sulla possibilità di riportare le diverse forme distinte per fonte (RICA, AACR2, RAK…)[20]. Significativamente, e quasi profeticamente, Guerrini conclude l'ultima risposta con le seguenti parole: «Speriamo nella collaborazione dei bibliotecari e dei lettori che lavorano in questo settore». Una speranza che OpenAcolit riaccende e, almeno potenzialmente, può soddisfare.

Riferimenti bibliografici

Associazione bibliotecari ecclesiastici italiani. 2000. *Acolit. Autori cattolici e opere liturgiche*, vol. 2: *Ordini religiosi*. Milano: Editrice Bibliografica.

Bianchini, Carlo, and Lucia Sardo. 2022. "Wikidata: A New Perspective towards Universal Bibliographic Control." *JLIS.It* 13, 1: 291-311. https://doi.org/10.4403/jlis.it-12725

Guerrini, Mauro. 2009. "Il catalogo "ecclesiastico" oggi tra Acolit, nuove RICA e nuovo soggettario." *Bollettino ABEI* 2. https://hdl.handle.net/2158/1132972.

Leoni, Cristiana. 2024. "Quindici anni di catalogazione in URBE: dalle Varianti locali (2009) alla Commissione sulle Varianti locali e la catalogazione in URBE (2024)." In *Parsifal: un modello di collaborazione bibliotecaria per condividere la conoscenza registrata*, a cura di Silvano Danieli, 127-35. Firenze: Firenze University Press. https://doi.org/10.36253/979-12-215-0356-2

Sardo, Lucia. 2021. "L'evoluzione dell'authority control." In *La trasmissione della conoscenza registrata: scritti in onore di Mauro Guerrini offerti dagli allievi*, a cura di Carlo Bianchini, e Lucia Sardo, 379-90. Milano: Editrice Bibliografica. https://www.doi.org/10.3302/0392-8586-202104-052-1

[20] Associazione bibliotecari ecclesiastici italiani 2000, IX-XI. La versione inglese è a pp. XXXIII-XXXV.

Una proposta per la gestione dei fondi personali in Wikidata

Tania Maio

Abstract: Cataloguing personal collections is challenging due to different document types requiring specific standards. Experiments show Wikidata is effective for describing collections as wholes and linking them in a network of relationships but less suited for item-level description, where a tailored Wikibase instance is preferable. A dedicated Wikidata WikiProject and a standard data model for personal collections will enhance data consistency, interoperability, and GLAM professionals' engagement, promoting richer, multilingual metadata and integration with catalogues and information sources.

Keywords: Wikidata, special collections, personal collections.

1. I fondi di persona, una definizione

Le biblioteche e gli archivi d'autore rappresentano un oggetto di studio privilegiato in ambito biblioteconomico. Il motivo di tale interesse risiede, in primo luogo, nella confluenza massiva di raccolte di questo tipo negli istituti culturali italiani verificatasi negli ultimi decenni, che ha portato le istituzioni MAB (Musei, Archivi, Biblioteche) ad accogliere decine di fondi personali e ad affrontare tutte le problematiche connesse alla loro gestione. Inoltre tali raccolte rivestono un ruolo chiave nella ricostruzione storica e nella valorizzazione della tradizione culturale italiana, in quanto testimoni dell'attività e degli interessi culturali di chi li ha prodotti.

Un'adeguata trattazione dei fondi personali richiede, in via preliminare, l'assunzione di una definizione condivisa, come quella proposta nelle *Linee guida sul trattamento dei fondi personali* a cura della Commissione AIB Biblioteche speciali, archivi e biblioteche d'autore:

> Per fondi personali si intendono complessi organici di materiali editi e/o inediti raccolti e/o prodotti da persone significative del mondo della cultura, delle professioni e delle arti prevalentemente dalla seconda metà del XIX secolo in poi (AIB 2019).

All'interno dei fondi personali, si distinguono diverse categorie di beni culturali tra cui: le biblioteche d'autore, raccolte di libri accorpati in maniera funzio-

Tania Maio, University of Salerno, Italy, tmaio@unisa.it, 0000-0003-0832-7529

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Tania Maio, *Una proposta per la gestione dei fondi personali in Wikidata*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.08, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 33-46, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

nale alla propria attività da un soggetto significativo per la comunità culturale; gli archivi di persona, il cui elemento caratterizzante e unificante è l'individuo che li ha prodotti e che contengono documenti legati al percorso di studi e alla formazione, all'attività professionale, alla vita privata e familiare; gli archivi culturali, raccolte (e non già archivi nel senso più stretto del termine) di documenti personali e di libri. I professionisti che trattano tali tipologie documentarie sanno bene come tali entità non sono sempre riconoscibili e riconducibili a un'unica e definita tipologia culturale, tanto è vero che spesso tali tipologie si integrano (Manfron 2011; Manfron 2012), il che rende più complesso il progetto che si sta presentando. Tuttavia, nonostante l'eterogeneità delle sue manifestazioni, l'elemento aggregatore di queste raccolte rimane l'individuo e la raccolta si manifesta come testimone degli interessi, delle attività e delle relazioni della persona nel contesto storico e culturale in cui ha operato.

2. La descrizione dei fondi di persona

Dopo aver chiarito a quali collezioni ci si riferisce, il presente contributo intende stimolare una riflessione condivisa su una delle criticità più rilevanti nel trattamento dei fondi personali: la descrizione complessiva del fondo, che, unitamente a quella del soggetto produttore, rappresenta un nodo ancora irrisolto nell'ambito della prassi biblioteconomica. Nonostante gli sforzi compiuti, non è ancora emersa una soluzione pienamente condivisa, normalizzata e soddisfacente a livello metodologico.

Negli ultimi decenni, si sono fatti significativi passi avanti sulla necessaria descrizione di esemplare, capace di restituire i dati della copia per il libro moderno, così come lo era già per il libro antico (Bernabè 2024, 53). Ne è una dimostrazione il capitolo 7 delle REICAT (Regole italiane di catalogazione) e le ottime linee guida sulla descrizione dell'esemplare adottate da istituzioni guida in questo campo, come la Biblioteca comunale dell'Archiginnasio di Bologna (Farinella 2018).

Tuttavia, se è vero che con il catalogo si può fare molto, ma non si può fare tutto (Petrucciani 2020, 35), è necessario pensare ad un superamento dell'approccio monodirezionale al fondo mediato dal catalogo. L'accesso offerto dal catalogo rimane infatti necessariamente limitato a una logica 'uno a uno', libro per libro, mentre risulta ancora carente la capacità di restituire una visione d'insieme della raccolta. La realizzazione di una scheda fondo e di una scheda soggetto produttore, strutturate secondo criteri condivisi e normalizzati, rimane un obiettivo largamente disatteso nella pratica biblioteconomica, pur essendo da tempo prassi consolidata nell'ambito archivistico. Tale ritardo può essere attribuito, almeno in parte, ai limiti operativi degli strumenti tuttora a disposizione di bibliotecari e archivisti:

> ogni fondo, non solo quelli personali, dovrebbe avere una sua descrizione dettagliata e fin dove possibile normalizzata, posta in collegamento con le descrizioni degli esemplari che ne fanno parte, da un lato, e con l'istituto che lo conserva, dall'altro. Un sistema informativo bibliotecario realmente adeguato,

> insomma, dovrebbe integrare tre componenti che oggi sono disarticolate, oltre che schiacciate dall'enfasi posta quasi soltanto sui record bibliografici: le informazioni sugli istituti bibliotecari (funzione oggi svolta, in parte, dall'Anagrafe), le descrizioni dei loro fondi, quelle degli esemplari e delle edizioni a cui appartengono (Petrucciani 2020, 36).

Per rendere possibile tale evoluzione, si impone un cambiamento di paradigma: il trattamento dei fondi personali richiede competenze sempre più interdisciplinari, che spaziano dalla bibliologia alla filologia, dalla storia del libro e dell'editoria alla biblioteconomia e all'archivistica. A queste si aggiungono le competenze digitali e in particolare quelle legate al web semantico e ai *linked data*, strumenti oggi imprescindibili per chi opera nei beni culturali, e che possono fornire un aiuto per la descrizione dei fondi personali:

> Le future generazioni di bibliotecari dovranno dunque essere preparate all'apprendimento costante, non solo praticando la programmazione informatica e utilizzando l'intelligenza artificiale, ma anche possedendo un adeguato background teorico nei modelli concettuali bibliografici. Risulta fondamentale anche mantenere uno sguardo attento e costante alle tendenze e ai cambiamenti culturali a livello globale, utilizzando un approccio incentrato sull'utente e un'attenzione a garantire la qualità e l'utilità dei metadati.
> [...] con l'aumento della necessità di organizzare quantità di dati e metadati sempre maggiori in modo da poterli trasformare efficacemente in informazione e conoscenza, sono emerse esigenze nuove che non possono più essere affrontate solo dal punto di vista tecnico-informatico e, di conseguenza, viene percepita la necessità di contestualizzare, standardizzare e valorizzare questa tipologia di attività nell'ambito LIS (Boccone 2022, 49).

L'approccio ai fondi di persona necessita di 'integrazione', prima di tutto di competenze professionali, come ricordava anche Alberto Petrucciani:

> Tra biblioteca, archivio, bibliografia, testimonianze scritte, fonti iconografiche, oggetti, e così via. Certamente integrazione con tutti gli altri fondi di persone in relazione con il 'nostro' o la 'nostra' (non esclusi, naturalmente, gli altri fondi conservati nello stesso istituto). Ma anche, più profondamente, integrazione tra il lavoro della biblioteca, dell'archivio, dell'istituto di conservazione, e il lavoro della ricerca, dello studio, della divulgazione (Petrucciani 2020, 33-4).

3. Nuove esigenze dell'utente, nuove funzioni del catalogo

Il mutamento di prospettiva nel trattamento dei fondi personali è strettamente connesso anche all'evoluzione del catalogo, che nel suo progressivo adattamento all'ambiente digitale ha visto modificarsi radicalmente funzioni e finalità, nel tentativo di rispondere alle esigenze di un'utenza sempre più abituata a modalità di ricerca esplorative e trasversali. Nato come strumento di tutela, con l'obiettivo primario di registrare e conservare le collezioni possedute dagli istituti, il

catalogo ha assunto presto la funzione di strumento di accesso al documento, orientato all'individuazione delle risorse da parte dell'utente.

Charles A. Cutter aveva definito tre funzioni fondamentali del catalogo: permettere il reperimento di un libro noto (per autore, titolo o soggetto), mostrare il posseduto della biblioteca in relazione a un determinato autore, soggetto o genere letterario, e facilitare la scelta di un libro attraverso la sua edizione (Cutter 1904). A questa prospettiva si è aggiunta, in tempi più recenti, la funzione di 'esplorazione', che, nel modello concettuale IFLA LRM, riconosce l'importanza delle relazioni tra entità bibliografiche e della scoperta contestuale, anche imprevedibile, delle risorse, valorizzando la dimensione serendipica della ricerca (Riva, Le Bœuf, Žumer 2017)[1]. La rivoluzione determinata dal navigare fra le informazioni impone di mettere in primo piano l'interesse dell'utente. In linea con tali trasformazioni, la *Dichiarazione dei Principi Internazionali di Catalogazione* (ICP) ribadisce che il catalogo dovrebbe essere uno strumento efficace e dinamico, capace di facilitare l'accesso e la navigazione tra domini informativi differenti, attraverso l'impiego di relazioni chiare e visibili tra entità bibliografiche e di autorità, sia all'interno del catalogo stesso sia in contesti esterni ad esso (Bertolini et al. 2016)[2].

L'utente deve essere quindi messo nelle condizioni di esplorare verso altri cataloghi e altri domini di conoscenza, attraverso punti di accesso molteplici, ma che arrivano alla fonte dell'informazione da lui desiderata. Per far sì che ciò avvenga è necessario superare le barriere tra standard descrittivi diversi e rendere possibili percorsi di ricerca trasversali, basati sulle relazioni fra entità e dati informativi. Tuttavia, l'implementazione di tali principi si scontra ancora oggi con i limiti strutturali degli strumenti descrittivi adottati in ambito archivistico e biblioteconomico, i quali non sono pienamente in grado di sfruttare le potenzialità offerte dai *linked data*, né di rendere visibili e interrogabili le relazioni semantiche tra entità culturali.

Infatti, come spiega Marilena Daquino:

> lo scambio di informazioni tra istituzioni affini richiede una prima fase di mappatura dispendiosa. In secondo luogo l'eredità di soluzioni tecnologiche – più o meno aperte e più o meno manutenibili – difficilmente permette di dialogare con fonti di dati esterne senza ulteriori costi e/o cambiamenti radicali. Non ultimo, il dialogo tra fonti richiede la revisione e l'aggiornamento

[1] «La funzione esplorare è la meno definita tra le funzioni utente. L'utente potrebbe trovarsi a navigare, mettere in relazione una risorsa con un'altra, fare collegamenti imprevisti o familiarizzare con le risorse disponibili per un uso futuro. La funzione esplorare riconosce l'importanza della serendipità nella ricerca d'informazione. Per facilitare questa funzione il sistema informativo cerca di agevolare la scoperta creando relazioni esplicite, fornendo informazioni contestuali e funzionalità di navigazione» (Riva, Le Bœuf, e Žumer 2017).

[2] Il catalogo dovrebbe essere uno strumento efficiente ed efficace che consente all'utente: di navigare ed esplorare: all'interno di un catalogo, tramite la disposizione logica dei dati bibliografici e d'autorità e la presentazione chiara delle relazioni tra entità oltre il catalogo, verso altri cataloghi e in contesti non legati alle biblioteche (Bertolini et al. 2016).

> delle basi di conoscenza, previa una sofferta fase di data cleansing. Il costo per aggiornare le descrizioni catalografiche è infatti dispendioso, in termini di risorse umane, expertise e tempo. Ciò comporta una drastica selezione delle informazioni da condividere con enti esterni. La conseguenza immediata è un netto ridimensionamento delle domande di ricerca e delle aspettative a cui l'integrazione potrà dare risposta (Daquino 2019).

In questa situazione Wikidata ben si presta, se non alla descrizione dei singoli esemplari che compongono un fondo personale, di sicuro alla descrizione del fondo nel suo insieme e del suo possessore nonché al collegamento del fondo con le entità a lui legate: la persona che lo ha costituito e/o che lo ha donato, l'ente che lo conserva, gli istituti che possiedono altre parti della stessa raccolta.

4. La descrizione dei fondi speciali in Wikidata

Sebbene la pubblicazione in Wikidata di dati relativi a monografie e periodici sia oggi relativamente agevole, grazie anche al supporto di progetti dedicati e alla presenza di modelli consolidati[3], la questione diventa più complessa per i materiali, documentali e non, delle collezioni speciali e dei fondi personali (libri rari, manoscritti, fotografie, carte d'archivio, oggetti, etc.) che hanno caratteristiche specifiche e in molti casi uniche. Varie sono le criticità in questo senso, ad esempio l'uso molto ampio di note non strutturate, la cui interpretazione richiede un lavoro analitico significativo per l'estrazione di entità e relazioni, la struttura complessa sia fisica sia testuale di tali materiali, l'uso di data model locali al posto di standard e programmi di descrizione collaborativa, la difficoltà di reperire entità correlate – come persone, famiglie o enti – nei repertori di autorità nazionali e internazionali, specialmente quando si tratta di soggetti legati a contesti territoriali locali.

Vari, anche se non molto numerosi, sono i progetti relativi alle collezioni speciali in Wikidata.

La Biblioteca Nazionale del Galles ha integrato i metadati di molte delle sue collezioni speciali in Wikidata, tra cui il Welsh Portrait Archive, la collezione di paesaggi gallesi, gli Aberystwyth Shipping Records, e, in particolare, i manoscritti Peniarth. Questi ultimi, comprendenti 560 codici contenenti testi fondamentali della letteratura gallese e riconosciuti dall'UNESCO nel programma *Memory of the World*, sono stati catalogati, digitalizzati e successivamente sono stati oggetto di un progetto esemplare di descrizione integrale in Wikidata. Nonostante l'ontologia per la descrizione dei manoscritti sia ancora in fase di sviluppo, sono state create nuove proprietà specifiche per rappresentarne adeguatamente le caratteristiche. Gli item sono stati collegati alle immagini digi-

[3] Per le monografie si consulti il Wikidata:WikiProject Books: <https://www.wikidata.org/wiki/Wikidata:WikiProject_Books>, mentre per le pubblicazioni periodiche il riferimento principale è Wikidata:WikiProject Periodicals: <https://www.wikidata.org/wiki/Wikidata:WikiProject_Periodicals>.

talizzate e resi interrogabili attraverso il servizio SPARQL query, che consente analisi visuali e statistiche relative, ad esempio, alle lingue, ai generi e ai soggetti trattati. Il progetto ha inoltre evidenziato le relazioni tra copisti ed esemplari, permettendo di tracciare nuove connessioni tra soggetti e opere. L'impiego dei *linked data* ha ampliato le potenzialità di ricerca e ha favorito l'integrazione delle collezioni con quelle di altri enti, come l'Università di Oxford (Evans 2023).

In ambito italiano, merita attenzione l'iniziativa della Fondazione "Ugo e Olga Levi" di Venezia, che nel 2019 ha avviato una collaborazione con il Dipartimento di Musicologia e Beni Culturali dell'Università di Pavia per migliorare la visibilità e il riuso dei dati descrittivi delle proprie collezioni, adottando Wikidata come strumento principale (Bianchini, Spinelli 2021).

Dopo un'attenta analisi delle raccolte possedute, è stato selezionato il Fondo Gambara, costituito da circa trecento ritratti di musicisti di fine Ottocento, corredati da immagini in pubblico dominio e metadati strutturati. A partire dalla descrizione dei disegni, sono state individuate le entità e le proprietà Wikidata rilevanti, con l'obiettivo di riutilizzare il modello sviluppato anche per la descrizione di altre collezioni della Fondazione e per la disseminazione di buone pratiche ad altre istituzioni culturali.

Nel contesto biblioteconomico, alcune sperimentazioni, anche portate avanti da chi scrive, hanno esplorato la possibilità di descrivere i fondi di persona in maniera analitica, arrivando alla creazione di un item Wikidata per ciascun esemplare. Tale approccio, se da un lato consente interrogazioni avanzate e visualizzazioni articolate, ha evidenziato i limiti intrinseci della piattaforma rispetto alla rappresentazione di collezioni speciali 'libro per libro'. In questi casi, l'uso di un'istanza di Wikibase – software open source su cui si basa Wikidata – risulta più adeguato. Anche OCLC, nel paper "Creating Library Linked Data with Wikibase" (2019), ha indicato Wikibase come strumento preferibile per la gestione di progetti che richiedano un'elevata personalizzazione, lo sviluppo di nuove entità e proprietà, o i cui elementi non soddisfano i criteri di notabilità previsti da Wikidata.

Ecco cosa scrivono i curatori del *position paper*:

> All'inizio del progetto pilota, il team del progetto OCLC ha valutato la possibilità di lavorare direttamente in Wikidata, ma ha optato per l'utilizzo di una nuova e separata istanza di Wikibase per la gestione dei dati del progetto. Questa decisione è stata motivata da:
> - interesse nel valutare il software, le sue opzioni di personalizzazione e configurazione e la sua scalabilità;
> - probabilità che il progetto pilota avrebbe proposto nuove entità e proprietà bibliografiche, che potrebbero non essere rilevanti o accettate per l'inclusione in Wikidata entro la durata del progetto;
> - consapevolezza che i requisiti di notabilità per Wikidata potrebbero differire da quelli per le risorse della comunità bibliotecaria (OCLC 2019).

Infine, anche il confronto con la comunità dei volontari di Wikidata ha confermato che la creazione di item relativi a ogni singolo esemplare, in assenza di

caratteristiche di unicità o rilevanza eccezionale, rischia di compromettere la natura generalista della piattaforma. In tal senso, solo collezioni di eccezionale valore storico e culturale, come i manoscritti Peniarth, giustificano l'adozione di un approccio descrittivo così granulare all'interno di Wikidata.

5. Descrivere un fondo di persona in Wikidata: il flusso di lavoro

L'analisi degli esempi precedenti ha permesso di delineare un flusso di lavoro strutturato, applicabile anche alla descrizione dei fondi di persona.

Questo flusso prende avvio con la fase di raccolta delle informazioni, durante la quale si procede alla compilazione dell'item in Wikidata. Tale fase non si limita alla mera trasposizione di dati già noti, ma richiede spesso un'attività ulteriore di ricerca e approfondimento, soprattutto in presenza di lacune documentarie. La creazione di un item consente di concentrare e rendere accessibili in un unico punto di accesso tutte le informazioni note su un fondo, contrastando così l'oblio che frequentemente colpisce tali dati. La dispersione delle informazioni all'interno delle biblioteche, spesso legate alla memoria individuale dei singoli bibliotecari, può così essere superata grazie alla condivisione su una piattaforma collaborativa.

La seconda fase del flusso di lavoro consiste nella strutturazione dei dati, durante la quale il professionista dell'informazione si impegna a modellare le informazioni raccolte secondo la struttura prevista dal *Wikiproject* di riferimento, quando presente. Trasformare le informazioni testuali in dati strutturati secondo le proprietà di Wikidata consente di valorizzare informazioni altrimenti relegate in campi non strutturati — come il campo note dei record catalografici — rendendole accessibili e intelligibili non solo all'utente finale, ma anche ad altri sistemi informativi.

A questo punto è possibile procedere alla pubblicazione dell'item, che può comportare anche la creazione di nuove entità e la proposta di nuove proprietà o valori utili a colmare lacune ontologiche, come nel caso dei manoscritti gallesi. La collaborazione tra professionisti con competenze differenti favorisce infatti un'evoluzione continua di Wikidata e della sua ontologia.

L'ultima fase del processo riguarda l'arricchimento dell'item e il suo collegamento a risorse esterne, come authority file, cataloghi o enciclopedie, amplificando il valore informativo dell'elemento e la sua interoperabilità.

6. La scheda di rilevazione dei fondi di persona

Il progetto che si propone in questo contributo è nato nel momento in cui la Commissione nazionale Biblioteche speciali, archivi e biblioteche d'autore ha iniziato a lavorare su una scheda di rilevazione dei fondi di persona normalizzata, che possa essere condivisa dagli istituti culturali italiani. Lo studio e la ricerca necessaria a tale scopo hanno naturalmente portato alla lettura della scheda di rilevazione dei fondi librari elaborata da Luigi Crocetti in occasione del progetto curato dalla Regione Toscana sul Censimento dei fondi librari

avviato nel 2001. La scheda di rilevazione è strutturata secondo un modello piuttosto ampio in grado di fornire i dati essenziali sul possessore/i, sulla data e sui modi di acquisizione, sulla storia del fondo, sulla sua consistenza e la tipologia dei materiali contenuti, sullo stato di ordinamento, sulle condizioni di accesso al fondo, il suo stato di conservazione, l'eventuale presenza di contrassegni identificativi, la bibliografia relativa (Desideri 2020, 63; Ricciardi, Calabri 2008, 75)[4]. Lo studio di tale strumento ha reso subito chiaro quanto la raccolta e la strutturazione delle informazioni ad opera del professionista dei beni culturali potesse essere propedeutica alla contestuale creazione di un item Wikidata relativo al fondo stesso. Tuttavia, le attuali descrizioni dei fondi di persona sulla piattaforma sono afflitte da una marcata frammentarietà, sia terminologica che strutturale, e mancanza di standardizzazione, che ostacolano la reperibilità e l'analisi dei dati.

7. Speciale o personale / Fondo o collezione?

Uno dei principali ostacoli nella descrizione dei fondi personali in Wikidata risiede nella questione terminologica. Sebbene il sintagma «fondo personale» sia frequentemente utilizzato nelle etichette di alcuni elementi (attualmente 79), non esiste come entità autonoma all'interno dell'ontologia, determinando così una significativa confusione e dispersione semantica. In assenza di un item specifico, questi fondi vengono spesso ricondotti alla classe più generica di «fondo speciale» (Q4431094), che conta 386 occorrenze, benché i due concetti non siano sovrapponibili. A ciò si aggiungono 10 item classificati come «biblioteca personale» (Q106402388) e 465 come «collezione bibliotecaria» (Q856592), tra cui figurano anche alcune biblioteche personali[5].

La confusione relativa alla terminologia da utilizzare quando si tratta di fondi di persona, non è presente solo in Wikidata, come dimostra l'eterogeneità dei termini proposti nei *thesauri* internazionali. Un'analisi dei principali vocabolari controllati mostra come il termine «fondo di persona» o «fondo personale» venga solitamente incluso nella più ampia categoria di «collezione» o «fondo speciale». Tuttavia, anche per questa categoria non esiste una condivisa connotazione dei contorni nemmeno nella pratica biblioteconomica angloamericana, dove tale concetto ha origine (Bernabè 2024, 53). Occorre rilevare che il concetto stesso di «fondo» non trova una diretta traduzione nella lingua inglese, che lo sostituisce con il termine più ampio *collection*. Il concetto di «collezione» è inoltre preferito nella tradizione biblioteconomica anglo-americana e tedesca,

[4] La scheda di rilevazione è disponibile in Wikimedia Commons <https://commons.wikimedia.org/wiki/File:Scheda_rilevazione_fondi_librari.pdf>.

[5] Se si estende l'analisi all'ambito archivistico, si rilevano oltre 100.000 item identificati come «fondo archivistico», 19 come «archivi di persona» (Q27032347), un solo elemento descritto come «archivio personale» (Q7170558) e 18 come «archivio di famiglia» (Q27032283). Per gli archivi si veda il Wikiproject Archival Description <https://www.wikidata.org/wiki/Wikidata:WikiProject_Archival_Description>.

mentre ritroviamo il termine «fondo» nel *thesaurus* adottato dalle biblioteche nazionali in Italia, Francia, Spagna.

Nell'Art and Architecture Thesaurus della Getty Foundation, le *special collections* sono definite come «collections of library materials separated from the general collection because they are rare and valuable», enfatizzandone la separazione fisica e concettuale dalle raccolte comuni. Diversamente, nel contesto italiano, queste raccolte non sono necessariamente separate, ma rappresentano spesso i fondi storici fondanti di una biblioteca, contribuendo in modo sostanziale alla sua identità. Il *Nuovo Soggettario* della Biblioteca Nazionale Centrale di Firenze definisce infatti i fondi speciali come «raccolte di libri e documenti riuniti insieme per affinità di argomento o provenienti da lasciti o donazioni». Il Library of Congress Subject Headings (LCSH) norma l'uso del sintagma *Special collections* per descrivere sia le collezioni speciali all'interno delle biblioteche sia le pratiche connesse alla loro gestione[6]. La Biblioteca Nazionale Tedesca, nel Gemeinsame Normdatei, adotta il termine *Spezialsammlung*, traducibile con collezione speciale, mentre la Biblioteca Nazionale di Francia privilegia *Fonds spéciaux*, allineandosi così alla tradizione italiana.

Ma esiste una differenza concettuale tra «fondo» e «collezione speciale»?

Anna Bernabè fornisce un utile approfondimento tra la scelta dei due termini:

> In sintesi, saranno considerati "patrimoni storici" quelli in dotazione alle biblioteche d'ateneo e soggetti a conservazione – quindi sia contenitori sia contenuti. In merito a questi ultimi, si opterà per il termine "fondi" e non "collezioni". Saranno intesi come "fondi speciali" quei complessi bibliografici o documentari caratterizzati da una specialità funzionale per la biblioteca e l'ateneo, che vada di pari passo con la specialità costitutiva (rarità o unicità, vetustà, provenienza specifica, compiutezza in sé ecc.). Per ciò che è speciale si manifesteranno ricadute gestionali sulle risorse umane e finanziarie, oltre che sugli spazi da destinare, e saranno adottate adeguate modalità di acquisizione, di conservazione e di fruizione, volte a distinguere quella porzione dal resto del patrimonio. Naturalmente il patrimonio speciale sarà a buon diritto riconducibile alla più ampia nozione di patrimonio storico della biblioteca di ateneo. Si potrà infine contemplare il raro caso delle "collezioni speciali", qualora esista una continuità strategicamente pianificata nell'acquisizione di materiali diversi rispetto alle collezioni correnti a supporto di didattica e ricerca: può essere il caso, per esempio, di collezioni in fieri di vinili o di pellicole d'epoca, di libri antichi o di libri d'artista (Bernabè 2025, 75).

[6] «Here are entered works describing special collections in libraries as well as the methods used to acquire, process, maintain, and provide service using these collections. This heading may be subdivided by the topic or form of the special collections, e.g., [Libraries--Special collections--Chinese literature; Libraries--Special collections--United States; Libraries--Special collections--Slides (Photography)» (LCSH).

Una volta stabilito che le raccolte oggetto di questo contributo, i fondi di persona o fondi personali, sono da intendersi come fondi e non come collezioni, occorre interrogarsi su come descriverle in Wikidata, considerando che esse rappresentano una sottocategoria più specifica rispetto ai più generici fondi speciali.

8. I fondi personali in Wikidata: una proposta di data model

La ricognizione effettuata in Wikidata evidenzia una forte eterogeneità terminologica e concettuale, segno dell'assenza di un vocabolario condiviso. Questa situazione rende evidente la necessità di costruire un set di termini controllati e di creare item specifici per la descrizione di fondi personali, biblioteche d'autore e archivi culturali. Tali item potranno essere descritti rifacendosi alle definizioni elaborate dalla Commissione nazionale Biblioteche speciali, archivi e biblioteche d'autore, garantendo una coerenza semantica che permetta l'utilizzo nelle varie versioni linguistiche dei progetti Wikimedia.

In questo contesto, la creazione di un Wikidata:Wikiproject dedicato ai fondi personali potrebbe costituire una risposta strutturata al problema, fungendo da punto di raccordo e fonte di buone pratiche per quanti desiderano inserire un fondo personale in Wikidata. Gli obiettivi del progetto includerebbero:

- l'inserimento, l'arricchimento e la valorizzazione dei dati sui fondi personali in Wikidata;
- l'implementazione e il mantenimento di ontologie e *thesauri* multilingue per la loro descrizione;
- l'interconnessione tra i cataloghi esistenti e Wikidata;
- l'integrazione dei dati all'interno di Wikipedia e dei progetti Wikimedia correlati.

In particolare, l'item «Fondo di persona/personale» (da intendersi come possibile sottoclasse di «fondo speciale» – Q4431094) dovrebbe essere descritto in base a un data model che rifletta i contenuti della scheda di rilevazione dei fondi personali a cui la Commissione Biblioteche speciali, archivi e biblioteche d'autore sta lavorando e che si ispira al documento curato da Luigi Crocetti. Ogni proprietà di Wikidata trova un corrispettivo nei campi previsti dalla scheda, permettendo una mappatura coerente tra le due strutture descrittive. La comunità di Wikidata sarà chiamata, inoltre, a valutare la creazione di nuove proprietà qualora quelle esistenti non siano sufficienti a restituire la complessità delle informazioni.

Il data model proposto rappresenta una base iniziale, pensata per mappare i campi fondamentali della scheda di rilevazione dei fondi personali; tuttavia, esso è da considerarsi aperto e in continua evoluzione. Saranno infatti i contributi, i suggerimenti e le esigenze emerse dalla comunità dei bibliotecari e dei professionisti dell'informazione a guidarne l'arricchimento e la progressiva articolazione, affinché possa rispondere in maniera sempre più efficace alla complessità, varietà e specificità dei fondi di persona.

Tabella 1 – Proposta di mappatura tra le proprietà Wikidata e scheda di rilevazione.

Proprietà	Descrizione	Valore
Istanza di (P31)	Tipologia di fondo	Biblioteca personale (Q106402388), archivio di persona (Q27032347), fondo personale (Q134886297) …
Creatore (P170)	Soggetto produttore / creatore della collezione	Q creatore
Prende il nome da (P138)	Soggetto produttore (in funzione eponima)	Q creatore
Proprietario (P127)	Ente ricevente (istituzione che possiede il fondo) con qualificatore data inizio (P580) e data fine (P582)	Più valori possibili
Luogo (P276)	Luogo in cui si trova la raccolta	Q del luogo
Data di fondazione o creazione (P571)	Data di acquisizione	Data
Parte di (P361)	Biblioteca o ente di collocazione	Q dell'ente
Tipologia contrattuale	Modalità di acquisizione	Comodato (Q1580844), deposito (Q2638043), compravendita (Q1054726), legato (Q211557), donazione (Q707482)
Livello di descrizione (P6224)	Livello di descrizione	Fondo archivistico (Q3052382), serie (Q3511132), unità archivistica (Q11723795)
Catalogo online (P8768)	Strumenti di ricerca	URL
Contiene elementi del tipo (P2670)	Ha elementi di questo tipo	Libro (Q571), periodico (Q1002697), documento (Q49848)
principale (P921)	Principali aree tematiche	Q degli argomenti
Dimensione della collezione o esposizione (P1436)	Estensione della collezione	Metri lineari o quantità di documenti
Restrizioni all'accesso (P7228)	Modalità di consultazione	Senza restrizioni (Q66739888), accesso completamente riservato (Q66739729), accesso parzialmente riservato (Q66739849)
Storico delle esposizioni (P608)	Eventi legati alla valorizzazione	Q degli eventi (mostre, giornate di studio, ecc.)

9. Conclusioni

L'integrazione dei fondi di persona in Wikidata non rappresenta solo una sfida tecnica o terminologica, ma si configura come una vera e propria opportunità strategica per la valorizzazione di tale patrimonio bibliografico e documentario conservato dalle istituzioni MAB.

Il lavoro in Wikidata permette di avere facile accesso a tools di interrogazione e visualizzazione, permettendo di esplorare le collezioni in modi nuovi; agisce da hub, collegando raccolte ed istituzioni nel web dei dati del patrimonio culturale, in continua espansione; rende possibile l'arricchimento dei cataloghi con le informazioni 'catturate' da altri domini di conoscenza.

La costruzione di un modello descrittivo condiviso dei fondi di persona, fondato su strumenti di rilevazione normalizzati, può garantire la creazione di dati di qualità, interoperabili e orientati al riuso. In questo scenario, il ruolo del bibliotecario si evolve e diventa mediatore attivo tra la tradizione catalografica e il web dei dati, capace di contribuire alla modellazione della conoscenza e ampliare la visibilità del patrimonio culturale attraverso strumenti aperti e partecipativi.

Riferimenti bibliografici

AIB. 2019. *Linee guida sul trattamento dei fondi personali*, a cura della Commissione nazionale biblioteche speciali, archivi e biblioteche d'autore (versione 15.1 – 31 marzo 2019). <https://www.aib.it/documenti/linee-guida-sul-trattamento-dei-fondi-personali/>.

Archives and Special Collections Linked Data Review Group, OCLC Research. 2020. "Archives and Special Collections Linked Data: Navigating between Notes and Nodes." <https://www.oclc.org/research/publications/2020/oclcresearch-archives-special-collections-linked-data.html>.

Association of Research Libraries. 2019. "ARL White Paper on Wikidata Opportunities and Recommendations." <https://www.arl.org/resources/arl-whitepaper-on-wikidata/>.

Baldacchini, Lorenzo, e Anna Manfron. 2015. "Dal libro raro e di pregio alla valorizzazione delle raccolte." In *Biblioteche e biblioteconomia. Principi e questioni*, a cura di Giovanni Solimine, e Paul Gabriele Weston, 315-50. Roma: Carocci editore.

Baldacchini, Lorenzo. 2023. *Le collezioni speciali: esperienze ed orizzonti : atti della giornata di studio promossa da Biblioteca nazionale centrale di Roma, Commissione nazionale AIB Biblioteche speciali, archivi e biblioteche d'autore, AIB Sezione Lazio (Roma, 14 ottobre 2022)*. Roma: AIB.

Bernabè, Anna. 2024. *Biblioteche, patrimonio culturale, Università: valutazione e impatto*. Roma: AIB. https://doi.org/10.53263/9788878124004

Bianchini, Carlo, e Pasquale Spinelli. 2020. "Wikidata at Fondazione Levi (Venice, Italy) : A Case Study for the Publication of Data about Fondo Gambara, a Collection of 202 Musicians' Portraits." *JLIS.it* 11, 3: 16-38. https://doi.org/10.4403/jlis.it-12648

Bianchini, Carlo, Stefano Bargioni, e Camillo Carlo Pellizzari di San Girolamo. 2021. "Beyond viaf: Wikidata as a complementary tool for authority control in libraries." *Information technology and libraries*, 40, 2. https://doi.org/10.6017/ITAL.V40I2.12959

Bianchini, Carlo. 2012. "Dagli OPAC ai library linked data: come cambiano le risposte ai bisogni degli utenti." *AIB studi* 52, 3: 303-23. https://doi.org/10.2426/aibstudi-8597

Boccone, Alessandra. 2022. "The Role of the Wikidata Librarian in a Renewed Bibliographical Universe: next Generation Metadata, next Generation Librarians." *JLIS.It* 13, 2: 45-57. https://doi.org/10.36253/jlis.it-460

Cutter, Charles Ammi. 1904[4]. *Rules for a Dictionary Catalog*. London: Library Association.

Daquino, Marilena. 2019. "Archivi Fotografici per la Storia dell'Arte e Semantic Web. Problemi, Risorse e Linee di Ricerca." *JLIS* 10, 2: 37-47. https://doi.org/10.4403/jlis.it-12533

Desideri, Laura. 2020. "La scheda-fondo di Luigi Crocetti." In *Storie d'autore, storie di persone. Fondi speciali tra conservazione e valorizzazione*, a cura di Francesca Ghersetti, Annantonia Martorano, e Elisabetta Zonca, 63-74. Roma: AIB.

Di Domenico, Giovanni. 2020. *Il privilegio della parola scritta. Gestione, conservazione e valorizzazione di carte e libri di persona*. Roma: AIB.

Evans, Jason, 2023. "Treasured Manuscript Collection gets the Wikidata Treatment, Digital Scholarship at The National Library of Wales." <https://datanlw.hcommons.org/2023/09/22/hello-world/>.

Farinella, Laura Tita. 2018. "Linee guida adottate in Archiginnasio per la descrizione degli esemplari." <http://badigit.comune.bologna.it/books/bollettino/pdf/2018-6.pdf>.

Galeffi, Agnese, María Violeta Bertolini, Robert Bothmann, Elena Escolano Rodríguez, e Dorothy McGarry. 2016. "Dichiarazione di Principi Internazionali di Catalogazione (ICP)." IFLA. <https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/icp/icp_2016-it.pdf>.

Ghersetti, Francesca, Annantonia Martorano, ed Elisabetta Zonca. 2020. *Storie d'autore, storie di persone. fondi speciali tra conservazione e valorizzazione*. Roma: AIB.

Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Kalan Knudson Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, Xiaoli Li, Marc McGee, Karen Miller, Honor Moody, Craig Thomas, and Holly Tomren. OCLC. 2019. "Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage." <https://www.oclc.org/research/publications/2019/oclcresearch-creating-library-linked-data-with-wikibase-project-passage.html>.

Guerrini, Mauro, Giuliano Genetasio, e Attilio Mauro Caproni. 2012. *I Principi internazionali di catalogazione (ICP): universo bibliografico e teoria catalografica all'inizio del XXI secolo*. Milano: Ed. bibliografica.

Manfron, Anna. 2008. "Dai libri alle carte: la gestione dei materiali anfibi." *Antologia Vieusseux* 14, 41-2: 63-74

Manfron, Anna. 2011. "Biblioteche e archivi d'autore: le relazioni da preservare." In *Spigolature d'archivio: contributi di archivistica e storia del progetto Una città per gli archivi*, a cura di Armando Antonelli, 323-43. Bologna: Bononia University press.

Manfron, Anna. 2012. "Biblioteca e archivio di persona: da fondo speciale a complesso documentario." In *Archivi di persona del Novecento: guida alla sopravvivenza di autori, documenti e addetti ai lavori*, a cura di Francesca Ghersetti, e Loretta Paro, 39-49. Treviso: Fondazione Benetton Studi Ricerche-Fondazione Giuseppe Mazzotti per la civiltà veneta.

Nepori, Francesca. 2017. "SBN tra funzioni catalografiche e aspirazioni bibliografiche." *Bibliothecae.it* 6 (giugno): 265-97. https://doi.org/10.6092/ISSN.2283-9364/7030

Petrucciani, Alberto. 2008. "Biblioteche d'autore in biblioteca: una catalogazione speciale?" *Antologia Vieusseux* 14, 41-2: 49-62.

Petrucciani, Alberto. 2020. "Fondi e collezioni personali: alcune questioni." In *Storie d'autore, storie di persone. Fondi speciali tra conservazione e valorizzazione*, a cura di Ghersetti Francesca, Martorano Annantonia, ed Elisabetta Zonca, 31-6. Roma: AIB.

Ricciardi, Paola, e Cecilia Calabri. 2008. "Le biblioteche d'autore nel Censimento dei fondi librari della Regione Toscana: tipologie e localizzazioni." *Antologia Vieusseux* 14, 41-2: 75-106.

Riva, Pat, Patrick Le Bœuf, and Maja Žumer. 2017. "IFLA Library Reference Model. Un modello concettuale per le informazioni bibliografiche." <https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712-it.pdf>.

Mapping UNIMARC to BIBFRAME: The SHARE Catalogue Knowledge Base on Wikibase.cloud

Claudio Forziati, Alessandra Moi

Abstract: SHARE Catalogue Mapping Knowledge Base is a project, hosted on Wikibase.Cloud, that aims to represent the mapping between the UNIMARC bibliographic format and the BIBFRAME ontology. This mapping, carried out by a librarian team between 2022 and 2023, is intended to enable SHARE Catalogue to adopt the Share Family's new LOD Platform, along with its bibliographic entity representation and the Linked Data Editor. The team also chose Wikibase technology to make the processed mapping accessible for widespread use in the professional community. Furthermore, the SHARE Catalogue team sought to experiment with a modeling of UNIMARC structured in statements, defining all the key elements of the format, their sources, their correspondence in external projects, and their matching with BIBFRAME specifications.

Keywords: Wikibase, Linked Open Data, UNIMARC, BIBFRAME, SHARE Catalogue.

1. Introduzione[1]

"SHARE Catalogue Mapping Knowledge Base" è un progetto nato come estensione dell'attività tecnica, svolta tra 2022 e 2023, utile all'aggiornamento del catalogo collettivo in Linked Open Data SHARE Catalogue. Questa attività, consistente in un'inedita mappatura tra il formato UNIMARC, utilizzato nei sistemi catalografici della maggior parte degli atenei aderenti alla convenzione SHARE (Scholarly Heritage and Access to Research)[2], e l'ontologia BIBFRAME nella sua versione 2.0 e successivi aggiornamenti, ha consentito l'adozione di una nuova piattaforma per la rappresentazione delle risorse bibliografiche, già in uso in iniziative internazionali come Share VDE[3]. Nella prima incarnazione di SHARE Catalogue, online dal 2016, la necessità di pubblicare i dati in RDF ha impli-

[1] Gli autori ringraziano le colleghe Annalisa Di Sabato (@Cult), Rossella Molisso (Università degli studi di Napoli Federico II) e Chiara Mugnano (Università degli studi di Salerno) per la partecipazione attiva al progetto di conversione e pubblicazione della mappatura su Wikibase.Cloud.

[2] Il testo della convenzione è reperibile a partire da Universities SHARE (Scholarly Heritage and Access to Research), <https://sharecampus.it/>.

[3] Share VDE, <https://www.svde.org/>.

Claudio Forziati, University of Naples Federico II, Italy, claudio.forziati@unina.it, 0000-0002-8976-0393
Alessandra Moi, @Cult, Italy, alessandra.moi@atcult.it, 0000-0003-0104-4999

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Claudio Forziati, Alessandra Moi, *Mapping UNIMARC to BIBFRAME: The SHARE Catalogue Knowledge Base on Wikibase.cloud,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.09, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 47-54, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

cato l'uso di diversi meccanismi di conversione, in particolare da UNIMARC a MARC 21, per giungere alla trasformazione in BIBFRAME. Seppur funzionale, questa doppia conversione ha reso più complesso il processo di pubblicazione dei dati e ha, di fatto, impedito un efficace aggiornamento di SHARE Catalogue alla versione 2 di BIBFRAME, in assenza innanzitutto di una traduzione aggiornata tra i due formati MARC. Per questo motivo si è reso necessario il lavoro di mappatura, condotto da un gruppo di lavoro formato da bibliotecari delle istituzioni aderenti alla convenzione e da analisti dell'azienda @Cult. Data la sua complessità, la mappatura ha visto la sua forma definitiva solo in seguito a diversi passaggi di lavorazione e di analisi, passaggi che ne consentissero un uso coerente con le caratteristiche tecniche di una rinnovata LOD platform. L'adozione della nuova piattaforma di rappresentazione dei dati ha pertanto costituito l'occasione per superare la macchinosità dei precedenti processi di conversione, sfruttando proprio il potenziale della trasformazione diretta tra UNIMARC e BIBFRAME[4].

Terminata la prima versione stabile della mappatura, periodicamente oggetto di aggiornamento, è emersa nel gruppo di lavoro una nuova esigenza, ovvero quella di rendere disponibile alla comunità professionale ciò che si era realizzato, tenendo conto altresì dei principi di apertura e interoperabilità a cui sia la convenzione sia i suoi membri si sono ispirati per realizzare le proprie iniziative, coerentemente inoltre con l'appartenenza alla più ampia famiglia di progetti che va sotto il nome di Share Family[5].

Allo scopo di garantire la possibilità di un ampio riuso, il gruppo ha ritenuto utile convertire il lavoro svolto dalla forma tabellare in cui era stato elaborato in una struttura basata su triple, utilizzando campi e sottocampi UNIMARC come soggetti, nel tentativo di mantenere integre le informazioni rilevanti e di arricchirle di contesto, fonti e collegamenti a dataset dello stesso ambito liberamente disponibili. L'esperienza maturata con Wikidata dai componenti del gruppo di lavoro ha suggerito l'adozione di Wikibase, estensione semantica di MediaWiki, come soluzione.

Wikibase è un software open source in grado di gestire dati strutturati in una knowledge base collaborativa centralizzata. Il data model di Wikibase offre uno schema semantico flessibile, non legato a un'ontologia specifica, che in tal modo favorisce l'integrazione di diverse tipologie di metadati. Una struttura di questo tipo permette di pubblicare i dati nel web semantico rendendoli riutilizzabili e valorizzandoli, per un uso diffuso a partire dalla molteplicità di formati in cui sono trasformabili automaticamente (JSON, RDF, Turtle ecc.).

2. La mappatura UNIMARC-BIBFRAME su Wikibase.Cloud

Negli ultimi anni, numerosi gruppi di ricerca e istituzioni di ambito GLAM hanno utilizzato Wikibase per la realizzazione di specifici progetti locali, uso in-

[4] Per una descrizione dettagliata dell'evoluzione di SHARE Catalogue, dalla prima all'attuale piattaforma LOD, si veda (Forziati et al. 2024).

[5] Share Family: Linked Data Ecosystem, <https://www.share-family.org/>.

coraggiato dalla stessa Wikimedia Foundation in favore di un 'alleggerimento' dei dati in Wikidata. Tra essi si ricorda, per vicinanza concettuale, la sperimentazione della Biblioteca Nazionale Centrale di Firenze sull'esportazione in LOD dei metadati di un campione di record in formato UNIMARC (Bergamin e Bacchi 2018).

A differenza del progetto fiorentino, basato su un'istanza locale di Wikibase, il gruppo di lavoro ha optato per la versione Wikibase.Cloud, avviata nel 2022 e fornita da Wikimedia Deutschland (WMDE)[6]. La scelta della versione *in cloud* rispetto all'installazione locale è stata motivata dalla forte operatività della piattaforma, dotata di strumenti di modifica e del query service preinstallati, nonché dall'assenza di processi e costi per manutenzione e aggiornamento per chi gestisce e arricchisce i wiki.

Oltre al setting dell'istanza e alla realizzazione delle pagine di documentazione, il passaggio preliminare è stato quello della modellizzazione di UNIMARC, oggetto di diverse proposte e discussioni all'interno del gruppo. La prima questione rilevante è stata decidere quali elementi (blocchi, campi, indicatori, sottocampi) fosse necessario rappresentare come elemento (item), esprimendo le loro caratteristiche tramite dichiarazioni in forma di triple.

In generale, si è cercato di fornire per tutti gli item una label formulata in tre lingue, italiano, inglese e francese, con, in aggiunta, le descrizioni per gli item relativi ai campi. Mentre per l'inglese ci si è potuti affidare alla documentazione ufficiale, in particolare agli aggiornamenti di UNIMARC Bibliographic 3rd edition (IFLA 2022), per la lingua francese ci si è avvalsi dei riferimenti presenti nella traduzione a cura del Comité français UNIMARC (CfU), reperibili nel sito del programma *Transition bibliographique* (Comité français UNIMARC, s.d.). Più complessa la scelta delle fonti per l'italiano, poiché il gruppo di lavoro si è trovato a dover mediare tra l'unico repertorio, piuttosto datato, liberamente disponibile online[7] e le traduzioni effettuate dal gruppo stesso sulla base del manuale ufficiale e dei citati aggiornamenti presenti nel sito dell'IFLA.

Oltre alle label nelle tre lingue, tutti gli item sono stati dotati di alias, particolarmente significativi per i sottocampi perché utili a disambiguare rispetto a tutti gli altri che, in UNIMARC, presentano la stessa denominazione.

Le dichiarazioni definiscono per ogni item la posizione gerarchica, in riferimento ad esempio al blocco o al campo di inclusione tramite proprietà reciproche[8], ma anche l'occorrenza, la ripetibilità e il riferimento all'aggiornamento utilizzato, proprietà quest'ultima con datatype 'Punto nel tempo' e alle cui dichiarazioni è stato associato il collegamento all'URL del documento o della pagina web utilizzati come fonte.

[6] La presentazione e la documentazione relative a Wikibase.Cloud sono reperibili agli indirizzi <https://meta.wikimedia.org/wiki/Wikibase/Wikibase.cloud> e <https://www.mediawiki.org/wiki/Wikibase>.

[7] Unimarc-IT. Unimarc bibliographic in italiano, <http://unimarc-it.wikidot.com/>.

[8] Si tratta delle proprietà 'comprende' <https://unimarc2bibframe.wikibase.cloud/wiki/Property:P3> ed 'è compreso in' <https://unimarc2bibframe.wikibase.cloud/wiki/Property:P4>.

Inoltre, sono state create proprietà utilizzabili come qualificatori per esprimere particolari condizioni d'uso in UNIMARC (ad esempio, un sottocampo con occorrenza *Mandatory* quando in un record bibliografico devono essere descritti specifici tipi di risorse).

Tabella 1 – Esempio di modellazione degli item di tipo campo.

Tipo Item	Campo	
Label	**Descrizione**	**Alias**
stringa IT, EN, FR	stringa IT, EN, FR	stringa IT, EN, FR
Proprietà	**Datatype**	**Valore**
istanza di (P1)	elemento	tag bibliografico (Q125)
è compreso in (P4)	elemento	blocco in cui è posizionato il campo
indicatore 1 (P11)	elemento	primo indicatore del campo o non definito (Q2753)
qualificatore: valore per l'indicatore (P13)	elemento	item che rappresentano i valori per lo specifico indicatore in UNIMARC
indicatore 2 (P12)	elemento	secondo indicatore del campo o non definito (Q2753)
qualificatore: valore per l'indicatore (P13)	elemento	item che rappresentano i valori per lo specifico indicatore in UNIMARC
occorrenza (P6)	elemento	obbligatorio (Q19)/opzionale (Q20)
qualificatore: condizione (P8)	stringa	particolari condizioni d'uso relative alla descrizione di risorse specifiche in UNIMARC
ripetibilità (P7)	elemento	ripetibile (Q21)/non ripetibile (Q22)
aggiornamento utilizzato (P5)	punto nel tempo	anno dell'aggiornamento utilizzato
fonte: URL di riferimento (P9)	URL	URL del documento/pagina web utilizzati come fonte per la definizione temporale del campo
mappatura descritta in forma tabellare all'URL (P14)	URL	tabella in pagina interna nel namespace *Project*
ha come conversione BIBFRAME	elemento	item della conversione, composto dagli elementi della tripla BIBFRAME
qualificatore: istruzione BIBFRAME aggiuntiva	stringa	indicazione della regola che completa la conversione

Tra gli item gestiti compaiono due tipologie inizialmente non previste: l'item indicatore e l'item valore dell'indicatore. La decisione iniziale per gli indicatori era di inserire i relativi valori tramite stringhe associate agli item dei campi UNIMARC. Tuttavia, in seguito ai test di estrazione dei dati con apposite query SPARQL, l'uso delle stringhe è stato abbandonato in favore di una gestione più complessa, che desse maggiore conto della granularità dei formati MARC.

In una prima fase, tutte le tipologie di item sono state create manualmente soprattutto per consentire ai componenti del gruppo di prendere dimestichezza con il nuovo ambiente, ma anche per poter verificare, su un numero ristretto di item, se le scelte fatte in sede di modellazione fossero corrette. Successivamente, la modalità manuale, comunque utile per correzioni o integrazioni specifiche, è stata affiancata da strumenti di editing massivo, tra cui il tool QuickStatements.

Allo stato attuale, è pressoché completata la creazione degli item relativi a UNIMARC, con i seguenti risultati[9]:

- Item totali: 2784, di cui,
  - item dei blocchi: 9,
  - item che sono istanza di (P1) tag bibliografico (Q125): 231,
  - item che sono istanza di (P1) sottocampo bibliografico (Q2059): 2091,
  - item che sono istanza di (P1) indicatore per i tag di UNIMARC Bibliographic (Q3300): 129,
  - item che sono istanza di (P1) valore assegnato a un indicatore (Q3836): 276,
  - triple totali: 90797.

Per quanto concerne UNIMARC, resta da completare la modellazione del LEADER e dei sottocampi contenenti valori posizionali, in quanto la loro articolazione richiede un'analisi dedicata supplementare e processi di duplicazione o, preferibilmente, di importazione dei vocabolari controllati che ne rappresentano i valori[10].

3. Arricchimento dei dati e conversione in BIBFRAME

Nell'ambito delle attività in corso, il gruppo di lavoro ha pianificato varie tipologie di arricchimento degli item. Tra queste va sicuramente citato l'inserimento degli URI di UNIMARC Element Sets[11] tramite una proprietà dedicata. Questo mapping, volto a garantire la più ampia interoperabilità del progetto con il contesto internazionale, non è tuttavia privo di complessità, dovuta principalmente a una diversa modellazione degli elementi di UNIMARC nei due progetti. Mentre nell'istanza di Wikibase.cloud, ciascun campo, indicatore e sottocampo sono individuati nella loro unicità, come singoli elementi, nell'UNIMARC Element Sets viene previsto un URI che identifica ciascun campo, indicatore e sottocampo in maniera concatenata; quindi, per esempio, al sottocampo $a del

[9] Rilevazione aggiornata al 18 aprile 2025.

[10] Il gruppo di lavoro sta testando la creazione di tabelle, in un *namespace* dedicato del wiki, contenenti sia i valori di riferimento che gli identificatori, rilevati da Wikidata e dai *Value vocabularies* presenti in iflastandards.info, come nell'esempio presente all'URL <https://unimarc2bibframe.wikibase.cloud/wiki/Project:UNIMARC_Value_Vocabularies/Visual_projections_etc._-_Form_of_release_-_videorecording>.

[11] UNIMARC Element Sets, <https://www.iflastandards.info/unimarc>.

campo 700 vengono attribuiti due diversi URI sulla base dei valori degli indicatori alternativamente utilizzati nel campo[12].

La parte sinora più impegnativa si è rivelata la creazione e l'inserimento delle corrispondenze BIBFRAME nei campi e sottocampi UNIMARC, parte che, per le sue implicazioni sull'efficacia del data model, ha spinto il gruppo di lavoro alla creazione di un'istanza di test, sempre *in cloud*, per valutare pro e contro di ogni scelta di modellazione per le classi, le proprietà e le relative conversioni BIBFRAME.

Pur non essendo ancora giunti a una scelta definitiva, l'approccio è sicuramente orientato, così come fatto in precedenza per UNIMARC, a una gestione 'per item' in tutti i casi questo sia possibile, rispetto all'inserimento di valori in forma di stringhe.

Nella mappatura UNIMARC-BIBFRAME 2 ci si trova davanti a casi come nella tabella 2:

Tabella 2 – Mappatura del campo UNIMARC 318.

318 - Action Note (R)	I - note - Note ; rdf:type rdf:resource="http://id.loc.gov/vocabulary/mnotetype/action"
Subfield Codes	
$k - Action Agent (R)	## - agent - Agent rdfs:label

Nell'istanza di test le conversioni utilizzate nella mappatura sono state definite attraverso item specifici, creati successivamente all'importazione di un set individuato di classi, sottoclassi e proprietà.

Come si può notare, la conversione nella tabella 2 investe due livelli: quello del campo, in cui è riportata una tripla completa di BIBFRAME, e quella del sottocampo che, richiamandosi alla tripla precedente, con il simbolo # dettaglia maggiormente l'informazione.

In entrambi i casi, si è optato per creare due elementi: l'item *I - note - Note* e l'item *## - agent - Agent - rdfs:label*, entrambi istanza di *conversione BIBFRAME* e collegati rispettivamente all'item del campo e all'item del sottocampo tramite la proprietà *ha come conversione BIBFRAME*.

I due item della conversione sono definiti tramite la relazione agli item dei singoli elementi che li compongono: nel caso di *I - note - Note* avremo dunque una proprietà *composto da* che relaziona l'intero oggetto della conversione agli item *Instance* (che è istanza di: classe BIBFRAME), *note* (istanza di: proprietà BIBFRAME) e *Note* (istanza di: classe BIBFRAME)[13].

Inoltre, sono state definite anche le eventuali istruzioni aggiuntive per la conversione tramite una proprietà che funziona da qualificatore e a cui è asso-

[12] Nel caso indicato sono <http://iflastandards.info/ns/unimarc/unimarcb/elements/7XX/U700_0a> e <https://www.iflastandards.info/unimarc/unimarcb/elements/7XX#U700_1a>.

[13] Il caso descritto della tripla *I-note-Note* è reperibile nel wiki di test all'URL <https://sharecat-test.wikibase.cloud/wiki/Item:Q47>.

ciato un datatype di tipo stringa, scelta inevitabile vista la grande varietà di tali istruzioni[14].

La scelta di esprimere le classi e le proprietà di BIBFRAME come item garantisce la possibilità di utilizzarli in tutte le triple della mappatura in cui sono impiegati. Inoltre, ciò permette di avere, tramite il query service SPARQL, un'immediata visione di insieme su come una specifica proprietà o classe venga modellizzata nella conversione dei formati MARC.

Ciascun item che costituisce quello della conversione è collegato alla descrizione nell'ontologia presente nel sito della Library of Congress, tramite l'inclusione nel progetto di identificatori esterni che puntano ad essa grazie ad un'apposita proprietà. Quest'ultima è sicuramente annoverabile tra le attività di arricchimento, perché pensata per fornire a tutti gli item che rappresentano BIBFRAME nel progetto gli elementi utili per una corretta definizione, tracciandone anche i casi d'uso e gli eventuali aggiornamenti.

Infine, una ulteriore attività di arricchimento e ampliamento del wiki può essere individuata nella progressiva creazione, in un *namespace* specifico dell'istanza, di tabelle corrispondenti a quelle della mappatura, opportunamente segmentate sulla base dei campi UNIMARC[15]. Questa attività ha un duplice scopo: quello di documentazione e tracciamento del lavoro originariamente svolto e, soprattutto, quello di consentire a chiunque di esportare le tabelle, alle medesime condizioni d'uso, verso propri file esterni tramite semplici funzioni applicabili ai dati presenti in pagine web contenenti tabelle o liste, tipicamente utilizzabili nei gestori di fogli di calcolo.

4. Conclusioni e prospettive

Lo sviluppo articolato del progetto di creazione di un wiki per la mappatura consente di affermare che esso sia andato, da diverse prospettive, in crescendo: crescita, professionale e tecnica, dei componenti del gruppo di lavoro, che hanno dovuto imparare approcci e metodologie inconsuete rispetto alle proprie pratiche quotidiane, ma altresì crescita del progetto stesso, inizialmente partito come una canonica attività di mappatura tecnica, benché inedita, dal formato UNIMARC all'ontologia BIBFRAME. Inoltre, in una fase avanzata del lavoro, si è acquisita la consapevolezza del suo potenziale, quello cioè di superare il contesto, definito e circoscritto, nel cui ambito è stato concepito (il catalogo LOD SHARE Catalogue).

La decisione di mettere a disposizione quanto fatto in un ambiente che fosse liberamente accessibile e, completata la modellazione della prima versione stabile della mappatura, collaborativo, con la creazione di account su richiesta all'esterno

[14] Si veda la conversione del sottocampo *318 $u* <https://sharecat-test.wikibase.cloud/wiki/Item:Q33>.

[15] Si veda ad esempio <https://unimarc2bibframe.wikibase.cloud/wiki/Project:Mapping_tables/UNIMARC_Bib_123>.

del gruppo di lavoro, è espressa sia dalla scelta della piattaforma sia dal progressivo processo di arricchimento e collegamento con basi di conoscenza esterne.

Nelle intenzioni del gruppo di lavoro i prossimi passi programmati hanno lo scopo di avvicinare, attraverso il wiki della mappatura, il contesto fortemente localizzato dell'uso di UNIMARC a quelle tendenze globali che richiedono ai dati bibliografici di essere interconnessi e interoperabili, condividendo riflessioni e scelte di modellazione con la comunità professionale che ha la stessa necessità di tradurre informazioni esposte in maniera descrittiva e testuale in istruzioni semanticamente rilevanti e correlate a tutti i contesti in cui esse sono applicabili.

In questa chiave, anche la scelta, più canonica, di condividere nel wiki le tabelle originali della mappatura ha l'intento di rendere distribuito e facilmente accessibile il lavoro, con un approccio positivamente problematico e con la convinzione che procedere verso una rappresentazione in linked data dei dati creati nei formati MARC, indipendentemente dagli strumenti utilizzati, richieda uno sviluppo armonico che non può essere raggiunto tramite una riflessione circoscritta ma che veda in una piattaforma di discussione viva tra i soggetti coinvolti il suo punto di forza creativo.

Del resto, cos'è un wiki (Wikipedia 2025) se non uno spazio collaborativo di discussione, condivisione e crescita comune?

Riferimenti bibliografici[16]

Bergamin, Giovanni, and Cristian Bacchi. 2018. "New Ways of Creating and Sharing Bibliographic Information: An Experiment of Using the Wikibase Data Model for UNIMARC Data." *JLIS.It* 9, 3: 3. https://doi.org/10.4403/jlis.it-12458

Comité français UNIMARC. s.d. "Manuel UNIMARC : format bibliographique." Transition bibliographique - Programme national. <https://www.transition-bibliographique.fr/unimarc/manuel-unimarc-format-bibliographique/> (2025-08-05).

Forziati, Claudio, Annalisa Di Sabato, Rossella Molisso, e Chiara Mugnano. 2024. "La nuova LOD Platform di SHARE Catalogue: un'evoluzione nel segno delle pratiche collaborative della Share Family." In *Parsifal. Un modello di collaborazione bibliotecaria per condividere la conoscenza registrata*, a cura di Silvano Danieli, 105-25. Firenze: Firenze University Press. https://doi.org/10.36253/979-12-215-0356-2.14

IFLA. 2022. "UNIMARC Bibliographic (3rd Ed.) Updates." IFLA. <https://www.ifla.org/g/unimarc-rg/unimarc-bibliographic-3rd-edition-with-updates/>.

Wikipedia. 2025. "Wiki." <https://it.wikipedia.org/w/index.php?title=Wiki&oldid=146124906>.

[16] Ultima consultazione dei siti web: 24 giugno 2026.

La bibliografia di Giacomo Caputo e il ruolo di Wikidata per la sua valorizzazione

Manuela Parrilli

Abstract: The private book collection of Giacomo Caputo (1901–1992), preserved at the Museum and Institute of Prehistory in Florence, offers valuable insight into the intellectual context and scientific production of a key figure in 20th-century Italian archaeology. The project aims to reconstruct and enhance Caputo's complete bibliography, including lesser-known production, such as reviews and methodological notes. Through Wikidata, the bibliography becomes a dynamic, interoperable ecosystem, fostering accessibility and new research paths while proposing a replicable model for the digital valorization of personal funds.

Keywords: private collections, valorization, dissemination, bibliography.

1. Premessa

Negli ultimi decenni, le modalità di valorizzazione degli archivi e delle biblioteche d'autore hanno conosciuto un significativo rinnovamento grazie alla diffusione di piattaforme collaborative e open source che hanno aperto nuove prospettive nella descrizione, accessibilità e interoperabilità dei dati bibliografici. Questi patrimoni documentari e bibliografici, custodi di materiali eterogenei per forma e contenuto e specchio di personalità rilevanti della nostra storia culturale, trovano oggi nuove opportunità di studio e valorizzazione, con ricadute importanti anche sul piano della ricerca. Il presente contributo intende illustrare un'esperienza concreta, sviluppata in seno al progetto di ricerca "ArLiCa - Il fondo archivistico e librario di Giacomo Caputo: archeologia e restauro architettonico in una biblioteca d'autore", finanziato dal Dipartimento di Storia, Archeologia, Geografia, Arte, Spettacolo (SAGAS) dell'Università di Firenze[1]. Il progetto, che affonda le sue radici nello studio della biblioteca personale dell'archeologo Giacomo Caputo, si è posto come obiettivo anche la riscoperta della sua produzione intellettuale, includendo quella parte più nascosta e meno nota

[1] Bando D.R. n. 419 del 02/05/2023 - Approvazione Decreto Rettorale n. 1103/2023 (Prot. n. 0243308 del 13 ottobre 2023). Responsabile del progetto Professoressa Valentina Sonzini.

Manuela Parrilli, manuparrilli@gmail.com, 0009-0002-8995-6641

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Manuela Parrilli, *La bibliografia di Giacomo Caputo e il ruolo di Wikidata per la sua valorizzazione*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.10, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 55-62, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

della sua bibliografia composta da materiale afferente alla cosiddetta letteratura grigia. L'iniziativa si propone come una buona pratica replicabile, atta a restituire nuova vita a patrimoni bibliografici sommersi, coniugando le metodologie tradizionali della biblioteconomia con le potenzialità dei linguaggi strutturati e degli strumenti digitali per la valorizzazione della conoscenza.

2. La biblioteca di Giacomo Caputo: profilo di un archeologo

Giacomo Caputo [Palma di Montechiaro (AG), 1901 – Firenze, 1992] rappresenta una delle figure più autorevoli dell'archeologia italiana del Novecento. Accademico dei Lincei e membro di numerose istituzioni scientifiche italiane e internazionali, Caputo ha rivestito importanti ruoli di responsabilità gestionale e amministrativa operando su un'ampia area geografica, dal Nord Africa all'Etruria. La sua formazione si consolidò attraverso le prime esperienze di scavo nell'isola di Lemno e, successivamente, in Sicilia sotto la direzione del suo maestro Paolo Orsi. A partire dagli anni Trenta, la sua carriera fu profondamente segnata dall'attività in Libia, dove assunse la direzione della Soprintendenza alle Antichità della Cirenaica, succedendo a Giacomo Guidi. I profondi legami con la Cirenaica non si interruppero con il secondo conflitto mondiale ma si rinsaldarono nel dopoguerra, quando Caputo fu nuovamente chiamato dalla British Military Administration a dirigere l'istituzione fino alla nascita del Regno Unito di Libia. L'esperienza maturata nel contesto nordafricano costituì uno snodo cruciale anche per la sua successiva attività in Italia: nel 1951 fu infatti nominato Soprintendente alle Antichità per l'Etruria, la Toscana e l'Umbria. In questo ruolo, Caputo seppe coniugare ricerca scientifica e interessi del vasto pubblico, promuovendo – tra le altre iniziative – una profonda riorganizzazione del Museo archeologico di Firenze, concepita per favorire l'accessibilità e la comprensione del patrimonio anche da parte di un pubblico non specialistico. Libero docente di Archeologia e Storia dell'Arte greca e romana presso l'Università di Firenze dal 1939 al 1952, Caputo non interruppe il proprio impegno scientifico nemmeno dopo il pensionamento, continuando a coordinare le ricerche italiane in Libia attraverso il "Gruppo di Ricerca per le Antichità dell'Africa Settentrionale" (Calloud 2012, 167-71).

Il profilo poliedrico di Caputo trova piena espressione nel fondo personale proveniente dalla sua abitazione e acquisito dall'Università di Firenze nel 2018. Attualmente conservato presso il Museo e Istituto Fiorentino di Preistoria, il *corpus* costituisce un insieme documentario di grande rilievo, composto di più di seimila esemplari tra monografie specializzate, estratti e periodici, oltre che ad una parte del suo archivio[2]. Tale bene, frutto del sedimentarsi della sua atti-

[2] Il materiale archivistico è attualmente conservato in undici scatoloni miscellanei nei quali la famiglia ha collocato i documenti di studio e di lavoro dell'archeologo non riversati presso il Centro di documentazione e ricerca sull'archeologia dell'Africa settentrionale Antonino Di Vita dell'Università di Macerata (dove è conservato il restante della documentazione) <https://cas.unimc.it/>.

vità accademica, della ricerca sul campo e degli interessi del soggetto produttore costituisce, nella sua specificità, un *unicum* eccezionale: lo studio del fondo è in grado di far emergere la figura poliedrica di Caputo e ricostruire il suo laboratorio intellettuale, indagando il suo *modus operandi* in relazione allo studio e all'approccio ai volumi, intercettando gusti personali e canali di approvvigionamento, il tutto messo a sistema con il contesto storico di riferimento. Attraverso l'analisi dei segni d'uso di ogni volume (postille, *ex libris*, dediche, inserti, segnalibri), fonti fondamentali per la ricostruzione del suo profilo di studioso, la biblioteca non solo fotografa l'autore nel suo laboratorio creativo, ma si apre anche all'indagine della fitta rete di contatti accademici, del clima culturale al di fuori delle mura della biblioteca stessa, della partecipazione al dibattito culturale del tempo (Buccino et al. 2025, 232-33).

3. La bibliografia di Giacomo Caputo: metodologia di indagine e trattamento dei dati

Nel progetto di studio del fondo Caputo, la ricostruzione della sua bibliografia completa ha rappresentato uno degli obiettivi principali. L'intento è stato quello di restituire un elenco esaustivo della produzione scientifica dell'archeologo non solo consultando fonti secondarie[3] o repertori bibliografici ufficiali[4], ma utilizzando anche – e soprattutto – i materiali conservati all'interno della sua biblioteca personale. Lo studio analitico di ciascun esemplare che compone il fondo ha infatti consentito di riportare alla luce numerosi contributi di Caputo che risultavano pressoché invisibili alla catalogazione tradizionale in quanto pubblicati su riviste minori, bollettini, atti di convegni, supplementi o opuscoli a circolazione limitata. Si tratta, nella maggior parte dei casi, di quella che viene definita letteratura grigia, pubblicazioni a carattere non convenzionale che, seppur inserite in contesti editoriali autorevoli, sfuggono alle maglie della indicizzazione sistematica dei cataloghi bibliotecari e dei principali *database* bibliografici a causa della loro natura volatile, della tiratura limitata o della diffusione non sistematica che spesso sfrutta canali personali, diversi da quelli che trattano la letteratura convenzionale (Alberani 2002, 325-26). Questa produzione sommersa assume un valore documentario di primaria importanza, poiché consente di cogliere con maggiore completezza l'evoluzione del pensiero scientifico dell'autore nel corso del tempo. Inoltre, diventa un prezioso tassello che contribuisce in modo significativo alla valorizzazione della sua biblioteca personale,

[3] Sono stati consultati i cataloghi delle più importanti biblioteche accademiche e archeologiche italiane e straniere: tra le altre, la Biblioteca di Archeologia e Storia dell'Arte (BiASa), la Biblioteca dell'Istituto Archeologico Germanico di Roma (DAI – Bibliothek), la Biblioteca della Scuola Italiana di Archeologia di Atene), nonché cataloghi commerciali (Amazon, Feltrinelli, Libraccio, Mondadori ecc.), banche dati *open access* ecc.

[4] La voce su Giacomo Caputo compilata da Irene Calloud, nipote di Giacomo Caputo, è stata utilizzata come base di partenza per compilare la bibliografia completa (si veda Calloud 2012, 167-79).

intesa non più solo come luogo di conservazione, ma come spazio dinamico di riflessione, di scrittura e di produzione del sapere.

Per tentare di ricostruire nel modo più completo possibile la produzione intellettuale dell'archeologo, la metodologia adottata si è basata su un approccio sistematico allo studio materiale del fondo. Ciascuna unità bibliografica, infatti, è stata analizzata con attenzione, a partire dall'esame dei segni d'uso – quali annotazioni, sottolineature, postille, ma anche dediche e inserti – considerati non come semplici elementi accessori ma anche come tracce significative delle pratiche di lettura del possessore e potenziali indizi rivelatori degli interessi tematici privilegiati. Tali elementi hanno spesso costituito il punto di partenza per individuare rimandi bibliografici indiretti o riferimenti espliciti a contributi firmati dallo stesso Caputo. Infatti, sebbene l'archeologo mostrasse un rapporto estremamente rispettoso nei confronti dell'oggetto-libro, intervenendo con moderazione sulle pagine, era solito segnalare con discrezione i passaggi che riteneva di particolare interesse, utilizzando piccole freccette, linee leggere o brevi tratti a margine. In alcuni casi, sottolineava l'occorrenza del proprio nome nei testi, un'abitudine che ha facilitato l'individuazione dei contributi: questi segni si distinguono a colpo d'occhio durante la consultazione del volume, permettendo di intercettare più agevolmente i riferimenti bibliografici segnalati e potenzialmente 'sommersi'. Una parte rilevante del lavoro, infatti, è stata l'analisi delle bibliografie e delle note a piè di pagina di ciascun materiale che compone il Fondo, grazie alle quali è stato possibile rintracciare alcuni dei numerosi contributi dell'archeologo non censiti nei repertori ufficiali[5]. Va sottolineato, inoltre, che la presenza di pubblicazioni di Caputo all'interno della biblioteca è sorprendentemente esigua (circa quaranta titoli tra monografie ed estratti sulle oltre seimila unità che compongono il fondo) rendendo necessaria un'attività di ricerca particolarmente meticolosa, tuttora in corso.

I dati raccolti nel corso dell'indagine – sia quelli emersi dall'analisi diretta del fondo, sia quelli individuati attraverso altri canali bibliografici – sono stati progressivamente inseriti in un file di testo, ordinato cronologicamente. In seguito, per ciascun contributo identificato, è stato condotto un accurato processo di verifica incrociata mediante la consultazione dei cataloghi online delle principali biblioteche accademiche e archeologiche, italiane e straniere, oltre che di banche dati e repertori bibliografici specialistici, ricorrendo alla consultazione diretta delle pubblicazioni presso varie sedi bibliotecarie italiane nei casi in cui le fonti digitali restituissero dati incongruenti o dubbiosi.

Nella fase successiva, il *corpus* bibliografico è stato rielaborato in un foglio di calcolo, scomponendo ogni citazione in campi distinti (autore, titolo, tipologia del contributo, rivista ospitante, annata, serie, volume, anno, fascicolo, numero di pagine, editore, luogo e data di pubblicazione) con l'obiettivo di predisporre un *dataset* normalizzato, leggibile tanto da un punto di vista umano quanto computazionale. Tale operazione ha consentito di creare una base dati coerente, destinata a una futura gestione automatizzata e integrabile in sistemi di catalogazione e valorizzazione digitale.

[5] Si rinvia alle fonti già esplicitate nella nota 3.

Una volta predisposti i dati in formato tabellare, è stato possibile procedere ad una fase fondamentale del lavoro: la pulizia e la normalizzazione dei dati raccolti tramite OpenRefine, uno strumento *open source* pensato per gestire, correggere e trasformare *dataset* in modo efficace. Per testare e calibrare le procedure di immissione dei dati in Wikidata, si è scelto di lavorare inizialmente su un sottoinsieme rappresentativo della bibliografia complessiva – composta da circa quattrocento contributi totali – selezionando esclusivamente le quaranta citazioni tratte da *Studi Etruschi*, la rivista sulla quale Giacomo Caputo ha pubblicato con maggiore continuità nel corso della sua carriera. È stato quindi predisposto un ulteriore foglio di calcolo contenente esclusivamente questi contributi, così da seguire con attenzione tutte le fasi del trattamento dei dati, riducendo al minimo il rischio di errori o dispersione derivanti dalla gestione simultanea di un numero troppo elevato di occorrenze.

L'utilizzo di OpenRefine ha permesso di individuare e correggere in modo sistematico e più rapido alcuni errori ricorrenti, varianti nei nomi, incongruenze nei formati, duplicati e altri elementi problematici agendo su larga scala, uniformando le voci e migliorando la qualità complessiva del dataset. Per ogni autore, tipologia del documento, rivista che ospita il contributo ed eventuali editori e luoghi di pubblicazione è stata effettuata la riconciliazione, associando ciascuno elemento a un *item* già esistente in Wikidata, o creandone uno nuovo nei casi in cui l'entità non fosse già presente[6]. Il risultato di questo lavoro è un *dataset* bibliografico coerente, strutturato, verificato e affidabile, pronto per essere integrato in ambienti digitali complessi.

La pulizia e la riconciliazione dei dati non vanno intese come mere operazioni tecniche, ma come un momento di riflessione critica e metodologica sulla natura e l'organizzazione delle informazioni bibliografiche. La possibilità di scomporre, segmentare e strutturare le informazioni consente infatti di restituire un insieme di dati controllati, interrogabili, riutilizzabili e collegabili ad altri *dataset* culturali, superando la dimensione statica delle bibliografie classiche, valorizzando i singoli contributi e gli elementi che li compongono, aprendo nuove strade per la ricerca (Boccone, Rivelli 2019, 236). La preparazione dei dati ha richiesto tempo e cura, ma ha rappresentato un passaggio indispensabile per poter effettuare le operazioni di trasferimento[7] da OpenRefine a Wikidata[8].

[6] Nel campione preso in esame, le tipologie documentarie principali – articolo scientifico, necrologio e recensione – erano già presenti in Wikidata, per cui non si è resa necessaria la creazione di nuove classi. La descrizione dei contributi è stata effettuata tramite le proprietà di base generalmente utilizzate in contesti bibliografici, come istanza di (P31), titolo (P1476), autore (P50), data di pubblicazione (P577), lingua (P407), pubblicato in (P1433), volume di un'opera (P478), pagine (P304), opera disponibile all'indirizzo (P953).

[7] Si ringrazia Camillo Carlo Pellizzari di San Girolamo per il prezioso aiuto fornito in questa delicata fase.

[8] A titolo esemplificativo, si rimanda ad alcuni dei contributi di Giacomo Caputo inseriti in Wikidata: Nuova testa fittile del tempio di Luni (Q134393756); Presentazione del Rilievo di Fiesole antica I (Q134393757); Nuova tomba etrusca a Quinto Fiorentino (Q134393758); Ciro Drago (Q134393759).

La scelta di adottare Wikidata come piattaforma per l'inserimento dei dati bibliografici è stata dettata da una valutazione consapevole delle sue potenzialità in termini di interoperabilità, sostenibilità e diffusione della conoscenza. Wikidata rappresenta oggi uno degli strumenti più avanzati e utili per la gestione di dati aperti e strutturati, grazie a un modello libero, collaborativo e multilingue, che consente la connessione di informazioni eterogenee all'interno di un'infrastruttura globale, leggibili sia da utenti umani che da sistemi automatizzati.

Nel caso della bibliografia di Giacomo Caputo, l'integrazione in Wikidata risponde a una molteplicità di istanze, tra le quali spicca l'esigenza di conferire visibilità e accessibilità a una parte consistente della sua produzione scientifica. Si tratta, come già detto, di quei contributi scientifici non convenzionali pubblicati sotto forma di estratti, articoli in rivista o atti congressuali, materiali spesso marginalizzati nei processi di catalogazione e pertanto difficilmente reperibili, consultabili o riutilizzabili dalla comunità scientifica e che rischierebbero di rimanere esclusi dai circuiti informativi tradizionali. Lo stesso progetto "Il fondo archivistico e librario di Giacomo Caputo: archeologia e restauro architettonico in una biblioteca d'autore" in cui si inserisce la ricostruzione della bibliografia completa dell'autore, a causa di vincoli logistici e temporali prevede esclusivamente la catalogazione delle monografie conservate nel fondo, tralasciando riviste, estratti, periodici e, di conseguenza, una parte significativa della produzione scientifica dell'autore. Inoltre, l'adozione di una piattaforma come Wikidata consente di attivare una modalità di valorizzazione trasversale dei dati, estendendone l'uso ben oltre la funzione descrittiva tradizionalmente associata alla bibliografia. L'organizzazione delle informazioni in formato strutturato e aperto le rende infatti riutilizzabili in contesti molteplici: dalla *data visualization* (tramite grafici, mappe concettuali, cronologie) alla costruzione di percorsi tematici, repertori digitali, strumenti per la didattica o per la divulgazione scientifica. In questo quadro, l'informazione bibliografica non viene più concepita come un semplice oggetto di consultazione, ma come un nodo all'interno di una rete semantica articolata, interrogabile, interconnessa. La natura *machine-readable* dei dati e la loro adesione a standard condivisi a livello internazionale assicurano un elevato grado di interoperabilità, rendendo possibile l'integrazione con altri *dataset* culturali. Questo approccio favorisce la costruzione di strumenti di ricerca in grado di facilitare il confronto tra dati, di integrare progressivamente conoscenze e risultati, di superare i confini disciplinari attraverso l'interazione tra saperi eterogenei. Wikidata si configura dunque non come un'alternativa ai cataloghi bibliotecari tradizionali, ma come un'infrastruttura ad essi parallela e complementare. Se i cataloghi tradizionali garantiscono un accesso puntuale alle risorse conservate, Wikidata consente di rappresentare le relazioni tra opere, autori, luoghi, concetti e istituzioni, superando le rigidità dei formati o delle ontologie chiuse. La sua architettura aperta e collaborativa, che si presta ad essere utilizzata in contesti e ambiti differenti, ne fa uno spazio dinamico per la modellizzazione e la condivisione della conoscenza, pienamente coerente con le istanze teoriche e metodologiche delle *digital humanities* contemporanee (Zhao 2023).

4. Conclusioni

L'esperienza di lavoro sul Fondo di Giacomo Caputo dimostra come la valorizzazione di una biblioteca d'autore possa costituire non solo un'occasione per riflettere sul ruolo della biblioteca come laboratorio creativo di uno studioso e sull'importanza della conservazione e della trasmissione della conoscenza, ma anche un'opportunità concreta per ricostruire e rendere accessibile la sua produzione scientifica. Con il giusto approccio e i giusti strumenti, infatti, bibliografia e biblioteca d'autore non si configurano esclusivamente come insiemi statici di materiali eterogenei scollegati fra loro, ma come mappe intellettuali dinamiche, capaci di restituire l'evoluzione di un pensiero, di illuminare le reti accademiche entro cui esso si è sviluppato e di documentare il dialogo continuo tra ricerca, metodologie e contesto storico[9].

L'analisi materiale dei volumi è stato il passaggio fondamentale, poiché ha permesso di individuare contributi significativi, spesso trascurati o difficilmente accessibili, restituendo così un quadro più completo e attendibile della sua attività intellettuale e mettendo in evidenza quanto della produzione scientifica di un autore possa rimanere invisibile se affidata esclusivamente ai circuiti catalografici tradizionali. Dopo aver raccolto i dati, l'utilizzo di OpenRefine ha reso possibile la costruzione di un dataset bibliografico normalizzato, interoperabile e leggibile sia da esseri umani sia da macchine. La scelta di una piattaforma aperta e collaborativa come Wikidata si è rivelata strategica per garantire una diffusione ampia, sostenibile e duratura dei dati raccolti e per favorire l'interconnessione tra informazioni culturali che spesso rimangono frammentate in compartimenti disciplinari separati.

Il progetto dimostra come l'integrazione tra metodologie tradizionali e strumenti digitali possa non solo restituire visibilità a patrimoni sommersi, ma anche attivare nuove modalità di conoscenza e partecipazione. In questa prospettiva, la valorizzazione del fondo Caputo si configura come un esempio concreto di gestione innovativa e replicabile del patrimonio documentario, capace di coniugare esigenze accademiche, scientifiche e divulgative. L'adozione di soluzioni *open source* e accessibili rende il modello esportabile anche in contesti con risorse limitate, come piccole istituzioni culturali o raccolte private. La qualità del lavoro, infatti, non dipende esclusivamente dalla disponibilità di strumenti tecnologici

[9] Trattandosi di un campione ridotto di citazioni bibliografiche non è stato possibile, in questa fase preliminare, trarre risultati significativi o delineare un quadro complessivo della produzione di Giacomo Caputo. L'analisi ha avuto soprattutto valore sperimentale e metodologico, volta a mettere a punto linee guida operative e testare le procedure di pulizia, riconciliazione e modellazione dei dati bibliografici, in vista dell'inserimento della produzione integrale di Caputo in Wikidata. Nonostante i limiti del campione analizzato, è stato comunque possibile evidenziare la continuità delle sue pubblicazioni sulla rivista *Studi Etruschi* e individuare alcune ricorrenze tematiche e di collaborazione editoriale. Questi dati, seppur parziali, confermano le potenzialità dell'approccio e costituiscono una base solida per futuri sviluppi su un corpus più ampio.

avanzati, ma dalla cura nell'osservazione, nella selezione e nella strutturazione dei dati: un'attività che resta, in ultima analisi, profondamente umana.

Più in generale, il progetto apre alla possibilità di costruire un ecosistema della conoscenza più aperto, trasversale e sostenibile, in cui le biblioteche d'autore non siano più soltanto oggetti di studio, ma risorse vive, interrogabili e integrabili nei flussi della ricerca contemporanea. L'utilizzo di Wikidata come strumento di modellizzazione della conoscenza si fa capace di restituire la complessità dei dati culturali e di attivare nuove relazioni tra testi, autori, istituzioni e luoghi, contribuendo alla costruzione condivisa di una memoria culturale più ricca, accessibile e interconnessa.

La valorizzazione della bibliografia di Giacomo Caputo è solo un punto di partenza; si auspica che altri progetti possano seguirne l'esempio, contribuendo alla creazione di un ecosistema della conoscenza aperto e collaborativo: le potenzialità di questo approccio, infatti, sono ancora in gran parte da esplorare, e meritano di essere approfondite attraverso ulteriori esperienze, in un'ottica di condivisione, riuso e arricchimento collettivo del patrimonio culturale.

Riferimenti bibliografici

Alberani, Vilma. 2002. "La "letteratura grigia" in rete è ancora "letteratura grigia"." *Bollettino AIB* 42, 3: 325-32.

Boccone, Alessandra, e Remo Rivelli. 2019. "I metadati bibliografici in Wikidata: Wikicite e il case study di «Bibliothecae.it»."*.Bibliothecae.it* 8, 1: 235-26.

Buccino, Laura, Manuela Parrilli, Lorenzo Sergi, Valentina Sonzini, e Emanuele Zamperini. 2025. "Dalla Libia a Firenze. Il fondo archivistico e librario di Giacomo Caputo. Un'esperienza di public engagement per la Bright Night 2024 dell'Università degli studi di Firenze." *AIB Studi* 64, 2: 231-47.

Calloud, Irene. 2012. "Giacomo Caputo." In *Dizionario biografico dei soprintendenti archeologi (1904-1974)*, a cura del Ministero per i beni e le attività culturali, Direzione generale per il paesaggio, le belle arti, l'architettura e l'arte contemporanee, e Centro studi per la storia del lavoro e delle comunità territoriali, 167-79. Bologna: Bononia University Press.

Fudie, Zhao. 2023. "A systematic review of Wikidata in Digital Humanities projects." *Digital Scholarship in the Humanities* 38, 2: 852-74. https://doi.org/10.1093/llc/fqac083

ProgrEFR: Wikidata al servizio dei programmi di ricerca dell'École française de Rome

Elena Avellino, Elisa Saltetto

Abstract: ProgrEFR aims to promote the research programmes of the Ecole française de Rome, create links between data containers of a complementary nature (LOD) and make the resources produced by EFR researchers accessible in an open data perspective. We recorded 26 research programmes in Wikidata, IdRef and HAL. An item identifier was created for each programme, allowing interoperability and obtaining different data and information depending on the environments of the various containers. Thanks to its adaptability, Wikidata turned into a hub giving access to all resources produced within the framework of EFR's programmes. IdRef essentially allows a description of the programme as a corporate name heading and HAL collects the relevant studies deposited by individual researchers.

Keywords: École française de Rome, research programs, Linked Open Data, data repository, open science.

1. I programmi di ricerca dell'École française de Rome e gli obiettivi del progetto ProgrEFR

L'École française de Rome è un ente pubblico a carattere scientifico, culturale e professionale posto sotto la tutela del Ministero francese dell'Insegnamento Superiore e della Ricerca (Ministère de l'enseignement supérieur et de la recherche), le cui missioni fondamentali sono la ricerca e la formazione alla ricerca nel campo delle scienze umane, in special modo dell'archeologia e delle discipline storiche, dalla preistoria fino ai nostri giorni, in uno 'spazio geografico' che comprende Roma, l'Italia, il Maghreb e i paesi del Sud-est europeo vicini al mar Adriatico.

L'École è promotrice di una vasta attività scientifica, la quale si esplica non solo attraverso le pubblicazioni (oltre una ventina di volumi all'anno nei diversi specifici ambiti di ricerca e una rivista a cadenza semestrale, i *Mélanges*), ma anche attraverso la partecipazione a diversi tipi di programmi e/o progetti di ricerca: dai progetti europei, ai progetti *Impulsion* e ai progetti finanziati dall'Agenzia nazionale per la ricerca francese (ANR), fino ai progetti finanziati dall'istituzione stessa, ovverosia i 'programmi strutturanti' (*programmes structurants*). Attualmente nella misura di ventisei, sono così definiti data la loro caratteristica

Elena Avellino, École française de Rome, Italy, elena.avellino@efrome.it
Elisa Saltetto, École française de Rome, Italy, elisa.saltetto@efrome.it

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Elena Avellino, Elisa Saltetto, *ProgrEFR: Wikidata al servizio dei programmi di ricerca dell'École française de Rome,* ©Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.11, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 63-68, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

di costituire, per l'appunto, la struttura cardine dell'attività di ricerca dell'École[1]. È compito del Consiglio scientifico dell'istituzione, ogni cinque anni, selezionare le migliori proposte per questo tipo di progetti, sulla base della qualità scientifica e della professionalità dei ricercatori coinvolti (il quinquennio in corso comprende gli anni 2022-2026).

Per quanto riguarda le tematiche oggetto di studio, essi si concentrano sugli ambiti di competenza storicamente propri dell'EFR (archeologia, storia, scienze sociali), entro le aree geografiche pertinenti[2], e prevedono la partecipazione sinergica di varie équipe di ricerca internazionali, in grado non solo di fornire un approccio interdisciplinare alla ricerca, ma anche di dare vita a una produzione scientifica assai eterogenea: articoli, monografie, atti di convegni, presentazioni, risorse prodotte nell'ambito dei seminari e dei cosiddetti 'atelier dottorali' (capaci di confezionare una serie di materiali utili per la didattica universitaria), blog, video, podcast, interviste, documentari.

Va tenuto conto, inoltre, del fatto che, in accordo con le direttive del Ministero della ricerca scientifica francese (Ministère de l'enseignement supérieur et de la recherche 2021), l'EFR aderisce al piano nazionale per l'*open science*[3], per cui diventa essenziale saper promuovere le ricerche della Scuola e, allo stesso tempo, favorire l'accessibilità ai dati prodotti dalle stesse[4], anche ai fini della valutazione da parte di specifici organismi di controllo. Il progetto ProgrEFR si inserisce esattamente in questa logica.

ProgrEFR prende vita dunque dall'attività dei ricercatori 'interni' all'istituzione ma, grazie al carattere dinamico delle loro ricerche, è in grado di oltrepassare i confini dell'École fino a coinvolgere pienamente anche gli specialisti delle discipline o delle tematiche oggetto dei programmi e, naturalmente, i professionisti della documentazione (bibliotecari, documentaristi, archivisti).

La registrazione dei programmi scientifici dell'EFR in Wikidata, 'oggetto principe' di ProgrEFR, ha consentito infatti la creazione di un *hub* capace, all'atto pratico, di agevolare l'accesso al complesso delle risorse prodotte nell'ambito dei programmi, sia per facilitarne il controllo e la valutazione, sia per accelerarne ed ampliarne la fruizione anche da parte di un pubblico più vasto.

2. Metodologia

La semplicità di utilizzo (e fruizione) di Wikidata ha fatto sì che per ogni programma potessero essere agevolmente inseriti durata, ente finanziatore, responsabili di progetto e tematiche di ricerca[5]. La stessa duttilità propria di Wikidata

1 Cfr. École française de Rome s.d.

2 Cfr. École française de Rome. s.d. "Axes scientifiques".

3 Réseau des Écoles françaises de l'étranger s.d. "Données de la recherche"

4 Réseau des Écoles françaises de l'étranger s.d. "Lettre d'information humanités numériques et science ouverte des Ecoles françaises de l'étranger".

5 Lista dei programmi dell'Ecole française de Rome in Wikidata. s.d.

ha fornito, inoltre, una risorsa ulteriore al progetto, ossia la possibilità di inserire metadati che consentono l'accesso diretto ai principali *repository* della ricerca di ambito francofono: da un lato IdRef [6], l'applicazione web che permette la creazione e la consultazione di record di autorità creati dalla rete bibliografica delle università francesi; dall'altro HAL[7], il deposito nazionale delle pubblicazioni della ricerca finanziate dalla Repubblica francese. A tutto ciò si è aggiunta una base pubblica Zotero (cfr. Zotero Public Library s.d.), creata *ad hoc* in biblioteca per ciascun programma, e che costituisce una raccolta di tutte le risorse disponibili sui programmi (siti web, contenuti, riferimenti bibliografici, podcast, locandine dei convegni e degli eventi, etc.).

Va sottolineato, infine, che, con questa procedura, ogni programma di ricerca è stato trattato – ed è 'diventato', di fatto – un 'ente/autore': in un unico 'contenitore', dunque, è ora possibile ritrovare tutte le pubblicazioni e le varie risorse prodotte da diversi autori e/o gruppi di ricerca, con un'estrema capacità di sintesi e una notevole velocità nel raggiungere il risultato della propria ricerca documentaria.

Lo possiamo osservare, ad esempio, con uno dei programmi presi in esame, ovverosia '*Dictamina. Transcription et mise en ligne de sources inédites sur la communication épistolaire italienne médiévale et sur son influence européenne (XIIe-XIVe siècle)*'[8]. Si tratta di un programma della sezione medievale dell'EFR, i cui responsabili provengono da diversi enti e laboratori di ricerca: il Centro di Ricerche Storiche dell'École des Hautes Etudes en Sciences Sociales e del Centro Nazionale per la Ricerca Scientifica e l'École Pratique des Hautes Études. Il progetto prevede di rendere disponibili collezioni inedite di lettere e di trattati teorici prodotti tra il XII e il XIV secolo in Italia, Francia, Spagna e Germania adottando le metodologie di pre-edizione tipiche del metodo filologico classico di impronta tedesca e italiana. Queste metodologie permettono di mettere insieme in tempi brevi delle edizioni non ancora 'critiche' *stricto sensu,* ma di qualità affidabile in quanto frutto di una prima fase di *recensio, collatio* ed *emendatio* di queste fonti poco conosciute e dall'alto potenziale per la ricerca storica. Tali 'pre-edizioni' costituiranno un corpus di trascrizioni che progressivamente saranno rese disponibili sul portale HAL dell'EFR e sul sito web dedicato al progetto.

Per quanto riguarda, più propriamente, il ruolo di ProgrEFR nella valorizzazione di questo progetto, la creazione dell'item 'Dictamina' in Wikidata (Q125959093) consente innanzitutto una definizione concisa quanto essenziale del programma stesso e della sua struttura di massima, rendendolo visibile e interrogabile anche mediante una ricerca in più lingue[9].Vi troviamo,

[6] Lista dei programmi dell'Ecole française de Rome in IdRef. s.d.

[7] Collezione HAL dell'Ecole française de Rome in HAL s.d.

[8] École française de Rome. s.d. "Dictamina. Transcription et mise en ligne de sources inédites sur la communication épistolaire italienne médiévale et sur son influence européenne (XIIe-XIVe siècle)".

[9] Programma Dictamina in Wikidata. s.d.

infatti, tutti i dati essenziali per un inquadramento esaustivo del progetto: tipologia (in questo caso programma strutturante), data di inizio e durata, istituzione principale di riferimento (EFR), responsabili, soggetti di ricerca (tematiche principali), link alla pagina ufficiale del programma. In secondo luogo, come possiamo osservare nella Fig. 1, sono presenti i legami agli identificativi IdRef e HAL che permetteranno di accedere alle risorse versate o recensite in questi database.

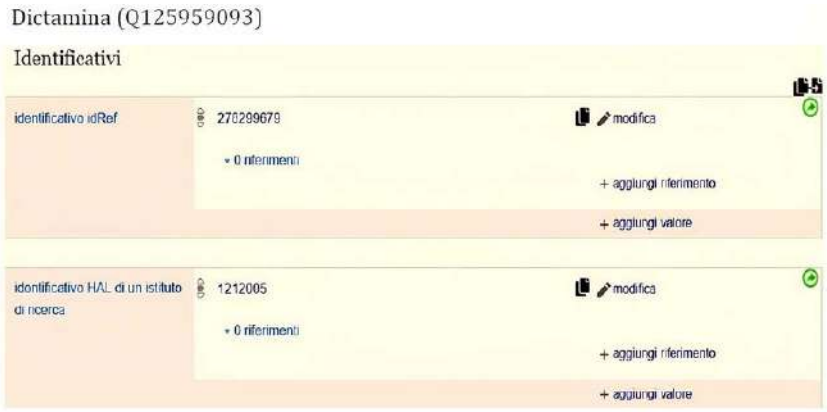


Figura 1 – Wikidata: chiavi di accesso a IdRef e HAL.

Infine, come '*external data available at URL*' nella scheda Wikidata, compare il link alla base di dati pubblica Zotero *Public Dictamina*[10], che raccoglie tutto il materiale prodotto nel contesto del programma, in quanto essa viene via via alimentata in biblioteca seguendo le varie attività dell'équipe di ricerca e le pubblicazioni che ne conseguono, in tutte le loro varie forme (video, blog, documenti a carattere divulgativo, etc.). Un esempio del contenuto di tal genere è costituito dal rinvio al blog del progetto (Dictamina 2022), aggiornato dai responsabili del programma stesso e che raccoglie non solo articoli scientifici, ma anche documenti legati a convegni o ad attività didattiche inerenti al programma quali gli atelier e i seminari dottorali.

L'esempio di Dictamina mostra che, funzionalmente all'obiettivo di diffusione e valorizzazione dei programmi scientifici dell'EFR, l'item Wikidata è perfettamente in grado di descrivere, in maniera essenziale, un progetto di ricerca trattandolo come 'ente-autore' e di racchiudere, in uno stesso luogo, i richiami fondamentali per la sua identificazione in una logica di interoperabilità e di dialogo tra diversi *repository* di dati e, nello stesso tempo, di dare accesso ad ulteriore bibliografia/sitografia di riferimento (cfr. Fig. 2).

[10] Zotero Public Dictamina. s.d.

Figura 2 – Risorse relative ai programmi accessibili tramite item Wikidata.

Di contro, in maniera speculare, situandoci, per esempio, sulla pagina IdRef di un programma (ne offriamo un prototipo ideale in fig. 3), possiamo osservare quali sono le funzionalità offerte da un altro tipo di applicazione (in questo caso volta all'interrogazione, consultazione e creazione di record di autorità). Anche qui viene fornita, in primo luogo, una descrizione del programma; seguono i rimandi alle schede IdRef relative ai responsabili e i link alla bibliografia; infine si elencano gli altri identificativi del programma stesso. Tra questi è assolutamente possibile inserire l'identificativo dell'item Wikidata – elemento che aggiunge un valore oggi fondamentale nella cornice di un'interoperabilità tra i sistemi che sia in grado di superare i confini delle piattaforme di riferimento nazionali.

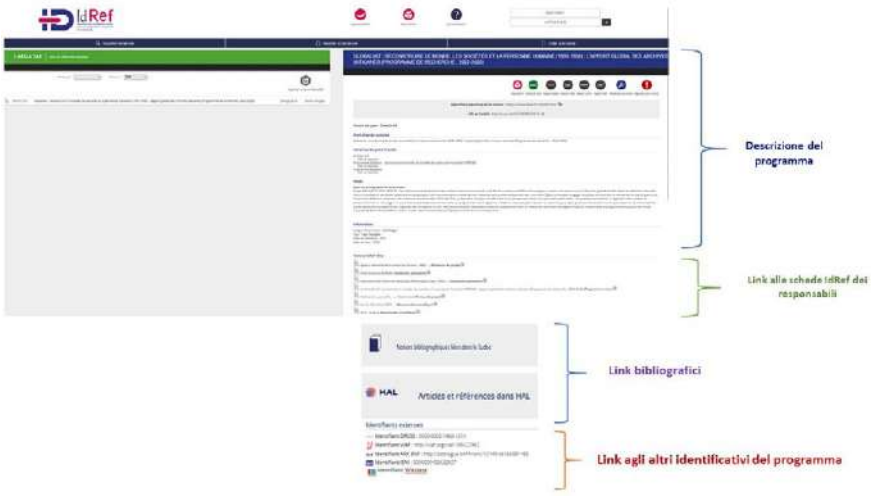


Figura 3 – Prototipo di scheda di programma in IdRef.

3. Conclusioni

ProgrEFR nasce, in un primo tempo, come progetto per descrivere e valorizzare i programmi dell'École française de Rome, ma diventa anche un mezzo utile in sede di valutazione poiché permette di descrivere e quantificare i risultati della ricerca sia di anno in anno che nella fase conclusiva del progetto, per esempio qualora sia necessario elaborare rapporti o relazioni. Esso, infatti, riesce pienamente ad offrire un accesso centralizzato e semplificato a tutti i contenuti che siano stati resi disponibili da un autore o da un ente-autore: dai semplici riferimenti bibliografici fino ai blog, passando per i siti web, i podcast e tutti i tipi di risorse che possano essere state prodotte. ProgrEFR si pone infine nella logica *dell'open science* poiché facilita l'accesso e la reperibilità della produzione scientifica dell'EFR.

Riferimenti bibliografici

Dictamina. 2022. In Hypothèses (blog). <https://dictamina.hypotheses.org/> (2025-04-29).

École française de Rome. s.d. "Axes scientifiques." <https://www.efrome.it/la-recherche/axes-scientifiques> (2025-04-29).

École française de Rome. s.d. "Dictamina. Transcription et mise en ligne de sources inédites sur la communication épistolaire italienne médiévale et sur son influence européenne (XIIe-XIVe siècle)." <https://www.efrome.it/it/la-ricerca/programmi/dettagli-programmi/dictamina> (2025-04-29).

École française de Rome. s.d. "Homepage." <https://www.efrome.it/> (2025-04-29).

HAL. s.d. "EFROME : la collection HAL de l'École française de Rome." <https://shs.hal.science/EFROME> (2025-04-29).

IdRef. s.d. "Lista dei programmi dell'École française de Rome." <https://data.idref.fr/sparql?default-graph-uri=&query=select+distinct+%3Fidref+%3Fpref+where+%7B%3Fidref+a+foaf%3AOrganization%3B+rdfs%3AseeAlso+%3Chttp%3A%2F%2Fwww.idref.fr%2F026436922%2Fid%3E%3B+skos%3AprefLabel+%3Fpref%7D+%23programmi+EFR+%28campo+510%29&format=text%2Fhtml&timeout=0&debug=on&run=+Run+Query+> (2025-04-29).

Ministère de l'enseignement supérieur et de la recherche. 2021 "Le Plan national pour la science ouverte 2021-2024 : vers une généralisation de la science ouverte en France." <https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021-2024-vers-une-generalisation-de-la-science-ouverte-en-48525> (2025-04-29).

Réseau des Écoles françaises de l'étranger. s.d. "Données de la recherche." <https://www.resefe.fr/fr/data> (2025-04-29).

Réseau des Écoles françaises de l'étranger. s.d. "Lettre d'information humanités numériques et science ouverte des Ecoles françaises de l'étranger." <https://www.resefe.fr/fr/newsletters> (2025-04-29).

Wikidata. s.d. "ASRB." <https://w.wiki/Asrb> (2025-04-29).

Wikidata. s.d. "Dictamina." <https://www.wikidata.org/wiki/Q125959093> (2025-04-29).

Zotero Public Dictamina. s.d. "Public Dictamina." <https://www.zotero.org/groups/5723587/public_progrefr/collections/M4HQ57DP> (2025-04-29).

Zotero Public Library. s.d. "Public ProgrEFR." <https://www.zotero.org/groups/5723587/public_progrefr/library> (2025-04-29).

SEZIONE 2

Studi classici / Classical Studies

Authoritative Practices and Collective Validation: Wikidata within the Collaborative Digital Edition of the *Greek Anthology*

Mathilde Verstraete, Maxime Guénette, Marcello Vitali-Rosati

Abstract: As collaborative digital platforms become more common in Cultural and Digital Humanities projects, the role and contributions of non-experts are shifting. Platforms such as Wikidata invite a reconsideration of where authority lies: not solely in the hands of experts, but also within a wider community of contributors who bring diverse forms of expertise. This development challenges hierarchies within academic and cultural institutions and raises new questions about validation, legitimation and collaborative knowledge production. In this paper, we explore these issues in the context of the *Anthologia Graeca* project, where the incorporation of Wikidata has become a way of rethinking editorial authority as a shared and collective process.

Keywords: Greek Anthology, Digital Philology, Authority, Wikidata, Collective Intelligence

1. Introduction

The management and preservation of research data in the Humanities increasingly raise questions concerning its sustainability, accessibility, sharing, and validation. In this context, Wikidata emerges as a powerful collaborative platform. By challenging traditional models in which researchers act simultaneously as producers and gatekeepers of authority, Wikidata reconfigures these issues and fosters new paradigms of collaborative knowledge production.

Within the framework of Digital Humanities [DH], which often emphasises open processes, data interoperability, and collective engagement, Wikidata functions as a knowledge base, enabling a sort of collective verification and semantic linking of information. Unlike traditional academic publishing, where authority can be centralised and restricted to established institutions or recognised experts, Wikidata operates through a model of continuous, multilingual, and community-based editing that promotes the global dissemination of free and accessible knowledge. This paradigmatic shift invites a fundamental rethinking of authority, editorial responsibility, and the epistemological foundations of data.

Mathilde Verstraete, Université de Montréal, Canada, mathilde.verstraete@umontreal.ca, 0000-0003-1642-8610
Maxime Guénette, Université de Montréal, Canada, maxime.guenette@umontreal.ca, 0009-0006-2076-1220
Marcello Vitali-Rosati, Université de Montréal, Canada, marcello.vitali.rosati@umontreal.ca, 0000-0001-6424-3229

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Mathilde Verstraete, Maxime Guénette, Marcello Vitali-Rosati, *Authoritative Practices and Collective Validation: Wikidata within the Collaborative Digital Edition of the Greek Anthology,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.13, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 71-83, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

How, then, can expert-led projects—whether developed by academics, government agencies or GLAM institutions [Galleries, Libraries, Archives and Museums]—work with a generalist platform such as Wikidata to generate new forms of knowledge? How do these hybrid models, which combine scholarly expertise with public participation, challenge traditional boundaries between academic and amateur contributors, and between knowledge production and validation?

In this article, we explore how the infrastructure and logic of Wikidata can be integrated into DH projects, with a focus on the *Anthologia Graeca* [AG] project, a collaborative digital edition of the *Greek Anthology*. We begin by situating Wikidata within the broader landscape of DH initiatives. We then turn to AG as a case study, analysing how Wikidata is embedded in the project's data model—particularly through our treatment of the keyword "authors"—and how this integration opens new spaces for knowledge production. Finally, drawing on those examples, we will reflect on the tensions between authority and collective intelligence that shape such initiatives, and the ways in which digital infrastructures can both challenge and extend traditional scholarly practices.

2. Wikidata and Digital Humanities

Since its creation in 2012, Wikidata has gradually become one of the most significant knowledge graphs on the web. As structured data plays an increasingly central role in the organisation, retrieval, and circulation of information online, Wikidata occupies a pivotal position in shaping how knowledge is represented, accessed, shared, and reused.

As a result, Wikidata has gained popularity within the field of DH. Although there has been—and to some extent, continues to be—skepticism about its quality and potential as a scholarly resource, recent studies have shown that it is now more frequently adopted across DH projects (Cook 2019; Zhao 2023). Within this context, it is primarily used as a content provider—a means to access, publish, or disseminate Linked Open Data [LOD] while avoiding many of the technical and financial constraints traditionally associated with the semantic web.

For institutions in the GLAM sector, Wikidata serves a somewhat different role and is mainly seen as a publishing platform and a tool for metadata curation. Because digitization strategies have historically been developed independently from one institution to another, accessibility and discoverability vary widely in the GLAM domain (Fagerving 2023). Wikidata is thus used either to publish digital identifiers for cultural heritage objects enriched with LOD or to enhance metadata by linking institutional records to Wikidata, improving visibility and interoperability (Candela et al. 2024).

Wikidata also plays a connective role. Its structure as a knowledge graph and its alignment with LOD principles allow it to act as a bridge between otherwise siloed datasets. By assigning persistent identifiers and promoting alignment across vocabularies and standards, Wikidata enables interoperability between different projects and institutions.

3. Wikidata as a linking hub

One of the key benefits of Wikidata, as identified by many DH projects and GLAM institutions, is in its function as a hub for linking heterogeneous external resources and datasets (Neubert 2017). This is achieved through the creation of external identifier properties, which connect entities across different databases to their corresponding Wikidata items.

Take, for example, the case of Megara, an ancient Greek πόλις (*pólis*) located in the northern part of the Isthmus of Corinth. Its Wikidata item (Q42307600)[1] is linked to a wide range of sources, including dictionaries, library catalogues and domain-specific databases on Greco-Roman antiquity such as Pleiades[2], ToposText[3] and MANTO[4]. Through such links, Wikidata facilitates cross-referencing and acts as a form of authority control (Fagerving 2023). As LOD is designed to enrich data with contextual relationships, it becomes much more efficient to enrich a dataset either directly through Wikidata or through cross-queries across databases linked to a common Q-item.

Some scholars have gone even further, suggesting that Wikidata should not only serve as a hub for linking external identifiers, but should function as *the* reference identifier (Van Veen 2019). The proliferation of competing identifiers for the same entity (such as Megara) is not simply redundant: each new authority file introduces another point of potential inconsistency, complicating reconciliation efforts over time. From this perspective, adopting Wikidata as a universal identifier could offer several benefits: a unified description model, a single SPARQL endpoint and API for querying, and a sustainable infrastructure for data access and storage.

While this position may seem more radical, it highlights another important aspect of Wikidata's relationship with external resources: that of reciprocal contribution. As demonstrated in a paper on the interaction between Wikidata and VIAF—a global platform that aggregates name authority files from multiple GLAM institutions—, bi-directional comparison and contribution can improve data quality on both sides (Bianchini, Bargioni, and Pellizzari Di San Girolamo 2021). In this way, such institutions can not only contribute to Wikidata, but also benefit from it.

4. Authority on Digital Platforms

This brings us to a central question: how can academic projects and Wikidata be mutually enriching, given the different systems of validation and recognition that operate within each other? While many Wikidata editors

[1] <https://www.wikidata.org/wiki/Q42307600>. All cited links were last accessed on June 11, 2026.

[2] <https://pleiades.stoa.org/places/570468>.

[3] <https://topostext.org/place/380233PMeg>.

[4] <https://resource.manto.unh.edu/9619397>.

have deep expertise in various disciplines, the platform's structure is built on principles of openness, peer consensus, and collaborative editing rather than formal academic credentials. In contrast, academia tends to associate authority with disciplinary expertise, institutional affiliation, and scholarly output. Similarly, institutions such as governments or GLAM organisations are often recognised as authoritative because of their perceived reliability and structured oversight. These different frameworks can create asymmetries in how contributions are valued, trusted, and legitimised: while academia often operates through top-down models of validation, collaborative platforms like Wikidata rely more on distributed forms of consensus. Understanding and negotiating these differences is key to enhance productive dialogue and collaboration between the two ecosystems.

However, we argue that this top-down hierarchy model is contextual and not always best suited for knowledge dissemination, particularly in the case of digital platforms. As collaborative, crowdsourced or peer-contributed infrastructures like Wikidata increasingly assert themselves as alternative forms of authority—both within and beyond academia—they challenge conventional assumptions about data reliability and the guarantee of expertise. This is mainly because, in most cases, anyone can contribute to these platforms, leading some to question the quality of the data (Santos, Schwabe, and Lifschitz 2024).

The status of Wikidata editors as authoritative actors is therefore often questioned. Many scholars and professionals consider them to be amateurs rather than experts[5]. Despite the increasing involvement of non-professionals in cultural initiatives, especially online, their contributions often remain undervalued or require validation by recognised professionals to be considered reliable. This tension highlights the ambiguous relationship between new actors in knowledge creation and traditional epistemic authorities who are still unsure about how to integrate them.

In Wikidata, this dynamic manifests in conflicting perceptions of authority. While the platform is designed to support collaborative knowledge production and open participation, its data is frequently considered trustworthy only when curated or approved by recognised experts. Yet, recent studies suggest that although Wikidata still has room for improvement, the platform is increasingly being recognised as a high-quality knowledge graph—one whose quality depends on context and must be assessed on a case-by-case basis (Piscopo and Simperl 2019; Shenoy et al. 2022; Zhao 2023). The platform's growing and active community contributes significantly to its data quality, while tools such as Shape Expressions (ShEx) Schemas are being implemented to enforce model conformity

[5] We define the term *amateur* as a socially engaged, non-professional actor who contributes to cultural, artistic or documentary activities outside formal institutional structures—often through participatory digital platforms. This figure is typically characterised by a high degree of self-taught expertise, an individual pursuit of excellence, a paradoxical sociability rooted in both emancipation and community, and a privileged relationship with digital platforms as new spaces of recognition and knowledge production. On the evolution of the term, see Severo (2021).

and internal consistency (Thornton et al. 2019, 2024)[6]. These community-driven standards not only support data consistency but also reflect the platform's commitment to a decentralised yet structured form of knowledge governance.

One of the most effective ways to ensure the quality and relevance of Wikidata is not only to assess it from the outside, but rather to engage directly with it. Through engagement with the platform—making statements, creating data models, correcting mistakes, and discussing ontologies—researchers, cultural institutions, and engaged amateurs alike help to shape the platform and its knowledge graph. Wikidata is not a finished product to be critiqued, but an open, iterative space where quality emerges through collective interaction.

Another way to reconceptualise authority in DH projects is to move beyond using Wikidata as a reference point, and to integrate it as a core infrastructural layer for data modelling, curation, and publication. In this way, the authoritative role traditionally held by academic experts is distributed and shared with a broader community of Wikidata contributors. Authority is thus reframed—not as a fixed attribute derived from institutional status, but as an emergent property of collaborative practices. Such reframing invites researchers to reconsider their own position—not above or outside the platform, but alongside a distributed network of contributors who collectively construct meaning, value, and trust in data.

In order to examine these dynamics more thoroughly, we now turn to our case study: the AG project. We will examine how the project builds its digital infrastructure around Wikidata, collective intelligence and participatory practices to challenge and reshape the status of authoritative figures within academic knowledge production.

5. A collaborative and digital edition of the *Greek Anthology*

Since 2014, Marcello Vitali-Rosati and his team have been developing a digital and collaborative edition of the *Greek Anthology* (Verstraete and Mellet 2024; Verstraete 2024; Vitali-Rosati et al. 2021, 2020; Mellet 2020). The project was born out of the need to index and render accessible this foundational corpus, which gathers nearly all the known epigrams of ancient Greek literature. From the beginning, the project sought both to provide access to these texts and to experiment with a digital editorial model suited to their fragmentary nature and complex transmission history.

Early stages of the project focused on exploring how digital tools could enhance both access to and understanding of that material. Our first prototype—a SPIP-based website—provided a foundation for collaborative enrichment and allowed us to begin identifying the technical and hermeneutic challenges of such an edition.

[6] Shape Expressions (ShEx) are a formal language used to describe and validate the structure of RDF data. In Wikidata, ShEx Schemas help ensure that items conform to expected data models by defining required properties and value types.

These early experiments revealed that designing a digital edition involves more than providing access to texts (Sahle 2016): it requires a coherent epistemological model to organise the relationships among texts, editions, translations, annotations, and contributors. As the project progressed, the platform evolved to support collaborative editing, allowing users to contribute translations, metadata, and commentary: that is when the platform *Anthologia Palatina*[7] was created. Yet, as the corpus expanded and new contributors joined, the limits of this infrastructure—particularly in terms of multilingual support and stable identification of entities—became increasingly apparent.

In response, a third platform was developed—*Anthologia Graeca*[8]. It is augmented by an API and its backend is implemented in Django, a Python-based web framework. Through the contributions of a wide range of collaborators, the platform offers a rich set of information for each epigram. Each of them is presented on a dedicated page, where one can access (1) its location in the *Palatine* manuscript (the *codex Palatinus graecus* 23, the principal testimony for the *Palatine Anthology*), retrieved via the IIIF protocol through the Heidelberg Library's annotation tool linked to its API[9], (2) multiple translations in various languages, (3) keywords covering authors, cities, and other themes, (4) commentaries, (5) internal and external references, and (6) cross-alignments between translations.

This iteration emphasises interoperability and collaboration, aligning with core principles of the Digital Classics community. Partnerships—with initiatives such as Perseus, Perseids, the Heidelberg Library, or the Liceo Classico "Cagnazzi" (Altamura, Bari)—highlight the project's orientation toward open scholarship and cultural heritage valorisation. The editorial model was also refined: the epigram became the fundamental editorial unit, and the platform adopted a semantic web approach grounded in the systematic use of Wikidata identifiers[10]. Wikidata's adoption thus marks a significant turn in the project's editorial framework. It addresses long-standing challenges around entity disambiguation and contributes to embedding each editorial act within a federated,

[7] <https://anthologia.ecrituresnumeriques.ca/home>.

[8] <https://anthologiagraeca.org>.

[9] See Heidelberg, Universitätsbibliothek, *Pal. gr.* 23 (<https://digi.ub.uni-heidelberg.de/diglit/cpgraec23 and not https://digi.ub.uni-heidelberg.de/dig-lit/cpgraec23>). The final folios of the manuscript (615-709) are held at the BNF, under the shelfmark *Paris. Suppl. gr.* 384 (<https://gallica.bnf.fr/ark:/12148/btv1b8470199g/>); they are not yet integrated into the platform.

[10] Data from previous platforms were automatically imported, although not all entities are yet linked to Wikidata—this reconciliation is ongoing, but reveals some epistemological problems. The project is unfinished—and necessarily so: this is not simply due to the usual incompleteness of digital infrastructure, but because of the anthology's very nature, a form that resists closure. Indeed, new readings, translations, annotations, and interconnections will always be possible. While nearly all epigrams are already available on the platform, some remain unfinished, and others await transcription or metadata enrichment. Annotation, translation, and semantic linking—particularly through Wikidata—open a boundless field of contribution. The platform is designed to remain open, both structurally and epistemologically, to future layers of meaning and interpretation.

multilingual knowledge network. This choice also reaffirms the project's original ambition: to develop a truly collaborative and open-ended editorial model.

This case-study focuses specifically on the project's use of keywords—covering entities such as (a) authors, (b) cities, and (c) other keywords collections divided in sections like deities, epithets, epiclesis etc. During the platform's development, it was decided that each keyword must be linked to a Wikidata item. This editorial choice initiated a comprehensive reconciliation of the platform's metadata with Wikidata's knowledge base.

5.1 The Authors of the *Greek Anthology*

From the new platform onwards, adding a Wikidata URI became a requirement for creating any new keyword. This decision, however, left a large portion of previously imported keywords unreconciled—particularly thematic literary motifs, cited personas, and editorial constructs (see *infra*). To address this, we launched a broader reconciliation effort, beginning with the authors, as they appeared to be both the most central and straightforward category to align with Wikidata. This assumption turned out to be only partly correct. While some authors could be easily matched, others presented more complex challenges.

On a methodological point of view, we started by exporting all author names from our platform into a CSV file, which served as both a diagnostic tool and a working basis for reconciliation. We matched names to existing Wikidata elements when possible, flagged ambiguous or missing entries, and created new ones when necessary. For each entry, we aimed to ensure multilingual labels (Latin, French, English, Italian, Ancient Greek). Even at this early stage, we encountered various issues: errors in our database (e.g., Phanias), duplicate or ambiguous entries (e.g., Diodorus) or uncertain attributions for common names (e.g., Dionysios, Archias). We undertook the philological work of identifying and disambiguating these entries and then pushed the revised dataset to Wikidata using OpenRefine.

The response from the Wikidata community was immediate and enlightening. Editors quickly pointed out methodological issues and helped correct errors, as documented in a discussion about our contribution[11]. Some errors highlighted our lack of familiarity with Wikidata conventions, and sparked further discussion about how to prevent such problems (such as preventing the ability to overwrite existing labels and aliases[12]). This process revealed both the strength of the Wikidata community and our own limitations.

Beyond the technical outcome, the process also triggered deeper editorial decisions. Several problematic or overlapping attributions had to be revisited and clarified through philological research. Below, we present a few representative cases that illustrate the nature of this work.

[11] <https://www.wikidata.org/wiki/User_talk:Enrico_malatesta#Epigrammatistes>.

[12] <https://phabricator.wikimedia.org/T157774>.

5.1.1 Phanias

Our dataset listed two authors under the name Phanias: "Phanias" (*Eresius*, TLG-1578) and "Phanias" (*Epigrammaticus*, TLG-1582). According to Gow and Page (1965, 464–75), as well as in the Budé edition, all extant epigrams should be attributed to the latter. The former, while historically attested, was erroneously credited due to editorial confusion. We corrected this by attributing all epigrams to the appropriate Phanias and updating our metadata accordingly.

5.1.2 Diodoros

Our database initially contained four variants of the name Diodoros: "Diodorus, Diodoros", "Diodore de Tarse, Diodoros le Grammairien", "Diodoros Zonas de Sardes", and a lowercase entry, "diodorus". In the Budé edition of the *Greek Anthology* several individuals are recognised under this name:

> *Diodoros*. Trois poètes de la Couronne de Philippe ont porté ce nom : Diodore Zonas, de Sardes, orateur célèbre du temps de la guerre de Mithridate; un autre Diodoros de Sardes; enfin, Diodoros de Tarse. On ne sait auquel attribuer les épigrammes où le gentilicium n'est pas spécifié (Waltz 2003, 142–43).

We reconcile this data with Perseus and kept three distinct *Diodorus* in our database: Diodorus of Sardis, Diodorus Zonas, Diodorus of Tarsus, along with a *catch-all* "Diodoros Epigrammaticus" for unattributed or conflated cases. The label "Epigrammaticus" was adopted (in line with Perseus conventions) for clarity, although it artificially unifies what are likely separate historical individuals. Where the manuscript tradition does not specify a gentilicium or distinguishing epithet, this pragmatic solution provides a stable point of reference.

5.1.3 Dionysios

The case of Dionysios illustrates the difficulties posed by common names in the manuscript tradition. Next to several epigrams, we find the simple attribution "Διονυσίου" (of Dionysios) without further clarification. Gow and Page (1965, 231) note that the name Dionysios is extremely common, and there is no internal evidence to determine how many distinct authors are represented in the *Greek Anthology*. Given this uncertainty, we retained existing individual Dionysios when information allowed (Dionysius of Andros, Dionysius of Cyzicus, Dionysius of Rhodes and Dionysius the Sophist), and introduced a generic "Dionysius" entity for epigrams with unresolved attribution. It is not a perfect solution, and it does not claim to resolve the uncertainty—but it does allow us to surface that uncertainty clearly, rather than pretending it does not exist. In that sense, it follows the same logic we adopted for the Diodoros entries: not trying to solve an unsolvable problem, but giving it space to be visible and documented. This raises the question of "how do we make ambiguity and uncertainty readable, especially in a digital context?"

5.1.4 Archias

Finally, the name Archias presented one of the most tangled cases. Our platform originally grouped all entries under a single name, aggregating multiple forms "Archias, Archias d'Antioche, Archias de Byzance, Archias de Macédoine, Archias de Mytilène, Archias of Macedon, Archias of Mytilene". Perseus lists a general "Archias the Epigrammatist" (TLG-0126). Wikidata likewise has an entry for "Archias the Epigrammatist"[13] which redirects to more specific entities where available. *Aulius Licinius Archias*, the subject of Cicero's *Pro Archia*, is distinct both on Perseus and on Wikidata[14].

The Budé edition adds further complexity, identifying five different individuals under the name Archias, across various volumes:

1) Archias of Antioch, also identified as A. Licinius Archias (Waltz 2003, 141; 1931, 184; 1960, 175; 1974, 281; Irigoin and Maltomini 2017, 69; Aubreton 1980, 302);
2) Archias of Byzantium, attributed to the *Garland of Philip* (Waltz 1960, 175);
3) Archias of Macedon, whose very existence is contested (Waltz 1960, 175);
4) Archias of Mytilene, possibly contemporary with Archias of Antioch (Waltz 1960, 175; 1974, 281);
5) Archias the Younger, a poorly known poet from the *Garland of Philip*, possibly the same as Archias of Antioch (Waltz 1974, 281; Irigoin and Maltomini 2017, 69).

To complicate matters further, Gow and Page, discussing epigrams attributed to "Archias of Byzantium", introduce yet another figure, "Archias Grammaticus": "Archias Grammaticus may be a different person from, or the same person as, the Byzantine; we simply do not know" (Gow and Page 1965, 434–35).

Faced with this level of ambiguity, we adopted a layered model: creating distinct Wikidata entries for historically or textually identifiable individuals and adding disambiguation notes within our platform for uncertain or overlapping cases. This approach allows us to remain consistent with scholarly conventions while aligning with established infrastructures (TLG, Perseus, Wikidata).

5.2 Some comments on the outcome and future work

The reconciliation work between the AG platform and Wikidata—initiated as a metadata cleaning project—has led to substantial improvements across our platform. We clarified and corrected numerous author attributions, inserted inline commentary where attribution remains uncertain (see e.g., *AG* 10.38[15]), created new Wikidata entities for previously unlisted epigrammatists, and contributed multilingual labels.

[13] <https://www.wikidata.org/wiki/Q80207093>.

[14] <https://www.wikidata.org/wiki/Q218251>.

[15] <https://anthologiagraeca.org/passages/urn:cts:greekLit:tlg7000.tlg001.ag:10.38/#comment-1087>.

In undertaking this process, we have not only deepened the integration between philological expertise and semantic web technologies but also brought to light a series of epistemological frictions and modeling challenges—revealing the extent to which editorial categories, rooted in interpretive scholarship, can resist or complicate their translation into structured, ontological frameworks like Wikidata. A particularly revealing case involved the labeling of pseudo-authors in Ancient Greek (see our discussion on Pseudo-Lucian[16]). Indeed, in January 2023, a dialogue unfolded regarding whether pseudo-authors in Wikidata should be labeled in Ancient Greek, and if so, how. One of our imported entries—initially labeled in Ancient Greek—was removed due to concerns about accuracy and conceptual anachronism. The crux of the debate lay in whether the prefix Ψευδο- (as used in Modern Greek for pseudo-authors) should be applied to Ancient Greek labels. Some contributors noted that this construction appears rarely in ancient sources and argued that its use would introduce modern editorial assumptions into ancient naming practices.

Finally, further work must be done in order to reconcile other keywords. However, the reconciliation process is revealing several limitations and epistemological tensions between the platform's existing editorial categories and the structure of Wikidata. Some keywords previously encoded in the AG platform do not align easily—or meaningfully—with Wikidata's data model. For instance, certain categories were created specifically for earlier editorial experiments, namely the *Plateforme Ouverte des Parcours d'imaginaires* [POP] for which we added *reading path*-keywords such as *Animal's epitaph* (on the AG platform[17] and on the POP[18]) or *Deadly gifts* (on the AG platform[19] and on the POP[20]). Whether we should include these in Wikidata is still unclear.

Other keywords correspond to broader or more abstract concepts—such as literary motifs or narrative structures—that resist straightforward alignment with Wikidata entities (e.g. *bad breath* or *drunkenness*). In the case of cited individuals, the situation is even more complex: the *Greek Anthology* includes references to numerous figures whose historical existence is uncertain or contested. Reconciling these names would demand a sustained philological and prosopographical effort, as some may be fictional, mythological, or only attested within this corpus see Floridi (2021) for broader epistemological considerations). In other cases, reconciliation is both feasible and necessary but requires careful, manual work—especially for editorial categories such as the various epigrams collections (see, by example, the Garlands of Meleager[21] or Philip[22]) or cited po-

[16] <https://www.wikidata.org/wiki/Talk:Q19558843#Nome_in_grc>.
[17] <https://anthologiagraeca.org/keywords/232/>.
[18] <https://pop.anthologiegrecque.org/#/parcours/308>.
[19] <https://anthologiagraeca.org/keywords/434/>.
[20] <https://pop.anthologiegrecque.org/#/parcours/300>.
[21] <https://anthologiagraeca.org/keywords/181/>.
[22] <https://anthologiagraeca.org/keywords/76/>.

ets. Overall, while the integration of Wikidata represents a significant step toward interoperability, the reconciliation of legacy data remains an ongoing process that highlights the complexity of mapping scientific knowledge to structured, external ontologies—especially when retroactively aligning two editorial infrastructures that were not originally designed to work together.

6. Conclusion

Wikidata is not just a technical tool for data storage and retrieval; it is a dynamic epistemic space where academic knowledge is collaboratively shaped, revised, and legitimised. By delegating (while taking part of) some of our curatorial authority to Wikidata, the AG project participates in a broader shift towards distributed models of knowledge production (Benkler, Shaw, and Mako Hill 2015; Bücheler et al. 2010). This delegation does not imply an abandonment of academic rigor, but rather a reconfiguration of expertise—one that recognises the value of collective intelligence (Lévy 1994) and the potential of community-driven platforms to maintain high-quality, multilingual, and interoperable data (Surowiecki 2004).

The implications of this shift are manifold. It invites researchers to rethink their role not only as producers of content, but also as facilitators of open, dialogical knowledge systems. It also raises critical questions: To what extent can academic standards be reconciled with the epistemological norms of a platform like Wikidata? How do we balance openness with accuracy, and participation with control? Far from erasing these tensions, the integration of Wikidata into our project foregrounds them—and in doing so, creates an opportunity for scholars to collectively reflect on the processes of validation, authorship, and authority.

Ultimately, we suggest that embracing platforms like Wikidata means accepting that academic knowledge is always provisional, situated, and co-produced. It is a move toward a more resilient and plural form of scholarship—one that recognises the potential of the many without losing sight of the responsibility of the few.

References

Aubreton, Robert. 1980. *Anthologie Grecque. Tome XIII: Anthologie de Planude*. Paris: Les Belles Lettres.

Benkler, Yochai, Aaron Shaw, and Benjamin Mako Hill. 2015. "Peer Production: A Form of Collective Intelligence." In *Handbook of Collective Intelligence*, edited by Thomas Malone, and Michael Bernstein, 175–204. Cambridge, Massachusetts: MIT Press.

Bianchini, Carlo, Stefano Bargioni, and Camillo Carlo Pellizzari Di San Girolamo. 2021. "Beyond VIAF: Wikidata as a Complementary Tool for Authority Control in Libraries." *Information Technology and Libraries* 40, 2. https://doi.org/10.6017/ital.v40i2.12959

Bücheler, Thierry, Rudolf M. Füchslin, Rolf Pfeifer, and Jan Henrik Sieg. 2010. "Crowdsourcing, Open Innovation and Collective Intelligence in the Scientific Method: A Research Agenda and Operational Framework." In *Artificial Life XII – Twelfth International Conference on the Synthesis and Simulation of Living Systems,*

Odense, Denmark, 19 August 2010 - 23 August 2010, 679–86. Odense: Danemark. https://doi.org/10.5167/uzh-42435

Candela, Gustavo, Mirjam Cuper, Olga Holownia, Nele Gabriëls, Milena Dobreva, and Mahendra Mahey. 2024. "A Systematic Review of Wikidata in GLAM Institutions: A Labs Approach." In *Linking Theory and Practice of Digital Libraries*, edited by Apostolos Antonacopoulos, Annika Hinze, Benjamin Piwowarski, Mickaël Coustaty, Giorgio Maria Di Nunzio, Francesco Gelati, and Nicholas Vanderschantz, 34–50. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72440-4_4

Cook, Stacey. 2019. "The Uses of Wikidata for Galleries, Libraries, Archives and Museums and Its Place in the Digital Humanities." *Comma* 2017, 2: 117–24. https://doi.org/10.3828/comma.2017.2.12

Fagerving, Alicia. 2023. "Wikidata for Authority Control: Sharing Museum Knowledge with the World." *Digital Humanities in the Nordic and Baltic Countries Publications* 5, 1: 222–39. https://doi.org/10.5617/dhnbpub.10665

Floridi, Lucia. 2021. "Speaking Names, Variant Readings, and Textual Revision in Greek Epigrams." *Incontri Di Filologia Classica* 19: 135–56. https://doi.org/10.13137/2464-8760/32054

Gow, Andrew Sydenham Farrar, and Denys Lionel Page. 1965. *The Greek Anthology. Hellenistic Epigrams*, vol. 2. Cambridge: Cambridge University Press.

Irigoin, Jean, et Francesca Maltomini. 2017. *Anthologie Grecque. Tome IX: Anthologie Palatine, Livre X*, 13 vols. Paris: Les Belles Lettres.

Lévy, Pierre. 1994. *L'intelligence collective : pour une anthropologie du cyberspace*. Paris: La Découverte.

Mellet, Margot. 2020. "Penser le palimpseste numérique. Le projet d'édition numérique collaborative de l'Anthologie palatine." *Captures* 5, 1. https://doi.org/10.7202/1073479ar

Neubert, Joachim. 2017. "Wikidata as a Linking Hub for Knowledge Organization Systems? Integrating an Authority Mapping into Wikidata and Learning Lessons for KOS Mappings." In *Proceedings of the 17th European Networked Knowledge Organization Systems Workshop Co-Located with the 21st International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017), Thessaloniki, Greece, September 21st, 2017*, edited by Philipp Mayr, Douglas Tudhope, Koraljka Golub, Christian Wartena, and Ernesto William De Luca, 1937: 14–25. CEUR Workshop Proceedings. CEUR-WS.org.

Piscopo, Alessandro, and Elena Simperl. 2019. "What We Talk about When We Talk about Wikidata Quality: A Literature Survey." In *Proceedings of the 15th International Symposium on Open Collaboration*, 1–11. Skövde Sweden: ACM. https://doi.org/10.1145/3306446.3340822

Sahle, Patrick. 2016. "2. What Is a Scholarly Digital Edition?" In *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll, and Elena Pierazzo, 19–39. Cambridge: Open Book Publishers (Digital Humanities Series). https://books.openedition.org/obp/3397

Santos, Veronica, Daniel Schwabe, and Sérgio Lifschitz. 2024. "Can You Trust Wikidata?" *Semantic Web* 15, 6: 2271–92. https://doi.org/10.3233/SW-243577

Severo, Marta. 2021. *L'impératif participatif: Institutions culturelles, amateurs et plateformes*. Bry-sur-Marne: Institut National de l'Audiovisuel (INA). https://doi.org/10.3917/ina.sever.2021.01

Shenoy, Kartik, Filip Ilievski, Daniel Garijo, Daniel Schwabe, and Pedro Szekely. 2022. "A Study of the Quality of Wikidata." *Journal of Web Semantics* 72, 100679. https://doi.org/10.1016/j.websem.2021.100679

Surowiecki, James. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. New York: Doubleday.

Thornton, Katherine, Harold Solbrig, Gregory S. Stupp, Jose Emilio Labra Gayo, Daniel Mietchen, Eric Prud'hommeaux, and Andra Waagmeester. 2019. "Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation." In *The Semantic Web*, edited by Pascal Hitzler, Miriam Fernández, Krzysztof Janowicz, Amrapali Zaveri, Alasdair J. G. Gray, Vanessa Lopez, Armin Haller, and Karl Hammar, 11503: 606–20. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-21348-0_39

Thornton, Katherine, Kenneth Seals Nutt, and Anne Chen. 2024. "Encoding Archaeological Data Models as Wikidata Schemas: Utilizing Shape Expressions to Structure Collaborative Linked Open Data for Digital Storytelling Within the International Dura-Europos Archive." *The International Journal of Technology, Knowledge, and Society* 21, 1: 69–83. https://doi.org/10.18848/1832-3669/CGP/v21i01/69-83

Van Veen, Theo. 2019. "Wikidata: From 'an' Identifier to 'the' Identifier." *Information Technology and Libraries* 38, 2: 72–81. https://doi.org/10.6017/ital.v38i2.10886

Verstraete, Mathilde. 2024. "Dynamiques auctoriales au sein du projet d'édition numérique collaborative de l'*Anthologie grecque*." *Sens Public*. https://doi.org/10.7202/1118945ar

Verstraete, Mathilde, and Margot Mellet. 2024. "Passés et présents anthologiques. Le projet d'édition numérique collaborative de l'Anthologie grecque." In *Communautés et pratiques d'écritures des patrimoines et des mémoires*, edited by Nicolas Sauret, and Marta Severo, 67–80. Paris: Presses universitaires de Paris-Nanterre.

Vitali-Rosati, Marcello, Servanne Monjour, Joana Casenave, Elsa Bouchard, and Margot Mellet. 2020. "Editorializing the Greek Anthology: The Palatin Manuscript as a Collective Imaginary." *Digital Humanities Quarterly* 14, 1. https://doi.org/10.63744/6tqtxmxjq3ej

Vitali-Rosati, Marcello, Margot Mellet, Servanne Monjour, Antoine Fauchié, Timothée Guicherd, David Larlet, and Enrico Agostini-Marchese. 2021. "L'épopée numérique de l'Anthologie grecque : entre questions épistémologiques, modèles techniques et dynamiques collaboratives." *Sens public*. https://doi.org/10.7202/1089649ar

Waltz, Pierre. 1931. *Anthologie Grecque. Tome III: Anthologie Palatine, Livre VI*, 13 vols. Paris: Les Belles Lettres.

Waltz, Pierre. 1960. *Anthologie Grecque. Tome V: Anthologie Palatine, Livre VII, Épigrammes 364-748*, 13 vols. Paris: Les Belles Lettres.

Waltz, Pierre. 1974. *Anthologie Grecque. Tome VIII: Anthologie Palatine, Livre IX, Épigrammes 359-827*, 13 vols. Paris: Les Belles Lettres.

Waltz, Pierre. 2003. *Anthologie Grecque. Tome II: Anthologie Palatine, Livre V*, 13 vols. Paris: Les Belles Lettres.

Zhao, Fudie. 2023. "A Systematic Review of Wikidata in Digital Humanities Projects." *Digital Scholarship in the Humanities* 38, 2: 852–74. https://doi.org/10.1093/llc/fqac083

Collecting and Detecting Ancient Greek Historians through Wikibase and Wikidata[1]

Leonardo D'Addario

Abstract: This paper presents the new Wikibase instance, *The Library of Polybius*, whose aim is to systematically collect and analyse data and metadata of the citations of ancient Greek writers, namely "lost" historians, in *The Histories of Polybius*</corsivo>. The database includes the original Greek text of the citations, metadata about the quoted authors (e.g., name, provenance, period) and their works (e.g., title, number of books, content), as well as references to the relevant sections of *The Histories* (book, chapter, paragraph). It also records the linguistic elements of the citations (e.g., the names of the quoted authors, nouns referring to their works, and verbs introducing citations). Once completed, the Wikibase will be queried through the SPARQL Query Service in order to shed light on Polybius' citation practice.

Keywords: ancient greek historiography, polybius, wikibase, digital classics.

1. Introduction

This paper explores the potential of Wikibase and Wikidata for cataloguing data and metadata on Ancient Greek historians. This study is part of my PhD research in Ancient Greek History at Leipzig University, conducted under the supervision of Monica Berti (Leipzig University) and Stefan Schorn (KU Leuven) within the framework of MECANO,[2] a doctoral network funded by the Horizon Europe program, which investigates the dynamics of canonization in Graeco-Roman texts. In particular, my research focuses on *The Histories* of Polybius.

The Histories of Polybius (206–124 BCE, approx.) originally consisted of 40 books, covering the period from 264 BCE (the outbreak of the First Punic War) to 146 BCE (the destruction of Carthage and Corinth) and documenting the rise of Rome as the main political power in the Mediterranean world. Unfortunately, while only the first five books of his work are fully preserved, the others are known only through quotations in later sources and collections of excerpts

[1] I would like to thank Camillo Carlo Pellizzari di San Gerolamo, who first introduced me to Wikidata and Wikibase and taught me about their usefulness for Classical studies. I also extend my thanks to my supervisor at Leipzig University, PD Dr. Monica Berti, for reading this work and providing many insightful suggestions. Any errors are obviously my own.

[2] <https://mecano-dn.eu/>.

Leonardo D'Addario, Leipzig University, Germany, leonardo.daddario@uni-leipzig.de, 0009-0004-5271-3878

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Leonardo D'Addario, *Collecting and Detecting Ancient Greek Historians through Wikibase and Wikidata*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.14, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 85-97, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

created during the Byzantine Empire. Despite this lacuna, Polybius' work is fundamental for Classical studies, as it is the only surviving example of Hellenistic historiography and the primary source for many of the so-called "lost historians."[3]

The expression "lost historians" refers to ancient historiographers whose works are lost and are now known only through verbatim quotations, paraphrases, and allusions in later sources. Following Schepens' (1997, 166–7 note 66) influential term, I will call sources quoting earlier writers "cover-texts."[4] Passages containing citations are often extracted from the cover-texts, gathered into collections, and studied as "fragments" (i.e., surviving parts of a lost whole). This approach underpins one of the greatest philological works of the last century: Felix Jacoby's *Die Fragmente der Griechischen Historiker*, the authoritative collection of fragments of the lost historiographers. However, as Hau (2024, 151–52) emphasises, it is crucial, when dealing with fragments, to take into account their context of transmission, since one cannot be certain how much the cover-text altered the quoted originals or whether the quotation is verbatim or a paraphrase. Hence, the need to study the quoting and bibliographic practices of the cover-texts.

Given Polybius' status as a source for many lost historiographers, my PhD project, "Detecting and Retrieving Lost Historians," examines *The Histories* as a cover-text. The main goals are: (1) to analyse the language that Polybius uses when citing earlier writers in order to provide insights into his quoting practice and to assess his reliability as a source for fragments, and (2) to establish whether Polybius' quoting practice reflects engagement with a specific canon of authoritative writers.

The first step of the project involves cataloguing all passages where Polybius quotes other authors and collecting information about them. This catalogue must include:

- The original Greek text of the quoting passages.
- Information about the quoted authors (e.g., name, provenance, period) and their works (e.g., title, number of books, content).
- The relevant linguistic elements of the quoting text–namely, the Greek forms of the authors' names and their works, as well as verbs introducing citations.

Although this research focuses on lost historiographers, authors from other genres whose works are cited by Polybius will be included in the collection. The analysis of the final catalogue may indeed reveal that Polybius' quoting habit

[3] According to the appendix of the Teubner edition, Polybius quotes forty-eight writers, among whom twenty-one are historiographers.

[4] Schepens (1997, 167 note 66) employs the term "cover-text" due to the triple meaning of the verb "to cover"— "to protect", "to conceal", and "to enclose." The cover-text functions in three interrelated ways: it preserves quotations of lost works (thereby "protecting" them from total oblivion); it conceals their original form, as most stylistic and contextual features of the quoted text are altered; and it encloses them within a new narrative framework, recontextualizing the quoted text to serve the aims of the quoting author.

varies based on literary genres. Citations of preserved authors will also be included, as the comparison of the quoted text with its original may help evaluate Polybius' accuracy and determine whether he uses specific verbs or expressions to quote verbatim or paraphrase.

As MECANO fosters the integration of traditional approaches with new digital methods, I am currently developing a Wikibase instance called *The Library of Polybius*[5] in order to collect the quoting passages of *The Histories* and their metadata according to the principles stated above. In this paper, I will discuss the data model I designed to describe each quoting passage through the statements system.

While many digital practices are now being developed for studying the indirect transmission of classical authors and works (cfr. Berti 2021; 2024), Wikibase's application for this specific research field remains unexplored. However, Wikibase offers an ideal digital environment for describing and analysing ancient Greek passages and their linguistic features.

Firstly, it allows for structuring data and metadata based on specific research questions. Through the statements system, one can describe each item by attributing to it the characteristics that need to be analysed. For instance, I designed a data model including statements specifying the relevant linguistic elements of each quotation, which are the very object of my research.

Secondly, the SPARQL Query Service allows for a fast and efficient analysis of data. In the specific case of *The Histories*, queries such as "quotations where Polybius uses the verb ἱστορέω" or "quotations where Polybius specifies the title of works" can help analyse Polybius' quoting practice (or *ratio laudandi*, as classical scholars usually call it).

Lastly, Wikibase allows for the creation of datasets which will remain available and reusable for future researchers according to the principles of Linked Open Data (LOD).[6] In fact, cataloguing quoting passages is a typically philological task, and most scholars would do it manually and then present only the outcome of their analysis of the catalogue. However, such catalogues usually remain unpublished, meaning other scholars can only reuse the research results presented in an academic work.

2. Data model for quoting passages

The following table (Fig. 1) shows the data model for quoting passages. This data model has been designed based on that of the Wikibase instance *Hypotheseis*, created by Camillo Carlo Pellizzari di San Gerolamo.[7] Each row represents the property-value pair. The first column specifies whether the pair

[5] <https://library-of-polybius.wikibase.cloud/wiki/Main_Page>. All links cited in this work were last accessed on 31 August 2025.

[6] On *Linked Open Data* (LOD) in Ancient World studies, see Cayless 2019.

[7] <https://hypotheseis.wikibase.cloud/wiki/Pagina_principale>.

is necessary or not. The central column includes the properties of each pair. The symbol [Q] means that the property-value pair is a qualifier. The symbol [R] means that the property-value pair is a reference. The third column includes the values of each pair. In the round brackets, the data type of each property is specified.

Necessity	Properties	Values
yes	instance of	quoting passage (item)
yes	quoting text	Greek text (string)
yes	[Q] transcribed from the expression	expression (item)
yes	[R] CTS URN	CTS URN (external identifier)
yes	[R] URL of the transcription	URL to Perseus Scafe Viewer (url)
yes	taken from	work (item)
yes	[Q] citation	citation (string)
yes	quoted author(s)	person (item)
yes	[Q] textual form	Greek text (string)
yes (if "textual form" has a value)	[Q] lemma(s)	lemma (string)
yes (if "textual form" has a value)	[Q] reference form	proper name / ethnic / generic noun (item)
yes	[R] determination method	explicit textual reference / inferred from the context (item)
yes (if "textual form" has a value)	[R] lemma reference(s)	*Logeion* URL (url)
yes	quoted work(s)	work (item)
yes (if there is the previous pair)	[Q] textual form	Greek text (string)
yes (if "textual form" has a value)	[Q] lemma(s)	lemma (string)
yes (if "textual form" has a value)	[Q] reference form	name / section (item)
yes	[R] determination method	explicit textual reference / inferred from the context (item)
yes (if "textual form" has a value)	[R] lemma reference(s)	*logeion* URL (url)
no	reporting verb(s)	lemma (string)
yes (if there is the statement)	[Q] textual form	Greek text (string)
yes (if there is the statement)	[Q] subject	person (item)
yes (if there is the statement)	[R] lemma reference(s)	*logeion* URL (url)

Figure 1 – Data model.

In order to discuss the data model in detail, a concrete example—notably, Q34 "Polyb. IV 20, 5"[8] —will be provided. In this passage, Polybius quotes the historian Ephorus and alludes to the content of the proem of his historiographi-

[8] <https://library-of-polybius.wikibase.cloud/wiki/Item:Q34>.

cal work: οὐ γὰρ ἡγητέον μουσικήν, ὡς Ἔφορός φησιν ἐν τῷ προοιμίῳ τῆς ὅλης πραγματείας, οὐδαμῶς ἁρμόζοντα λόγον αὐτῷ ῥίψας, ἐπ᾽ ἀπάτῃ καὶ γοητείᾳ παρεισῆχθαι τοῖς ἀνθρώποις· ("For we must not suppose, as Ephorus, in the Preface to his History, making a hasty assertion quite unworthy of him, says, that music was introduced by men merely for the purpose of beguiling and bewitching.")[9] In the following paragraphs, I will show and explain the statements describing the item, specifying their relevance to my research questions and providing guidelines for their correct use.

2.1 Quoting text, expression, and work

Let us look at the first three statement describing Q34 "Polyb. IV 20, 5" (Fig. 2). The item Q34 is an instance of Q3 "quoting passage", which is an entity. As shown in the data model (Fig. 1), the data type of P7 "quoting text" is a string. Therefore, this property has the ancient Greek text of the quoting passage as its value. The qualifier of the second statement, P8 "transcribed from the expression," specifies the critical edition from which the ancient Greek text has been transcribed. This is notably the Teubner edition by Büttner-Wobst, which is openly available as an XML file on the GitHub repository of the Perseus Digital Library.[10] The ancient Greek text has been copied from this XML file and pasted as the value of P7 "quoting text". The first reference provides the URL to retrieve the exact passage of *The Histories* in the Perseus Digital Library Scaife Viewer, which is a digital reading environment for text collections.[11] The second reference specifies the CTS URN of the quoting passage (urn:cts:greekLit:tlg0543.tlg001.perseus-grc2:4.20.5), a standardized identifier for texts in digital environments designed to uniquely and persistently reference specific passages.[12]

In P8 "transcribed from the expression", *expression* denotes Polybius' Greek text as edited by Büttner-Wobst. The term *expression* is employed according to the IFLA LRM model.[13] This conceptual entity-relationship model specifically distinguishes *expression* and *work* as different entities.[14] *Work* is defined as an "intellectual or artistic content of a distinct creation"—in our case, the abstract and intangible idea of *The Histories*, which exists independently of language, form, or physical format. On the other hand, *expression* is defined as "a distinct

[9] All translations of Polybius' text are from the Loeb edition.

[10] <https://github.com/PerseusDL/canonical-greekLit/tree/master/data/tlg0543/tlg001>.

[11] <https://scaife.perseus.org/>.

[12] On the use of CTS URNs for citing and retrieving Classical sources, see Babeu 2019, and Backwell, Christopher, and Smith 2019. The URN "urn:cts:greekLit:tlg0543.tlg001.perseus-grc2:4.20.5" follows the Canonical Text Services standard ("cts"), where "tlg0543" identifies Polybius, "tlg001" refers to *The Histories*, "perseus-grc2" denotes the specific edition used—the Perseus XML file of the Büttner-Wobst's Teubner edition, and 4.20.5 references Book 4, Chapter 20, Section 5.

[13] As for the IFLA LRM model, see Riva, Le Boeuf, and Žumer 2017.

[14] I use the italic when I refer to *work* and *expression* as entities of the IFLA LRM model.

combination of signs conveying intellectual and artistic content"–here, the specific version of Polybius' text as edited by the Teubner editors.[15] Consequently, while Q5 "Polybii Historiae (ed. Büttner-Wobst)" is an instance of *expression*, the value of the third statement, Q8 "Πολυβίου Ἱστορίαι," is an instance of *work*. The original ancient Greek title is used here to preserve the historical identity of the *work*, as Polybius conceived the idea of *The Histories* in ancient Greek. Additionally, the original Greek title ensures disambiguation, as translated titles are assigned to *expressions*.[16]

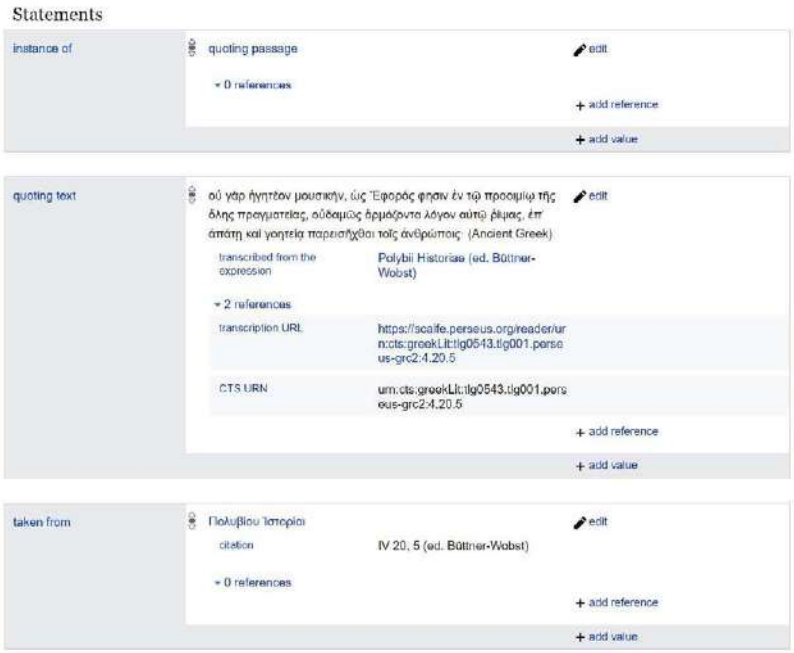


Figure 2 – First three statements of Q34.

The third statement ("taken from – Πολυβίου Ἱστορίαι") specifies the source of the quoting passage. P11 "taken from" has the *work* as its value, because the quoting passage is a part of the intellectual and artistic content of *The Histories*. The statement also includes a qualifier specifying in which section of *The Histories* the quoting passage is found. The qualifier's value, "IV 20,

[15] The definitions of *expression* and *work* are taken from Riva, Le Boeuf, and Žumer 2017, 21–2.

[16] Because the original titles of ancient works were not fixed, the ancient Greek titles used in this study align with those assigned by Jacoby in *Die Fragmente der Griechischen Historiker* and, for authors who are not lost historians, the the ones provided in the standard critical editions. Since works by different authors may share the same title (e.g., Ἱστορίαι), the genitive form of the author's name is occasionally used for disambiguation.

5 (ed. Büttner-Wobst)," refers to the *expression*. As a matter of fact, although the book division of *The Histories* could be considered as a characteristic of the *work* (since Polybius himself conceived it), the paragraph and subparagraph numbering belongs to the *expression*—specifically, the Teubner edition's textual configuration.

Before moving to the next section, a clarification about the criteria for selecting the quoting passage is required. In general, a quoting passage is defined as a section of *The Histories* where Polybius refers to an author and their work. This implies that all passages naming a specific author but not referring to the content of their work (e.g., anecdote about the author's life) will be excluded from the Wikibase. Fortunately, this almost never happens in Polybius, who cites other authors just to express his—mostly critical—judgement about their literary and historical value. However, there are some exceptions, such as Aratus of Sicyon. He wrote his *Memoirs*, a fundamental source for Polybius. At the same time, he was also a politician of the Achaean Legue, and this is why Polybius mentions him multiple times. Therefore, *The Library of Polybius* will include only passages referring to Aratus of Sicyon as the writer of the *Memoirs*.

Additionally, since the Wikibase allows for a maximum length of four hundred characters for a string, other principles are strictly observed to delimit the quoting passage. Notably, the quoting passage will consist of a meaningful sentence and stop at the period or high dot. It is not important if Polybius continues describing the content of the lost work after the punctuation, as my research focuses exclusively on his quoting language—specifically, authors' names, works' titles, and verbs introducing quotations. If the next sentence contains any of these three linguistic elements, it will be considered as another quoting passage. Consider Polyb. II 56, 6–7:

[6] βουλόμενος δὴ διασαφεῖν τὴν ὠμότητα τὴν Ἀντιγόνου καὶ Μακεδόνων, ἅμα δὲ τούτοις τὴν Ἀράτου καὶ τῶν Ἀχαιῶν, φησὶ τοὺς Μαντινέας γενομένους ὑποχειρίους μεγάλοις περιπεσεῖν ἀτυχήμασι, καὶ τὴν ἀρχαιοτάτην καὶ μεγίστην πόλιν τῶν κατὰ τὴν Ἀρκαδίαν τηλικαύταις παλαῖσαι συμφοραῖς ὥστε πάντας εἰς ἐπίστασιν καὶ δάκρυα τοὺς Ἕλληνας ἀγαγεῖν. [7] σπουδάζων δ' εἰς ἔλεον ἐκκαλεῖσθαι τοὺς ἀναγινώσκοντας καὶ συμπαθεῖς ποιεῖν τοῖς λεγομένοις, εἰσάγει περιπλοκὰς γυναικῶν καὶ κόμας διερριμμένας καὶ μαστῶν ἐκβολάς, πρὸς δὲ τούτοις δάκρυα καὶ θρήνους ἀνδρῶν καὶ γυναικῶν ἀναμὶξ τέκνοις καὶ γονεῦσι γηραιοῖς ἀπαγομένων.

[6] Wishing, for instance, to insist on the cruelty of Antigonus and the Macedonians and also on that of Aratus and the Achaeans, he [scil. Phylarchus] tells us that the Mantineans, when they surrendered, were exposed to terrible sufferings and that such were the misfortunes that overtook this, the most ancient and greatest city in Arcadia, as to impress deeply and move to tears all the Greeks. [7] In his eagerness to arouse the pity and attention of his readers he treats us to a picture of clinging women with their hair disheveled and their breasts bare, or again of crowds of both sexes together with their children and aged parents weeping and lamenting as they are led away to slavery.

According to the principles stated above, subparagraph 6 and subparagraph 7 will constitute two different quoting passages and therefore two different Wikibase items—namely, Q47 "Polyb. II 56, 6" and Q48 "Polyb. II 56, 7".

2.2 Quoted author(s) and quoted work(s)

These statements (Fig. 3) identify the quoted author and work and describe metadata about the quotation—notably, its linguistic characteristics. The value of P13 "quoted author(s)" is the item Q35 "Ephorus of Cyme", which is an instance of person. On the other hand, the value of P15 "quoted work(s)" is Q36 "Ἐφόρου Ἱστορίαι", which is an instance of *work*.

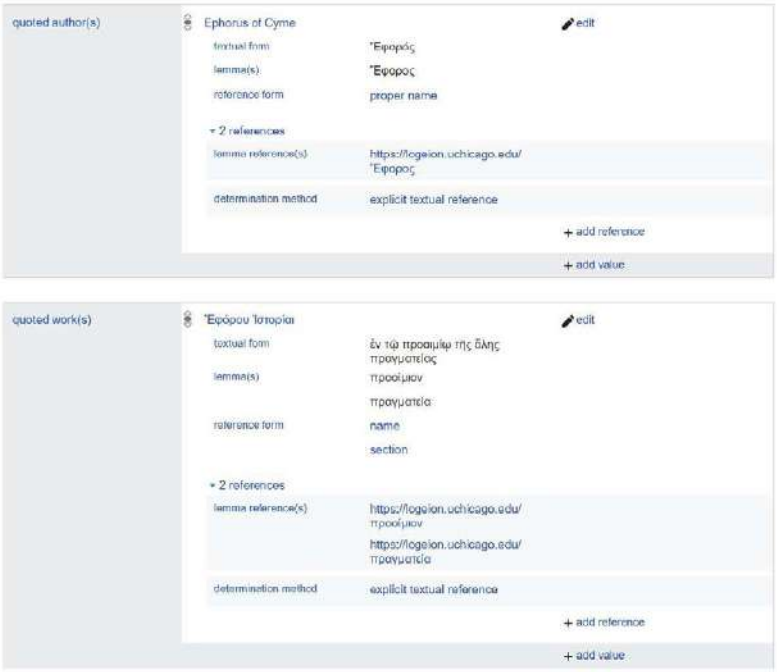


Figure 3 – Quoted author(s) and work(s) of Q34.

The first qualifier of both statements highlights the exact textual form of the quotations. The value of P16 "textual form" is a string representing Polybius' words referring to the quoted author and work (Ἔφορός and ἐν τῷ προοιμίῳ τῆς ὅλης πραγματείας). The second qualifier specifies the lemma of Polybius' quoting words. In the statement about the quoted author, the lemma is Ἔφορος. The statement about the quoted work has instead two lemmas, as Polybius' text refers both to Ephorus' work (πραγματείας) and to a specific section of it (προοιμίῳ). The last qualifier describes the quotation form. As shown in the data model (Fig. 1), P17 "reference form" is the same, but the items constituting its possible values change based on the statement.

As for "quoted work(s)", the data model (Fig. 1) shows that the possible values for P17 "reference form" are Q40 "name" and Q41 "section". In Q34 (Fig. 3), they are both used. However, there are examples where only one of these values

is required. Consider "Polyb. X 44, 1 (Q45)": Αἰνείας δὲ βουληθεὶς διορθώσασθαι τὴν τοιαύτην ἀπορίαν, ὁ *τὰ περὶ τῶν Στρατηγικῶν* ὑπομνήματα συντεταγμένος, βραχὺ μέν τι προεβίβασε, τοῦ γε μὴν δέοντος ἀκμὴν πάμπολυ τὸ κατὰ τὴν ἐπίνοιαν ἀπελείφθη ("Aeneas, the author of *the work on strategy*, wishing to find a remedy for the difficulty, advanced matters a little, but his device still fell far short of our requirements, as can be seen from this description of it.") Here, Polybius only refers to Aeneas' work (τὰ περὶ τῶν Στρατηγικῶν ὑπομνήματα); the qualifier thus requires only "name"[17] as its value.

As for "quoted author(s)", the possible values of P17 "reference form" are Q37 "proper name", Q38 "ethnic", and Q39 "generic noun". In Q34 (Fig. 3), Polybius refers to Ephorus only by his proper name, meaning that only Q37 "proper name" is used. In Q44 "Polyb. I 3, 2", P17 will instead have both the values Q37 "proper names" and Q38 "ethnic", as Polybius specifies the provenance of the quoted historian: ταῦτα δ' ἔστι συνεχῆ τοῖς τελευταίοις τῆς παρ' *(Ἀρ)άτου Σικυωνίου* συντάξεως (Trans. These events immediately succeed those related at the end of the work of *Aratus of Sicyon*.) Q39 "generic noun" is used when the quoted authors are identified by a noun indicating the literary genre of their work, such as in Q43 "Polyb. XII 13, 7": οὐ γὰρ ἂν *Ἀρχέδικος ὁ κωμῳδιογράφος* ἔλεγε ταῦτα μόνος ("For in that case not only *Archedicus, the comic poet*, would… have said this.") Here, Polybius refers to Archedicus both by his proper name (Ἀρχέδικος) and by a noun indicating his role as comedy writer (κωμῳδιογράφος).

Both statements in Fig. 3 have two references. The data model (Fig. 1) shows that the property P14 "determination method" has two items as possible values: Q33 "explicit textual reference", and Q32 "inferred from the context". This reference is necessary to indicate how the Wikibase editor can verify that the quoted authors and works match those declared in the statements. As a matter of fact, while many quoting passages name the author and the work, there are cases where this must be inferred from the context. Consider Q42 "Polyb. XVI 19, 1": Μετὰ δὲ ταῦτά φησι καταπροτερουμένην τὴν φάλαγγα ταῖς εὐχειρίαις καὶ πιεζομένην ὑπὸ τῶν Αἰτωλῶν ἀναχωρεῖν ἐπὶ πόδα, τὰ (δὲ) θηρία τοὺς ἐγκλίνοντας ἐκδεχόμενα καὶ συμπίπτοντα τοῖς πολμίοις μεγάλην παρέχεσθαι χρείαν ("Next he states that the phalanx, proving inferior in fighting power and pressed hard by the Aetolians, retreated slowly, but that the elephants were of great service in receiving them in their retreat and engaging the enemy.") Here, Polybius does not specify neither the author's name nor his work's. Let us take a look at the statements about the author and work of Q42 (Fig. 4).

The broader context of the passages clarifies that Polybius is quoting the work of Zeno of Rhodes. The Wikibase editor knows that and must specify it in the reference.[18] It is worth noting that in such a case, the editor must use the

[17] Since Polybius refers to the same work using different nouns—which cannot always be considered titles—the item has been labelled as "name" rather than "title."

[18] Q32 "inferred from the context" is also used when only the section of the work only is specified (e.g., Q4), and when the author is identified only through an ethnic, or generic noun (e.g., Q46).

statements about the quoted authors and works, but the qualifier indicating the textual form of the quotations will have "no value".

Returning to the statements about quoted author(s) and work(s) of Q34 (Fig. 3), the final references include URLs linking to the lemma entries in *Logeion*, an open access database of Latin and Ancient Greek dictionaries.[19]

Figure 4 – Quoted author(s) and work(s) of Q42.

2.3 Reporting verb(s)

The final statement of Q34 (Fig. 4) specifies the "reporting verb(s)." These are verbs associated with the content of the quoted works, such as those denoting the act of saying, writing, or composing. The value of P21 "reporting verb(s)" is a string representing the lemma of the verb. This statement includes two qualifiers. While the first specifies the verb's form as it appears in Polybius' text, the second identifies the verb's subject. The subject qualifier is essential for quoting passages where multiple verbs are attributed to multiple authors. Consider Q43 "Polyb. XII 13, 7", where λέγω has Archedicus as its subject, and φημί refers to Timaeus: οὐ γὰρ ἂν Ἀρχέδικος ὁ κωμῳδιογράφος ἔλεγε ταῦτα μόνος περὶ Δημοχάρους, ὡς Τίμαιός φησιν ("For in that case not only Archedicus, the comic poet, would, as Timaeus asserts, have said this.") Once the Wikibase is completed, this data structure will enable users to investigate, through the SPARQL Query Service, which verbs Polybius uses for different authors.

If we look at the data model (Fig. 1) again, we will notice that the statement about the reporting verbs is not necessary. This is why there are also quoting passages where Polybius does not use a reporting verb, such as Q44 "Polyb. I 3, 2," cited above: ταῦτα δ' ἔστι συνεχῆ τοῖς τελευταίοις τῆς παρ' (Ἀρ)άτου

[19] <https://logeion.uchicago.edu/λόγος>.

Σικυωνίου συντάξεως ("These events immediately succeed those related at the end of the work of Aratus of Sicyon.") In such cases, the Wikibase editor must omit the statement entirely, since there is no verb to analyse. The special item "no value" must not be used here, as it implies that the verb should be there, but is missing.

Figure 5 – Reporting verb(s) of Q34.

3. Conclusion: further developments

As more items are added to the Wikibase, the data model will evolve to accommodate new and more specific types of quoting passages. Therefore, additions and improvements to the data model are currently being considered. In this section, I will discuss some of them.

First, the data model is being refined to highlight other linguistic elements that may prove relevant for this research. For instance, in Q45 "Polyb. X 44, 1": Αἰνείας δὲ βουληθεὶς διορθώσασθαι τὴν τοιαύτην ἀπορίαν, ὁ τὰ περὶ τῶν Στρατηγικῶν ὑπομνήματα συντεταγμένος (Trans. Aeneas, the author of the work on strategy, wishing to find a remedy for the difficulty), it may prove useful to create statements identifying the use of the preposition περί with the genitive to indicate the title or content of a specific work.

Another considered improvement involves automating the integration of Wikidata's entries about ancient authors. Wikidata is indeed a very rich database, with many well-curated entries about the ancient world, such as those for Ephorus and his *Histories*.[20] Therefore, integrating them into the Wikibase rather than simply creating a link to the Wikidata item via a property would be highly beneficial and would save considerable time for the Wikibase editor.

As suggested during the conference, the data model could also benefit from the implementation of Wikibase Lexemes. In the current data model (Fig. 1), properties identifying lexical elements have string as their data type. However, P16 "textual form" could be assigned the lexeme data type. In this manner, the exact word used by Polybius—e.g., πραγματείας in Q34—would be

[20] Ephorus Wikidata item: <https://www.wikidata.org/wiki/Q313791>. Ephorus' *Histories* Wikidata item: <https://www.wikidata.org/wiki/Q42190991>.

linked to a lexeme page where users can find all relevant grammatical information (gender, number, and case), the lemma, and the reference form it represents (e.g., whether it is a name indicating a work or the section of a work). With this change, the statements about quoted author(s) and work(s) will only require the "textual form" statement as a qualifier. As for the verbs, P21 "reporting verb(s)" should also be assigned the lexeme data type, and its value should be the exact form of the verb used by Polybius. Therefore, in Q34, P21 "reporting verb(s)" should have the lexeme φησιν as value, and the P16 "textual form" qualifier should be eliminated. However, the qualifier specifying the subject of the verb must remain, as this information cannot be included in the lexeme data.

A further addition under consideration is the creation of a statement to classify each quoting passage according to its designation in Jacoby's *Die Fragmente der Griechischen Historiker*. However, this presents a difficulty, as the passages documented in the database do not always correspond exactly to the fragments as they are presented in the Jacoby collection. It is therefore necessary to determine how to address this issue before implementing such a statement.

The ultimate goal of the database is to detect patterns within the data to gain relevant insights into Polybius's quoting language and practice, which will be interpreted as part of my PhD project. Therefore, once the database contains a sufficient amount of data, SPARQL queries will be performed. The following are examples of queries that may prove useful for this research:

- Which authors are referred to by a generic noun?
- Which authors are referred to by an ethnic?
- Which works are referred to both by name and by section?
- Which verbs are used to introduce citations of specific authors?

In conclusion, the main objective of *The Library of Polybius* is to adapt a traditional philological approach to a digital environment, demonstrating how new computational methods can enhance classical scholarly approaches. Although the Wikibase is still under development, this paper has demonstrated its effectiveness in systematically cataloguing data and metadata of Polybius's quoting passages. The statements not only specify which authors, and which works Polybius quotes but also highlight elements of his quoting language and briefly describe them. Through this method, one not only creates structured data that is openly available to anyone according to the LOD principles, but classical philologists can also address and analyse traditional questions—such as the *ratio laudandi* of a cover-text—in a more efficient way by querying these data.[21]

[21] This work is supported by the MSCA Doctoral Networks 2022 program (HORIZON-MSCA-2022-DN-01) (Project MECANO. The Mechanics of Canon Formation and the Transmission of Knowledge from Greco-Roman Antiquity. Grant agreement number 101120349).

References

Babeu, Alison. 2019. "The Perseus Catalog: Of FRBR, Finding Aids, Linked Data, and Open Greek and Latin." In *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, 53–72. Berlin: De Gruyter Saur.

Berti, Monica. 2021. *Digital Editions of Historical Fragmentary Texts*. Heidelberg: Propylaeum (Digital Classics Books 5.).

Berti, Monica. 2024. "Digital Practice for Studying the Indirect Transmission of Classical Authors and Works." In *Treasuries of Literature: Anthologies, Lexica, Scholia and the Indirect Tradition of Classical Texts in the Greek World*, edited by Federico Favi, and Virginia Mastellari, 189–206. Berlin, Boston: De Gruyter.

Blackwell, Christopher, and Neel Smith. 2019. "Homer and the Papyri: The Canonical Text in Antiquity." In *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, 73–92. Berlin: De Gruyter Saur.

Büttner-Wobst, Theodor, edited by. 1882–1905[2]. Polybius, *Historiae*, 4 vols. Leipzig: B. G. Teubner (Bibliotheca Scriptorum Graecorum et Romanorum Teubneriana).

Cayless, Hugh A. 2019 "Sustaining Linked Ancient World Data." In *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, 35–50. Berlin: De Gruyter Saur.

Hau, Lisa I. 2024 "Athenaeus as Covertext for Hellenistic Historiography: The Fragments of Polybius as Case Study." In *Fragmente einer fragmentierten Welt: Zur Problematik des Umgangs mit Fragmenten in der gegenwärtigen klassisch-philologischen Forschung*, edited by Felix Neuerburg, Theodoros Tsiampokalos, and Piotr Wozniczka, 151–76. Berlin: De Gruyter.

Jacoby, Felix, edited by. 1923–1958; 1994–present. *Die Fragmente der Griechischen Historiker*, multiple vols. Berlin: Weidmann; Leiden: Brill.

Paton, W. R., W. Walbank, and Christian Habicht, edited by. 2010–2012. Polybius, *The Histories*, translated by W. R. Paton, revised by F. W. Walbank and Christian Habicht, 6 vols. Cambridge (MA): Harvard University Press (Loeb Classical Library 128, 137–38, 159, 160–61).

Riva, Pat, Patrick Le Bœuf, and Maja Žumer. 2017. *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*. The Hague: International Federation of Library Associations and Institutions.

Schepens, Guido. 1997. "Jacoby's *FGrHist*: Problems, Methods, Prospects." In *Collecting Fragments – Fragmente Sammeln*, edited by Glenn W. Most, 144–72. Göttingen: Vandenhoeck & Ruprecht.

Reimagining Digital Gazetteers: A Wikidata-Powered Approach

Maxime Guénette

Abstract: This paper explores the potential of Wikidata as a multilingual, collaborative knowledge base for producing digital gazetteers within a Linked Open Data framework. Through two case studies — Temples in Roman Britain and the International (Digital) Dura-Europos Archive (IDEA) — we show how Wikidata can support gazetteer projects of varying scales while improving their interoperability and opening them to broader communities of contributors. We also address recurring concerns: data quality, the technical barriers to participation, and the postcolonial critiques that any infrastructure built on collaborative consensus must confront. Together, these cases support the argument that Wikidata offers a realistic and adaptable platform for digital gazetteers — one capable of sustaining both rigorous scholarly use and more inclusive practices of heritage engagement.

Keywords: Wikidata, Digital Gazetteer, International (Digital) Dura-Europos Archive, Roman temples, Linked Open Data.

1. Introduction

Over the years, the Humanities have shown a growing interest in cataloguing places and the ways in which they are perceived, represented, and experienced through gazetteers. In its most simplistic state, a gazetteer consists of a list of places with basic attributes like names, feature types and coordinates (Berman, Mostern, and Southall 2016). Because of their minimalist structure, gazetteers have sometimes been deemed relevant as an ancillary product of visual cartography with limited academic value (Horne 2020a, 215).

With the rise of the web in the 1990s, gazetteers underwent major transformations, shifting for the first time from paper format to the digital environment. Freed from the constraints of print, digital catalogues are more flexible and less limited in size, allowing researchers to add virtually any information about a place more efficiently. Unlike printed versions, digital catalogues can be dynamic, evolving alongside research to reflect the latest knowledge in a field. When published under open licenses such as CC BY 4.0, which allow the reuse and modification of data with certain restrictions, they can also serve as a foundation for future research projects and initiatives. In this way, digital gazetteers can be seen as living documents that can be improved, expanded and repurposed over time (Horne 2020a).

Maxime Guénette, Université de Montréal, Canada, maxime.guenette@umontreal.ca, 0009-0006-2076-1220

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Maxime Guénette, *Reimagining Digital Gazetteers: A Wikidata-Powered Approach,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.15, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 99-110, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

One of the most widely used and long-lived digital gazetteers in the field of Classics is Pleiades[1]. Initially developed as a digital successor to the Barrington Atlas of the Greek and Roman World, Pleiades records tens of thousands of places of pre-modern cultures, documenting name variations, geographic coordinates, and historical metadata. Maintained by academic contributors, all new entries are subject to a rigorous peer-review process before publication. Expanding beyond its print predecessor, Pleiades offers structured metadata, including preferred titles, contributor lists, coordinates, multilingual historical names, place types, and linked bibliographic references (Elliott and Gillies 2009).

A growing number of digital gazetteers, including Pleiades, are using Linked Open Data (LOD) to enrich, improve and publish their data (Grunewald and Mostern 2024). The overall aim of LOD is to provide structured datasets that are freely accessible, interlinked, and cross-searchable on the web, allowing it to function less like a sea of data populated by isolated islands and more like a dynamic map able to establish connections between places, people, events, objects and so on. By adopting LOD principles, these gazetteers facilitate interoperability between different datasets, enabling scholars to integrate and contextualize information from multiple sources more easily and with greater precision.

A key characteristic of these LOD-based gazetteers is their reliance on Unique Resource Identifiers (URIs) to disambiguate places. These URIs provide persistent, machine-readable identifiers that enable precise citation, linking, and integration across multiple datasets. By assigning stable URIs to places, gazetteers provide a reliable mechanism for referencing spatial entities while promoting semantic interoperability across platforms. This approach not only enhances the accuracy and traceability of historical research but also stimulates new possibilities for interdisciplinary collaboration, allowing scholars to draw connections between different, often disconnected fields and uncover patterns that were previously hidden across isolated datasets (Grossner, Janowicz, and Keßler 2016; Horne 2020b).

Although several projects such as Recogito[2] and Peripleo[3] have aimed to facilitate the creation and access of spatial LOD, significant barriers remain for much of the scientific community. While these tools are freely available, the broader implementation of LOD often involves hidden or long-term costs — such as technical infrastructure, data hosting, maintenance, and the need for specialized skills to manage complex ontologies or align vocabularies. These challenges are particularly apparent within smaller or unfunded projects, which might not have the institutional support necessary to sustain LOD workflows over time (Cayless 2019; Geser 2016; Middle 2022).

This paper argues that Wikidata can significantly reduce barriers for digital gazetteers published as LOD. It offers a shared, interoperable framework that makes publication easier and ensures that gazetteers remain visible and reusable across disciplines.

[1] <https://pleiades.stoa.org/>. All cited links were last accessed on June 11, 2026.

[2] <https://recogito.pelagios.org>.

[3] <https://github.com/britishlibrary/peripleo-lanc>.

2. Wikidata as a platform for digital gazetteers

Since its inception in late 2012, Wikidata has grown into one of the largest public knowledge graphs on the web. As a structured, multilingual, and openly accessible knowledge base built on LOD principles, Wikidata allows scholars to query and connect their datasets to a wider ecosystem of historical, geographic, and bibliographic resources. Its collaborative editing model means that data is continuously enriched and corrected by a broad community of contributors — a strength that also introduces well-documented risks of inconsistency and uneven coverage. (Vrandečić, Pintscher, and Krötzsch 2023).

Due to these benefits, Wikidata has become increasingly popular in research, especially in Digital Humanities (DH). A recent survey of Wikidata's use in DH projects reveals that it is often viewed as a content provider for a wide range of topics, primarily used to annotate and enrich materials, or as a hub for linking various datasets. However, while the quality of Wikidata's data can vary from one subject to another, most projects tend to only consume data rather than contribute or correct it (Cook 2019; Zhao 2023). Thus, only a few projects are using Wikidata as a platform to create, store and share their data directly or use it to publish LOD. This highlights the need for more active contributions to improve data quality, ensuring that Wikidata can reach its full potential as a reliable and comprehensive resource for global knowledge, freely available to all.

Schmidt, Thiery and Trognitz have argued that Wikidata can serve as an accessible entry point into LOD practice for archaeologists who lack extensive technical training (Schmidt, Thiery, and Trognitz 2022). By using Wikidata as a publishing platform, DH projects contribute to the expansion and refinement of scientific data, enhancing free public knowledge while simultaneously integrating their datasets into the LOD ecosystem. The authors also point out how Wikidata adheres to FAIR (Findable, Accessible, Interoperable, Reusable) data principles by ensuring that research data is openly available, structured in a machine-readable format, linked to external resources, and continuously updated through community collaboration(Wilkinson et al. 2016).

Additionally, Wikidata fulfills all five of the key characteristics for modern gazetteers listed by Peter Bol (Bol 2011). First, it has extensive support for temporal data with properties such as time period (P2348)[4], start time (P580)[5] or end time (P582)[6]. Second, its multilingual structure ensures that data entered by editors is available in multiple languages, which strengthens global knowledge. Third, Wikidata is designed for interoperability: it links its entries to a wide range of external resources via their URIs, including authority files, encyclopedic platforms, and specialist databases such as Trismegistos Geo (P1958)[7], Pleiades (P1584)[8] or

[4] <https://www.wikidata.org/wiki/Property:P2348>.
[5] <https://www.wikidata.org/wiki/Property:P580>.
[6] <https://www.wikidata.org/wiki/Property:P582>.
[7] <https://www.wikidata.org/wiki/Property:P1958>.
[8] <https://www.wikidata.org/wiki/Property:P1584>.

ToposText place (P8068)[9]. Fourth, its data model is extensible, accommodating continual growth, under which users are encouraged to create or modify items and properties while leveraging an open API and SPARQL queries for enhanced adaptability. Lastly, long-term sustainability is realized through the institutional support of the Wikimedia Foundation, open licensing (CC0), and decentralized contributions, which guarantee data integrity, scalability, and accessibility over time.

Building on Bol's original list, later work identified two additional requirements for digital gazetteers adopting LOD: (1) improve methods for formally linking places with temporal entities such as historical periods and events, and (2) integrate place and period while accounting for spatial and temporal uncertainty (Grossner, Janowicz, and Keßler 2016). Both challenges can be addressed through Wikidata's data model. The first is supported by properties like significant event (P793)[10] or time period (P2348)[11], which allow places to be meaningfully connected to historical contexts. The second is handled through a combination of temporal properties, links to external temporal gazetteers like PeriodO (P9350)[12] or with qualifiers such as uncertainty (Q13649246)[13] and approximately (Q60070514)[14].

Despite its many benefits, Wikidata also comes with some challenges. While analysing Wikidata's integration in DH projects, Fudie Zhao categorized them in two main categories: technical implementation issues and concerns about data quality (Zhao 2023; Cook 2019). Anne Hunnell Chen identifies further challenges for archaeologists and GLAM institutions, most notably potential vandalism due to Wikidata's open editorial model (Chen 2023).

However, such vandalism appears to be rare in archaeological projects. Built-in mechanisms, such as ShEx (Shape Expressions) schemas, help safeguard metadata, ensuring data integrity and adherence to predefined data models (Thornton, Seals Nutt, and Chen 2024). Moreover, a recent study emphasized Wikidata's dynamic nature and the commitment of its editors to produce high-quality data (Shenoy et al. 2022). The DH community has a role to play here, bringing methodological experience that can inform more deliberate, domain-specific applications of Wikidata. That said, vandalism is not the only source of instability. Legitimate changes — new scholarship, shifting research priorities, varied contexts of use, or the ambiguities of partially overlapping places — can equally introduce data drift, with consequences for the reliability of identifiers and statements. In such cases, complementary strategies like version control and timestamping offer more targeted responses.

[9] <https://www.wikidata.org/wiki/Property:P8068>.

[10] <https://www.wikidata.org/wiki/Property:P793>.

[11] <https://www.wikidata.org/wiki/Property:P2348>.

[12] <https://www.wikidata.org/wiki/Property:P9350>.

[13] <https://www.wikidata.org/wiki/Q13649246>.

[14] <https://www.wikidata.org/wiki/Q60070514>.

Therefore although Wikidata presents certain challenges for research and creating digital gazetteers, its structured, multilingual, and open framework offers spatial humanities scholarship a powerful means of integrating diverse datasets and fostering cross-disciplinary collaboration. The following case studies — Temples in Roman Britain and the International (Digital) Dura-Europos Archive — demonstrate this potential in practice.

3. Temples in Roman Britain

The first case study, developed as part of my doctoral thesis, investigates the transformation of temples and sanctuaries in Roman Britain through a digital gazetteer published directly in Wikidata. By analyzing their spatial distribution and relationships with imperial power and local networks, we reconsider their role within the broader framework of globalization (Lambert 2024; Versluys 2014). This approach builds on postcolonial critiques of Romanization, defined as the impact of Rome on its conquered territories and populations, emphasizing connectivity, hybridity, and the agency of local communities.

The first step of our thesis consists in compiling a gazetteer of temples and shrines in Roman Britain to better understand their characteristics and transformations from the late Iron Age through the Roman period. This also serves as a foundation for analyzing the spatial relationships between these cult places and key elements of the built and natural environment—such as urban centers, military installations, and topographical features—thereby offering new insights into their function and significance within the broader landscape.

By using Wikidata as the backbone of our digital gazetteer, we ensure that our dataset remains easily accessible, interoperable, and linked with external resources. We also build on data already added and curated in Wikidata, such as cities' coordinates, historical persons etc. Moreover, we also contribute to free knowledge that can be reusable and expandable over time as new Roman temples and sanctuaries are frequently discovered in the United Kingdom. Given that, publishing the gazetteer on Wikidata allows it to evolve as a living document—one that can be continuously enriched and updated by the wider community as new data emerges or existing records are refined.

At the core of our gazetteer is a master CSV (Comma-Separated Values) file. We expand the more simplistic tripartite model of gazetteers—names, types, coordinates—to make the most out of the richer possibilities offered by digital infrastructures. Our dataset incorporates additional attributes such as alternative names (aliases), dates of foundation and abandonment, references to external resources like Pleiades, and key archaeological information. This richer structure reflects the specific demands of the Roman British corpus, where a site may be known under multiple names, attested across a long chronological span, and documented in datasets that predate our own.

Once the data is collected, it must be mapped onto Wikidata's data model — a process that requires both technical decisions and community engagement. In practice, this involves reconciling our gazetteer's attributes with corresponding

Wikidata properties and linking corresponding values to existing or newly created Wikidata items. To facilitate this process, we rely on tools such as OpenRefine and QuickStatements, which allow for efficient data reconciliation and batch uploads (Delpeuch et al. 2025; "QuickStatements 3.0" 2025).

Once our dataset is uploaded to Wikidata, it becomes possible to explore it more systematically using SPARQL queries. However, while the Wikidata Query Service can generate web maps, Wikidata is not intended for comprehensive spatial analysis. For more advanced spatial exploration, researchers typically turn to Geographical Information System (GIS) softwares such as ArcGIS or QGIS which offer robust spatial functionalities.

This workflow often results in data silos as datasets are either not published online, locked behind paywalls, or are not encoded as LOD. ublishing a gazetteer on Wikidata addresses part of this problem, but we argue that full integration with other Semantic Web standards and open-access tools is necessary to go further. Such an approach would help dismantle existing silos, promoting greater interoperability and significantly enhancing the usability of digital gazetteers for future spatial and historical analysis.

To bridge the gap between Wikidata and GIS, we are using the SPARQLing Unicorn QGIS Plugin[15] (Thiery and Homburg 2025). Despite the growing community around Wikidata and other LOD initiatives for spatial projects, there is a lack of user-friendly, free and open tools available to researchers. The SPARQLing Unicorn plugin addresses this issue by enabling users to import GeoJSON layers into QGIS directly from SPARQL endpoints—not only from Wikidata, but also from other LOD projects such as Kerameikos, Nomisma, and Roman Open Data.

For instance, by executing SPARQL queries directly within QGIS, we can retrieve data on all Roman temples in Britain, including their geographic coordinates and associated attributes, and conduct spatial analyses such as proximity measurements, clustering, or point pattern analysis. Additionally, the gazetteer includes a wide range of attributes, such as temporal attestation, architectural forms, and associated topography, which can be filtered through more complex SPARQL queries. For example, one might query all Romano-Celtic temples in London attested in the second century.

In order to investigate how sacred spaces interacted with their environment—cities, military forts, roads, and more—across time and space, we adopt a deep mapping approach (Bodenhamer, Corrigan, and Harris 2015, 2021). Deep maps provide nuanced insights into how sacred spaces were positioned and functioned within the larger landscape. In practice, the ability to overlay multiple spatial datasets—temples, settlements, roads, natural features—within a GIS environment greatly enhances our capacity to explore patterns of proximity, visibility, and connectivity. To support this, we rely on complementary and external gazetteers, such as Pleiades, DARE, and Itiner-e, to enrich our mapping process.

[15] <https://github.com/sparqlunicorn/sparqlunicornGoesGIS>.

Nevertheless, working with external datasets is not always straightforward. Many gazetteers like The Rural Settlement of Roman Britain (Allen et al. 2015) only partially (or do not follow at all) Linked Open Data standards, which can limit interoperability and data reuse over time. To overcome this, we chose to contribute to the LOD ecosystem in two ways. Firstly, for archaeological sites and temples missing in Wikidata but listed in external gazetteers and resources, we create new entries directly on the platform to ensure their inclusion in the LOD ecosystem. Secondly, if an archaeological site or sacred space is already in Wikidata, we verify and improve existing records by cross-referencing them with reliable external sources, updating any missing or outdated information, and ensuring proper links to related datasets. This helps maintain data accuracy, enriches the resource, and strengthens its utility for ongoing research (Elliott 2024).

This case study demonstrates that Wikidata is a workable platform for digital gazetteers in Roman archaeology — not only because it makes data openly accessible and interoperable, but because it can absorb new discoveries without editorial disruption. The integration of Semantic Web standards, and in particular tools like the SPARQLing Unicorn QGIS plugin, makes it possible to move directly from structured Wikidata queries to spatial analysis in GIS environments — allowing us to examine the relationships between sacred spaces and their landscapes without switching between incompatible data ecosystems.

4. The International (Digital) Dura-Europos Archive

IDEA was established in 2020 to build a digital urban gazetteer for Dura-Europos, an ancient city in modern Syria. Founded around 300 BCE by the Seleucids on the west bank of the Euphrates near Salihiyeh, the city passed successively under Parthian and then Roman control before its destruction by the Sassanid Persians around 256 CE. The remarkable preservation of Dura-Europos is largely due to a defensive strategy implemented during one of its sieges: the interior of the city's western rampart was reinforced with tightly packed earth and debris. This desperate action, together with the hot, dry climate, created ideal conditions for the archaeological preservation of the city.

The first remains of Dura-Europos were uncovered in 1920 by the English captain M. C. Murphy during a period of territorial jostling between Western powers following the collapse of the Ottoman Empire. This "discovery" was followed by a one-day excavation by the archaeologist James Henry Breasted[16]. Further investigation of the city was carried out by Franz Cumont between 1922 and 1923. After a hiatus due to regional unrest, renewed excavations jointly financed by the *Académie des inscriptions et belles-lettres* and Yale University resumed in 1928 as a series of campaigns that would span the next decade. Finally, new

[16] While scientific exploration of the ruins began in the 1920s under the auspices of foreign-led missions, the remains of the ancient city had long been known to the inhabitants of Salhiyeh and the surrounding region.

excavations began in 1986 with the *Mission Franco-Syrienne de Dura-Europos*, but were interrupted in 2011 by the Syrian civil war (Baird 2018).

Since the outbreak of the civil war, Dura-Europos has been mostly inaccessible and subject to extensive looting (Baird 2020). Moreover, because it was common practice to share artefacts in early twentieth century archaeological excavations, most of the objects unearthed at Dura-Europos are now scattered across more than ten museums and archives worldwide.

Therefore, IDEA is developing digital archival methods and content to make archaeological knowledge about Dura accessible to both scholars and the public. Its urban, site-specific gazetteer has a concrete aim: to bring together artifacts and archaeological records from a city whose material remains are now scattered across multiple collections, languages, and institutions.

Although digital gazetteers have become progressively valuable resources in DH, they often lack the needed granularity in archaeological research. For instance, simply stating that a specific artifact was found at Dura-Europos is informative, but vague. Since archaeologists have divided the city into blocks, houses, rooms, and even specific walls, it is much more accurate to state the exact and precise location of an artifact. With Wikidata's flexible and extensible data model, IDEA can capture this level of granularity. For example, we can specify that a certain artifact, for example the Altar to Gods of Palmyra (Q100299148)[17], was found by Maurice Pillet (Q3301265)[18] in 1929 in the temple of Bêl's courtyard (Q109303232)[19] and is now in the Yale University Art Gallery (Q1568434)[20] in New Haven. By using Wikidata, IDEA is thus building a and granular knowledge graph of archaeological and archival artifacts directly through LOD.

Nevertheless, choosing Wikidata is not just about creating granular Linked Open Data—it's also a conscious decision rooted in the political and historical context of Dura-Europos. Because the site's modern history is deeply entangled with imperialism and colonial-era archaeology, IDEA has two broader goals. First, it aims to clarify and standardize the often confusing placenames and references that emerged from the massive early 20th-century excavations. Second, by making archival information more accessible and searchable, IDEA works to support cultural heritage recovery—especially by helping Syrians reconnect with their past that is too often stored far from its place of origin.

Placename inconsistency in Dura-Europos is something well-known to researchers. Even the name of the city itself may vary depending on the scholarly tradition (Dura, Doura, Doura-Europos, Europos-Doura). Anne Chen identifies three main causes for placenames inconsistency: change over time, multilingualism and differences of interpretation (Chen 2024, 10). A striking example that

[17] <https://www.wikidata.org/wiki/Q100299148>.

[18] <https://www.wikidata.org/wiki/Q3301265>.

[19] <https://www.wikidata.org/wiki/Q109303232>.

[20] <https://www.wikidata.org/wiki/Q1568434>.

addresses these three causes is the temple of Bel (Q3517541)[21] in the north-west corner of the city. The temple was originally called the "Temple of the Palmyrene Gods" by early excavators but is now more commonly referred to as the "Temple of Bel." Multilingualism further contributes to this inconsistency, as the temple's name appears in various languages and scripts, such as Temple of Bel, Tempel des Baal, and معبد بل. Additionally, differing scholarly interpretations of the temple's primary deity have led to multiple names, including "Temple of Zeus," "Temple of the Palmyrene Gods," and "Temple of the Oriental Gods." Wikidata helps to centralise all these variants in one place to disambiguate confusing naming traditions rooted in colonialism.

Furthermore, as previously stated, Dura's metadata and archaeological publications are mainly in Western languages as it was excavated by French and American teams. Although metadata is accessible online, it remains undiscoverable for Syrian due to the lack of Arabic translation, making it even harder for them to access their own cultural heritage through digital surrogates. It also fails to take into consideration Syrian perspectives and interpretations of Dura-Europos. For example, recent ethnographic research has shown that Dura is known as *Alsur* (السور "The Wall") by Syrians living in the vicinity of the city, a name that was unknown to Western scholarship (Baird and Almohamad 2025). The same goes for the Palmyrene Gate (the main gate of the city leading to Palmyra) which is locally called *Bab al-Hawa* (باب الهوا, "The Gate of the Winds"). IDEA is thus using Wikidata to increase Dura's discoverability to Syrians by translating related items in Arabic and by incorporating local naming traditions into scholarship.

This case study demonstrates the potential of Wikidata as a promising platform for reconstructing fragmented archaeological and archival data in a site-specific gazetteer. By leveraging LOD, IDEA not only enables spatial metadata granularity related to Dura-Europos but also harmonizes heterogeneous and discordant placenaming conventions rooted in colonial scholarship. Furthermore, by providing Arabic translations and regional naming conventions, IDEA allows Syrian researchers as well as the broader Arabic public easier access to their own cultural heritage, which might otherwise be restricted to Western academic spheres. In doing so, IDEA illustrates how Wikidata can bridge historical data silos while addressing broader issues of archival visibility and postcolonial reinterrogation in archaeology.

5. Conclusion

This article has traced how Wikidata can function as more than a data repository — as an infrastructure that shapes how archaeological knowledge is structured, linked, and made available across languages and institutions. The two case studies bear this out in different ways: the Temples in Roman Britain gazetteer shows what Wikidata makes possible at scale, while IDEA shows how its data

[21] <https://www.wikidata.org/wiki/Q3517541>.

model can accommodate the fine-grained spatial and archival demands of site-specific archaeology. These examples demonstrate the potential of Wikidata not only as a repository, but as a space for critical engagement with historical data.

As digital gazetteers continue to evolve, we believe that the adoption of platforms such as Wikidata will play an important role in shaping the future of DH and LOD. By addressing challenges such as data quality, technical barriers, and inclusivity, scholars can take advantage of the full potential of digital gazetteers to advance knowledge, promote collaboration, and democratize access to cultural heritage.

References

Allen, Martyn, Tom Brindle, Alex Smith, Julian D. Richards, Tim Evans, Neil Holbrook, Michael Fulford, and Nathan Blick. 2015. "The Rural Settlement of Roman Britain: An Online Resource." Archaeology Data Service. https://doi.org/10.5284/1030449.

Baird, Jennifer A. 2018. *Dura-Europos*. London: Bloomsbury Academic.

Baird, Jennifer A. 2020. "The Ruination of Dura-Europos." *Theoretical Roman Archaeology Journal* 3 (1): 1-20. https://doi.org/10.16995/traj.421.

Baird, Jennifer A., and Adnan Almohamad. 2025. "The Ruins That Remain: Remembering Dura-Europos in Salhiyeh." In *Dura-Europos: Past, Present, Future*, edited by Lisa Brody and Anne Hunnell Chen, forthcoming. Turnhout: Brepols.

Berman, Merrick Lex, Ruth Mostern, and Humphrey Southall. 2016. "Introduction." In *Placing Names: Enriching and Integrating Gazetteers*, edited by Merrick Lex Berman, Ruth Mostern, and Humphrey Southall, 1-12. Bloomington: Indiana University Press.

Bodenhamer, David J., John Corrigan, and Trevor M. Harris, edited by. 2015. *Deep Maps and Spatial Narratives*. Bloomington: Indiana University Press. https://doi.org/10.2307/j.ctt1zxxzr2.

Bodenhamer, David J., John Corrigan, and Trevor M. Harris, edited by. 2021. *Making Deep Maps: Foundations, Approaches, and Methods*. Abingdon: Routledge.

Bol, Peter K. 2011. "What Do Humanists Want? What Do Humanists Need? What Might Humanists Get?" In *GeoHumanities*, edited by Michael Dear, Jim Ketchum, Sarah Luria, and Doug Richardson, 296-308. London: Routledge.

Cayless, Hugh A. 2019. "Sustaining Linked Ancient World Data." In *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, edited by Monica Berti, 35-50. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110599572-004.

Chen, Anne Hunnell. 2023. "Collaborative Curation of Digital Archaeological Archives: Promise, Prospects, and Challenges." In *Shaping Archaeological Archives: Dialogues Between Fieldwork, Museum Collections, and Private Archives*, edited by Rubina Raja, 33-46. Turnhout: Brepols. https://doi.org/10.1484/M.ARC-EB.5.133495.

Chen, Anne Hunnell. 2024. "What's in a Name? The Role of Digital Gazetteers for Postcolonial Legacy Archaeology." In *Trends in Archive Archaeology: Current Research on Archival Material from Fieldwork and Its Implications for Archaeological Practice*, edited by Jon M. Frey and Rubina Raja, 9-18. Turnhout: Brepols. https://doi.org/10.1484/M.ARC-EB.5.142129.

Cook, Stacey. 2019. "The Uses of Wikidata for Galleries, Libraries, Archives and Museums and Its Place in the Digital Humanities." *Comma* 2017 (2): 117-24. https://doi.org/10.3828/comma.2017.2.12.

Delpeuch, Antonin, Tom Morris, David Huynh, Weblate (bot), Stefano Mazzocchi, Jacky, Thad Guidry, et al. 2025. "OpenRefine." https://doi.org/10.5281/zenodo.10689569.

Elliott, Tom. 2024. "Wikidata Wednesdays." *Pleiades: A Gazetteer of Past Places*. https://pleiades.stoa.org/news/blog/wikidata-wednesdays.

Elliott, Tom, and Sean Gillies. 2009. "Digital Geography and Classics." *Digital Humanities Quarterly* 3 (1).

Geser, Guntram. 2016. *ARIADNE WP15 Study: Towards a Web of Archaeological Linked Open Data*. http://legacy.ariadne-infrastructure.eu/wp-content/uploads/2019/01/ARIADNE_archaeological_LOD_study_10-2016-1.pdf.

Grossner, Karl, Krzysztof Janowicz, and Carsten Keßler. 2016. "Place, Period, and Setting for Linked Data Gazetteers." In *Placing Names: Enriching and Integrating Gazetteers*, edited by Merrick Lex Berman, Ruth Mostern, and Humphrey Southall, 80-96. Bloomington: Indiana University Press.

Grunewald, Susan, and Ruth Mostern. 2024. "Working with Named Places: How and Why to Build a Gazetteer." *Programming Historian* 13. https://doi.org/10.46430/phen0117.

Horne, Ryan. 2020a. "Beyond Lists: Digital Gazetteers and Digital History." *The Historian* 82 (1): 37-50. https://doi.org/10.1080/00182370.2020.1725992.

Horne, Ryan. 2020b. "Mapping Power: Using HGIS and Linked Open Data to Study Ancient Greek Garrison Communities." In *Historical Geography, GIScience and Textual Analysis*, edited by Charles Travis, Francis Ludlow, and Ferenc Gyuris, 213-27. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-37569-0_13.

Lambert, Danielle Hyeonah. 2024. *Decolonizing Roman Imperialism: The Study of Rome, Romanization, and the Postcolonial Lens*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009491044.

Middle, Sarah. 2022. "Investigating Linked Data Usability for Ancient World Research." PhD thesis, London: The Open University. https://doi.org/10.21954/ou.ro.00014b1f.

"QuickStatements 3.0." 2025. https://qs-dev.toolforge.org/.

Schmidt, Sophie C., Florian Thiery, and Martina Trognitz. 2022. "Practices of Linked Open Data in Archaeology and Their Realisation in Wikidata." *Digital* 2 (3): 333-64. https://doi.org/10.3390/digital2030019.

Shenoy, Kartik, Filip Ilievski, Daniel Garijo, Daniel Schwabe, and Pedro Szekely. 2022. "A Study of the Quality of Wikidata." *Journal of Web Semantics* 72: 100679. https://doi.org/10.1016/j.websem.2021.100679.

Thiery, Florian, and Timo Homburg. 2025. "SPARQLing Unicorn QGIS Plugin." *Squirrel Papers*. https://doi.org/10.5281/zenodo.14874994.

Thornton, Katherine, Kenneth Seals Nutt, and Anne Chen. 2024. "Encoding Archaeological Data Models as Wikidata Schemas: Utilizing Shape Expressions to Structure Collaborative Linked Open Data for Digital Storytelling Within the International Dura-Europos Archive." *The International Journal of Technology, Knowledge, and Society* 21 (1): 69-83. https://doi.org/10.18848/1832-3669/CGP/v21i01/69-83.

Versluys, Miguel John. 2014. "Understanding Objects in Motion. An Archaeological Dialogue on Romanization." *Archaeological Dialogues* 21 (1): 1-20.

Vrandečić, Denny, Lydia Pintscher, and Markus Krötzsch. 2023. "Wikidata: The Making Of." In *Companion Proceedings of the ACM Web Conference 2023, Austin TX USA*, 615-24. Austin: ACM. https://doi.org/10.1145/3543873.3585579.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Zhao, Fudie. 2023. "A Systematic Review of Wikidata in Digital Humanities Projects." *Digital Scholarship in the Humanities* 38 (2): 852-74. https://doi.org/10.1093/llc/fqac083.

SEZIONE 3

Storia, arti, territorio /
History, Arts and Cultural Heritage

Developing *e*Viterbo: Challenges and Strategies for Open Linked Data in the Humanities

Alice Santiago Faria

Abstract: *e*Viterbo is a platform that combines a MediaWiki-based encyclopaedia with an open, structured linked data database, built on a Wikibase suite. Designed as a collaborative research tool, it operates under a CC BY-SA 4.0 license, ensuring openness and accessibility. Its structured data is intentionally designed to be shared with Wikidata, amplifying its potential for interoperability. This chapter examines the conceptual and technical decisions behind the creation of *e*Viterbo, including the development of its infrastructure, data model, and collaborative workflows. It also discusses the challenges faced in building and maintaining a platform of this nature to be constructed and subsequently sustained within a humanities-oriented environment.

Keywords: Wikibase, Linked Data, Ontology, Collaborative research, Humanities.

1. Introduction[1]

*e*Viterbo (WD_ID:Q116234460[2]) is an online encyclopaedic platform that combines text in MediaWiki-based pages with an open, structured database of linked data based on a Wikibase suite. Designed as a collaborative research tool, it operates under a Creative Commons licence and is intended to allow re-use of its content and data, ensuring openness and accessibility. Its structured data is deliberately designed to be shared with Wikidata, increasing its potential for interoperability and global knowledge integration.

This text examines the conceptual and technical choices that have shaped the creation and development of *e*Viterbo. In some aspects, it reiterates the report by Faria and Correia (2022), while in others, it complements it. The main aim is to reflect on the potential of Wikimedia projects for academia, but also to make note of some of the setbacks that research in the humanities encoun-

[1] This work is funded by national funds through the Fundação para a Ciência e a Tecnologia (FCT), I. P., under the CEECIND/00044/2017 and 2023.11076.TENURE.255. It also had the support of CHAM (NOVA FCSH/UAc), through the strategic project sponsored by FCT (UIDB/04666).

[2] <https://www.wikidata.org/wiki/Q116234460>.

Alice Santiago Faria, NOVA University Lisbon, Portugal, alicesantiagofaria@fcsh.unl.pt, 0000-0002-5006-4067

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

ters when working with such tools in practice. The purpose is to contribute to the development of synergies between academia and the Wikimedia projects.

The text is divided into two main sections, each comprising two parts. The initial section provides a concise introduction to the primary objectives of *e*Viterbo, as well as a brief historical account of its development to date. The subsequent section presents and discusses the general system and collaborative workflows developed under a research project funded by the Portuguese Foundation for Science and Technology. In addition, it examines the challenges faced in building and maintaining a platform of this kind in a humanities-oriented environment in Portugal. The second section of the text introduces plans for the future, specifically a second version, which has been tentatively named *e*Viterbo 2.0, following the recommendations outlined in the aforementioned report. This subsection will address two primary issues, as outlined in the report: the migration of the database to Wikibase.cloud and the transformation of the ontology.

Finally, the text will conclude with a brief reflection on the future of the platform, including strategies for its adaptation, evolution, and maintenance in the near future.

2. *e*Viterbo

1.1 Main goal and first steps for an open platform in the humanities

*e*Viterbo[3] is rooted in the *Art, History and Heritage* group and the *Heritage and Current Challenges* thematic line of the CHAM-Centre for the Humanities, Faculdade de Ciências Sociais e Humanas, Universidade NOVA de Lisboa (hereafter, CHAM-NOVA FCSH). Its primary goal is to support academic research in architectural history, urbanism or built environment history in general, art history, decorative arts, and other connected fields, concerning Portugal and related countries, while also serving the broader community by providing an online platform that consolidates and facilitates access to information on shared cultural heritage. The term "related countries" is used to refer to nations or territories that have been connected to Portugal through its colonial past, Portuguese-speaking or not, but also to countries linked by actors, by knowledge, practices, family, or any other type of ties.

The Portuguese General Regime for Archives and Archival Heritage sets its chronological scope. Since a considerable part of its data is biographical, a privacy policy regarding personal data has been established in accordance with the general regulation[4]. In short, this means that data will not be published until 30 years after the date of death of the biographed to whom the

[3] <https://eviterbo.fcsh.unl.pt/>. All webpages mentioned in the text were consulted and archived at https://arquivo.pt/ , 27 june 2026.

[4] <https://eviterbo.fcsh.unl.pt/wiki/EViterbo:Pol%C3%ADtica_de_privacidade>.

documents relate. If the date of death is not known, the data will only be published 40 years after the date of the documents.

The platform was named *e*Viterbo as a tribute to Francisco de Sousa Viterbo, author of several works dating from the transition from the 19th to the 20th century, which remain fundamental in the areas of study of the *Art, History, and Memory* research group within CHAM-NOVA FCSH.

*e*Viterbo was conceived as a collaborative initiative and tool designed to bring together researchers and research made in CHAM-NOVA FCSH. Its first objective was to consolidate individual pieces of information into a collective repository, facilitating the identification of shared interests and potential avenues for collaboration. This implied gathering dispersed data across various results of research done by researchers over the years, sharing it among peers, engaging in discussions, identifying diverse meanings and applications, recognizing informational gaps, going into the archives to solve information gaps, and attempting to co-complete, co-edit, and re-share the data.

Wikipedia, the most comprehensive, digital, and collaborative encyclopaedia to date, emerged as a reference and Wikimedia toolbox followed. The reasons behind the use of Wikimedia software, Mediawiki, were numerous: an easily recognisable and collaborative tool, open software, easy connection between structured and unstructured data, easy to edit, possibility of a closed environment as a way to control the dataset observing the principles of FAIR data—findable, accessible, interoperable, reusable data—, align to be shared with Wikidata. All *e*Viterbo content, available under a Creative Commons Attribution-ShareAlike 4.0[5] licence, can be freely reused, bridging the gap between academia (Lindeman 2025 and see Wikifair[6]) and society.

The reasons behind doing a controlled wiki instance are mostly related to academic obligations of productivity to which each researcher is obliged, but also with notability issues on some of Wikimedia project, namely Wikipedia, and control of data (see also FactGrid blog[7] and documentation on this subject). Finally, but importantly, this solution would accept the incompleteness of historical data, even if, in practice, some caused disturbances in the infrastructure.

Based on previous experiences, several of the researchers involved in *e*Viterbo team anticipated the challenge of maintaining a sustained academic interest in updating the platform's content. Consequently, various strategic decisions were made to ensure continued engagement from the academic community and, in turn, foster the platform's growth. These measures included the attribution of authorship, rigorous quality control through peer-reviewed entries, the assignment of an International Standard Book Number (ISBN),

[5] <https://creativecommons.org/licenses/by-sa/4.0/>.

[6] <https://meta.wikimedia.org/wiki/WikiFAIR>.

[7] <https://blog.factgrid.de/archives/1591>.

the use of URI's and the incorporation of Digital Object Identifiers (DOI) for individual entries.

Additional decisions shaped the structure of the platform, both from a team and a user perspective, some of which were not easy to make. These included accepting that *e*Viterbo would remain an experimental tool and would never be fully finished; that its development would proceed gradually and often in segments, depending on the availability of researchers acting as volunteer contributors or funding for research projects; the use (or not) of discussion pages that would later become public; accepting all variants of the Portuguese language, with the choice of which to use left to the authors; and welcoming all contributions that followed the platform guidelines and were accepted through a peer-reviewed process (applied only to pages with an attributed DOI, as will be explained later in the next section).

It is important to emphasise that the objective of the project was never to insert "a lot of data" into Wikidata, but rather to contribute with quality data. The extraction of data was not a factor that was taken into consideration.

1.2 Development under TechNetEmpire Research Project

*e*Viterbo first steps were taken in 2017, at the same time as the application of the research project TechNetEMPIRE, *Technoscientific Networks in the construction of the built environment in the Portuguese Empire (1647-1871)*, which was supported by FCT, Fundação para Ciência e Tecnologia (PTDC/ART-DAQ/31959/2017) but only started with a complete team in April 2019. The principle behind TechNetEMPIRE was to gather scattered knowledge, bringing together researchers working in different geographical spaces and chronologies of the former Portuguese Empire, aiming to examine the technical and scientific networks of the built environment (including engineering, architecture, cartography, carpentry, masonry etc.) in the Portuguese Colonial Empire. The work began with the financial support of CHAM and the volunteer efforts of a Professor from the Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Aveiro University, despite not knowing the results of the funding competition.

The group aimed to use Francisco de Sousa Viterbo's bibliography as a test (Viterbo 1892, 1898, 1899, 1903, 1917). The goal was to adopt a prosopographical/biographical method for analysing architects, engineers, artists, craftsmen, and institutions. The 19 volumes of the *History of Scientific, Literary and Artistic Institutions* by José Silvestre Ribeiro (1872–1914) were also used as a starting point, although not systematically transposed like Sousa Viterbo's bibliography. Simultaneously, the need to standardise concepts and terminology led to a parallel project within the research group focused on vocabulary, which ultimately resulted in a glossary based on Rafael Bluteau's early 18th-century dictionary (Bluteau 1712). The project of a digital encyclopaedia was born, based on three pillars: people, institutions, and the glossary (Faria and Correia 2022, 5–14, 18–20).

The general model of the platform (Fig. 1) and data modelling comes from this first stage. The most profound change was after the first attempt using Cargo extension to deal with the structured data. Wikibase was installed in the system at the end of 2019, with the curation of data and the importance of ensuring the quality of structured data being the main reasons. In practice, it means that *e*Viterbo structured data passed from a table-based database (cargo) to a triple-based database (wikibase), which breaks down all knowledge into three-part statements. Nevertheless, since it has two types of data—content and structural data—*e*Viterbo it still uses two database systems: a relational database, MySQL, for managing MediaWiki content, and a graph knowledge system, Wikibase, to manage structured data (Faria and Correia 2022, 5–14).

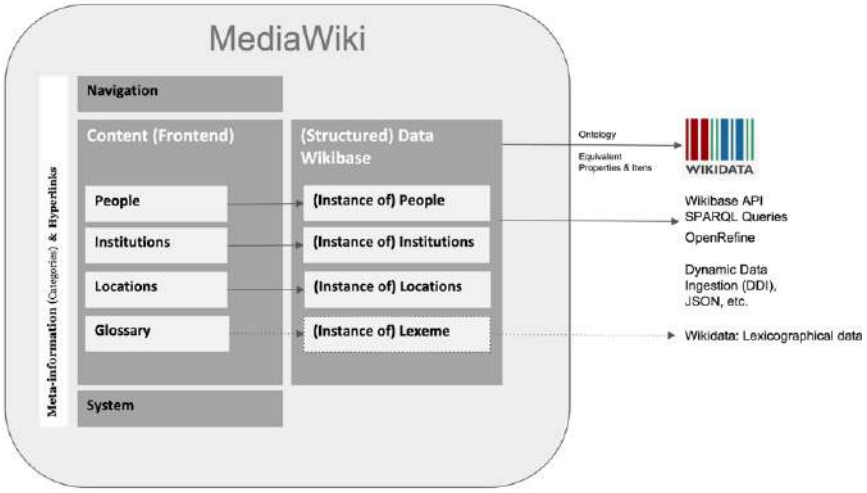


Figure 1 – Wiki *e*Viterbo general model.

Viterbo's information is organised into four main types of pages: content, data, navigation, and system. Each type, except for data pages, includes meta-information through "Categories", which group pages by topic or technical aspects, serving as navigation pages (Fig. 2). The platform is highly interconnected through interwiki links, with meta-information acting as a content aggregator. Just as a reference, by the end of TechNetEMPIRE project, there were more than 391 navigation pages constituted by aggregation using categories: the "People" category was grouped by chronology ("Active at [century]") across 10 pages; by "Institution" in 146 pages; by profession in 83 pages; and by work in 109 pages (a test not conducted systematically). Additionally, 32 main geographical categories, which also aggregate persons and institutions but are divided into subcategories with varying numbers (Ethiopia has no subcategories, whereas Brazil has 19, for instance).

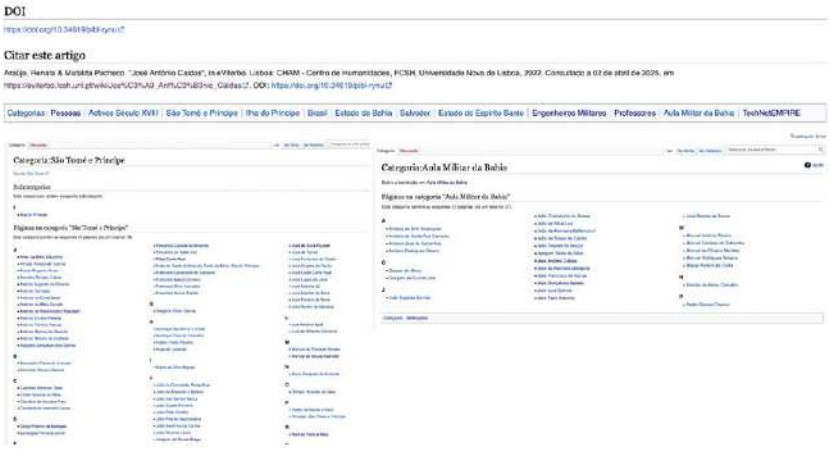


Figure 2 – Categories as presented on the page of José António Caldas[8] (top) and as navigation and aggregation pages linked by geography (left) and training institution (right), both departing from José António Caldas page.

Figure 3 – Example content + data pages. José António Caldas (Q794[9]).

System pages cover everything related to the platform itself, including user management, software details, extension settings, maintenance, and templates.

Content and data are connected in two different ways (Fig. 3). On the one hand, property "P30: local url"[10] directs the data to its source, the content page; on the other hand, dynamic infoboxes, created through transclusion, that is, inclusion of the content of data pages in content pages, using pre-defined templates

8 <https://eviterbo.fcsh.unl.pt/wiki/Jos%C3%A9_Ant%C3%B3nio_Caldas>.

9 <https://eviterbo.fcsh.unl.pt/wiki/Item:Q794>.

10 <https://eviterbo.fcsh.unl.pt/wiki/Property:P30>.

(Faria and Correia 2022, 16–23, 30–31). The role and importance of infoboxes within the system have evolved as the platform and associated workflows have changed. However, infoboxes continue to serve as a valuable tool for maintaining the connection between researchers, readers, and structured data.

Aligning the ontology with equivalent items and properties to Wikidata was done, even with doubts and a conscious need for remodelling the ontology and the hypothesis that in some cases this alignment might not be wanted, from *e*Viterbo team or even from a Wikidata community point of view. Exporting data to Wikidata remains a future plan. Exporting data using JSON (JavaScript Object Notation) and SPARQL queries was also previewed. However, JSON proved possible but "too messy", and SPARQL queries never worked, even after several attempts, using different methods.

The initial phase of the collaborative workflows of data was already described, departing from secondary bibliography, data was introduced manually in the platform. The information was not transcribed but rather resumed and enriched with other known sources and hyperlinks to the digitised items existing in the digital collection of the National Library of Portugal or other collections with digital items, in a work that was non-exhaustive. It aimed to have a picture of the amount of information that had been updated since Francisco Viterbo's publications.

As the research team had no prior experience working with MediaWiki software or editing in Wikimedia tools or communities (only the project coordinator had two years of experience, but with a low number of edits), it became necessary to produce several "how to..." guides.[11] These guides were updated regularly to reflect any changes made to the platform, which would impact the researcher's workflow. Over time, the more technical instructions disappeared, and new pages were created to assist researchers in writing entries consistently (both are still available in the system; see Faria and Correia 2022, 35–36). Some researchers just gave up the direct insertion of data into the platform, and "templates" in a textual format were created for their use.

The second stage involved searching for new data, assembling information from research conducted over the years by the team, as well as from archives and library catalogues, using relevant keywords. The retrieval of data from archives was done manually, putting together an open bibliographical database using Zotero[12]. Tables assembled 2877 persons and 105 institutions, scattered throughout the globe. Analysing the "names" of people and institutions, crossing over geography and chronology (categories) enabled researchers to identify entries to be written in collaboration, since the team has historical research mostly done in "pieces." For example, the area of work of the project principal investigator (PI) mainly was the Indian Ocean and the China Sea geographies

[11] See example of this guides here: https://eviterbo.fcsh.unl.pt/wiki/Gui%C3%A3o_de_cria%C3%A7%C3%A3o_de_p%C3%A1gina_-_Pessoas.

[12] <https://www.zotero.org/groups/2273153/technetempire>.

during the long 19th century, and the area of work of the co-principal investigator (Co-PI) is Brazil in the 18th century. From the "new names" and categories taken from this list, content pages were "automatically" created, and the number of pages more than doubled. From this point on, all data insertion was done manually and continued throughout the project.

After each researcher started to write entries designated to each one, but at the same time, to add information across the platform. The goal was to share information and collaborate. This process is transparent in the "view history" of each page, even if the team decided not to use the "discussion" page, also available in MediaWiki, because there was a conscience that even if, at the time, the platform was private, it would become public. Optional unit courses that allow students to experience the work in research projects were created at FCSH, allowing bachelor's and master's students to be involved. One researcher was able to incorporate some entries from *e*Viterbo as a practical assignment of a regular curriculum unit course, also at FCSH. A total of 13 students were integrated into the team, and some became co-authors of entries (for example, Francisco Mendes Esculca[13]).

The Content page[14] is the source of the structured data, containing in itself the multiple sources used by researchers. The pages were created and edited simultaneously, as the recommendation was that every piece of structured data should have a clear source on the content page. In the structure data page[15], the property P30[16] and the infoboxes establish connections between the two parts, as mentioned. Wikidata alignment and Wikipedia links were created by searching manually for matches.

After the author considered a page finished, its full content was peer-reviewed. Depending on the revisions, corrections would be made by the authors or by core team members (PI, Junior researcher, or contracted master's student). After a DOI is requested from the Universidade NOVA de Lisboa, it would be attributed and inserted on the page.

This workflow created a multitude of pages in various types of development, for example, a finished page with an attributed DOI (Luis Serrão Pimentel[17]), or pages with content only from Viterbo bibliography sources created manually (Afonso Anes[18]). This caused stress to researchers and strangeness to visitors, but ultimately, this is merely the outcome of an ongoing, open, and collaborative research process. Perhaps using banners to indicate the "status" of the page, as in Wikipedia, could address this issue.

[13] <https://eviterbo.fcsh.unl.pt/wiki/Francisco_Mendes_Esculca>.

[14] <https://eviterbo.fcsh.unl.pt/wiki/Jos%C3%A9_Ant%C3%B3nio_Caldas>.

[15] <https://eviterbo.fcsh.unl.pt/wiki/Item:Q794>.

[16] <https://eviterbo.fcsh.unl.pt/wiki/Property:P30>.

[17] <https://eviterbo.fcsh.unl.pt/wiki/Lu%C3%ADs_Serr%C3%A3o_Pimentel>.

[18] <https://eviterbo.fcsh.unl.pt/wiki/Afonso_Anes_(6)>.

Following the conclusion of TechNetEMPIRE project, a technical report was written which asserted the creation of *e*Viterbo 2.0, a possibility that had long been in discussion. Two possible solutions were identified: migrating the infrastructure to Wikibase.cloud or creating a new, enhanced installation of the Wikibase Suite and subsequently importing the data from the initial version to the latter. The next part of this text follows these steps.

2. *e*Viterbo 2.0

2.1 Exploring the possibility of moving to Wikibase.cloud

The primary reason for seeking an alternative system is the instability and constant need of maintenance of the existing one (Faria and Correia 2022, 25–30). For the moment, while MediaWiki itself is stable, Wikibase is not (yet). Moreover, the numerous extensions required for the system to function introduce additional complexity and demand continuous maintenance—an undertaking that NOVA FCSH is not equipped to support. The faculty and the affiliated research centers lack the technical resources to provide ongoing maintenance for all research systems. Given that Wikibase.cloud operates as a closed system with stability ensured by Wikimedia Deutschland, it was considered a viable alternative.

This section of the text addresses the preliminary exploration of the feasibility of migrating the *e*Viterbo project to Wikibase.cloud. It follows the steps and recommendations of the 2022 report and was first presented at the Wikidata Days in Lisbon (Faria 2024). The process involved practical trials, conceptual questions, and reflections on the documentation and support available to users.

The first step was made to create content pages, navigation pages, and basic data structures within Wikibase.cloud. At first glance, the implementation seemed promising. However, issues quickly arose—most notably, the category pages did not function as expected. It remains unclear why, but it seems to be a limitation of the system. Nevertheless, the initial outcomes suggested that a migration might be technically possible, despite several unresolved questions.

One key issue was whether storing content pages, some of which have, as mentioned, have attributed DOIs implying permanent weblinks. Wikibase.cloud uses servers from Wikimedia Deutschland and Wikibase does not use URI's, Uniform Resource Identifiers, as identifiers (Dobriy and Polleres 2022) so permanent links are probably not a priority, and that's understandable. A potential solution could involve separating content—retaining it within NOVA FCSH servers—, from structured data—hosted on Wikibase.cloud—, though this approach would introduce its own set of challenges.

A major concern throughout the implementation of *e*Viterbo, reported in 2022, was the mechanisms for ensuring data quality. *Quality Constraints* extension, which enables validation and control over the data model is possible when using a Wikibase Suite (even if, *Quality Constraints* extension never worked in *e*Viterbo system, the team was not able to understand why). Unfortunately,

Wikibase.cloud, as a closed system, does not currently support the installation of this extension. This limitation alone was difficult to verify, as the available documentation is inconsistent. The Wikibase FAQ[19] (version 16 September 2024) implied that the extension is supported, help pages link to general Wikibase documentation rather than information tailored to Wikibase.cloud[20]. In the absence of formal quality constraints, efforts were made to simulate constraint enforcement by using the *Property Constraints Portal* and studying third-party examples (Ferranti et al. 2022). Some projects on Wikibase.cloud appeared to make use of property constraints (for example, "property constraint"[21] in LGBTd), but these proved to be either misleading or incomplete. Ultimately, it became clear that without the official extension, property constraints cannot function properly. As a result, confirming this constraint required a time-consuming process of trial and error, reflecting a broader issue with documentation quality and usability for new users once again. Therefore, the approach previously used in *e*Viterbo—manual validation and the reference that an item should be used in a specific property, supported by guides[22] and documentation[23]—seems the only way possible, but it is far from ideal in terms of scalability or reliability[24].

The integration of OpenRefine as a tool for modelling and exporting data was also a recommendation (Faria and Correia 2022, 32, 34–35). This aspect was not tested in wikibase.cloud, for two main reasons: the structured data framework within Wikibase.cloud was very incomplete, and the author of this text does not have the needed familiarity with OpenRefine. Nevertheless, a review of the available documentation was conducted, and while some online articles suggest compatibility, they often fail to clearly distinguish between Wikibase and Wikibase.cloud.

A broader need for tools that easily support data manipulation and extraction seems clear. The query service is available with SPARQL, which can also be challenging to use for a humanities research team, and the process of exporting data is not explicitly outlined, nor is it clear whether this is even possible.

The need and usability posed by "empty placeholders" in Wikibase is also discussed in the report, and its importance is underlined and recommended (Faria and Correia 2022, 37, 58). In Wikibase.Cloud, *Cradle* offers a possible solution by providing structured forms for manual data entry. However, *Cradle* is not natively integrated into Wikibase.cloud, it is a tool created by Magnus

[19] <https://www.mediawiki.org/wiki/Wikibase/FAQ>.

[20] In the meantime, specific documentation for Wikibase.cloud has been created.

[21] <https://lgbtdb.wikibase.cloud/wiki/Property:P5>.

[22] <https://eviterbo.fcsh.unl.pt/wiki/Guia_de_itens>.

[23] <https://eviterbo.fcsh.unl.pt/wiki/Gui%C3%A3o_de_reda%C3%A7%C3%A3o_-_Pessoas>.

[24] This approach was also used in Katherine Harloe, Amara Thornton, James Baker, Ammandeep K. Mahal, and Sharon Howard, *Beyond 'Notability': Re-evaluating Women's Work in Archaeology, History and Heritage in Britain, 1870–1950* (https://beyond-notability.wikibase.cloud, 2021-2024). Check for example "woman" item that mentions in the description "To be used in 'assigned gender' (P3)".

Manske and functions via external connection (see the section about cradle in Wikibase/Creating and deleting data[25]). This may explain the limited success in implementing it effectively. Whether *Cradle* alone can guarantee data quality is still uncertain and would require further testing and evaluation. Nevertheless, others have already reported problems in its use (see Software Collaboration for Wikidata/Call Submissions[26]).

While the experience of moving eViterbo into Wikibase.cloud was a short one, and focused on specific aspects, it was clear that moving eViterbo was not the best solution. Nevertheless, Wikibase.cloud shows strong potential, either as a host for researchers without technical support or as a sandbox for complex projects. However, to become truly viable for projects with structured data in the humanities, the platform needs to offer: specific and targeted documentation for non-technical users in the humanities community; clear, built-in methods for ensuring data quality; and finally, user-friendly tools for data analysis, visualisation, and export. So, for the time being, the best option for *e*Viterbo is to remain a self-hosted installation of Wikibase (Wikibase Suite).

2.2 Ontology 2.0

The data model currently used in *e*Viterbo has been presented and discussed (Faria and Correia 2022, 46–57). Several issues in its ontology were pointed out, and several suggestions were made for their resolution. The main problems were that some properties were redundant, and some items should be used as qualifiers. The second version departs from there. Most of the redundancies were caused by the overlap of properties with information embedded in the item information, such as name or different spellings, overlapping with the label (name), aliases and description associated with each item. All evident cases were solved, and cases where doubts persist will be discussed. Nevertheless, it should be noted that these redundancies exist in Wikidata, and it is not that evident that this is an actual problem[27].

The objective of this section of the text is not to propose or discuss the complete ontology, but rather to provoke critical reflection on the challenges and possibilities that have been identified. Rather than offering a comprehensive representation of the entire ontology, a case study, *e*Viterbo item Q794[28], will be used as an example to describe the ontology, elucidate the complexities of the system, and discuss the alignments with Wikidata.

[25] <https://www.mediawiki.org/wiki/Wikibase/Creating_and_deleting_data>.

[26] <https://meta.wikimedia.org/wiki/Software_Collaboration_for_Wikidata/Call_Submissions>.

[27] See for example the case of Álvaro Siza Vieira, WD_ID: Q251365 where there several properties with overlapping values: "name is native language"; "birth name"; and "given name" and "family name".

[28] <https://eviterbo.fcsh.unl.pt/wiki/Item:Q794>.

It is possible to find online several documents on the basics of the Wikibase data model. Wikibase data model primer[29], presents the most basic concepts that will be used in this text: *Item* (with a Q identifier), and *Property* (with a P identifier), organising the information in triples "item-property-value" (Varvantakis 2025; to understand the differences and similarities to a RDF model see Dobriy and Polleres 2022; Shimizu et al. 2022). Values can be of several types, from string, to date, or number, and when the triple is joined by a *qualifier* and a reference, it becomes a *statement*[30]. In *e*Viterbo, the references are not included in each statement, but references are a property in the form of a URI that links to the content page. Content pages contain all the references for all values in each item.

In the following text and in the diagram presented below (Fig. 4) *e*Viterbo items and properties are presented with just a Q or a P and Wikidata equivalent item as a precedent "WD_ID:." The majority of the labels have been translated from Portuguese into English to help the understanding of the structure. When labels were very long, only the identifier was used.

The case study being analysed Q794[31] (José António Caldas), "has a Property" P15[32] (Instance of), that "has a value" Person (Q3148[33]) this equals in Wikidata to WD_ID:Q10308779[34] (José António Caldas) "has a Property" WD_ID:P31[35] that "has a value" Human WD_ID:Q5[36]. This is the most basic and important triple on Wikibase, which defines the fundamental nature of each item.

However, as shown in Figure 4, not all eViterbo items have "instances"[37]. This is because one of the considered possibilities, which was not previously recommended and has not yet been implemented, involves the use of subclasses (WD_ID: P279[38]), "parts of" (WD_ID: P361[39]), or subproperties of (WD_ID: P1647[40]) in the new ontology (see wikidata basic membership properties page[41]). This will make considerable changes to the data model, namely in categorising these items. By looking at examples such as "sex or gender" (WD_ID:

[29] <https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer>.

[30] There are many more documentation pages such as wikibase data model, which discusses more conceptually the data model, linking to other documentation about wikidata, (see for example: Wikidata/Notes/Entities and Snaks). This is a common issue experienced by beginners, who often find it challenging to identify the appropriate point of departure.

[31] <https://eviterbo.fcsh.unl.pt/wiki/Item:Q794>.

[32] <https://eviterbo.fcsh.unl.pt/wiki/Property:P15>.

[33] <https://eviterbo.fcsh.unl.pt/wiki/Item:Q3148>.

[34] <https://www.wikidata.org/wiki/Q10308779>.

[35] <https://www.wikidata.org/wiki/Property:P31>.

[36] <https://www.wikidata.org/wiki/Q5>.

[37] For a better view of this figure check the slides of the presentation at the Conference Wikidata and Research, June 2025, Florence, Italy available on Wikimedia Commons.

[38] <https://www.wikidata.org/wiki/Property:P279>.

[39] <https://www.wikidata.org/wiki/Property:P361>.

[40] <https://www.wikidata.org/wiki/Property:P1647>.

[41] <https://www.wikidata.org/wiki/Help:Basic_membership_properties>.

P21[42]), it's possible to observe the use of items such as "Wikidata property for items about people" (WD_ID: Q18608871[43]), or "curated Wikidata property" (Q121928698[44]) to establish classes or subclasses, and this would fit perfectly *e*Viterbo non-classified items about people, such as sex or gender (P6[45]) or religion (P27[46]). Classes, can also be used to classify institutions (Type of institution P36[47]) or locations, as discussed in the following paragraphs. However, the absence of a unifying structure in Wikidata poses a significant challenge in identifying specific elements. While the advantages of using classes seem clear, how to implement them is still being analysed.

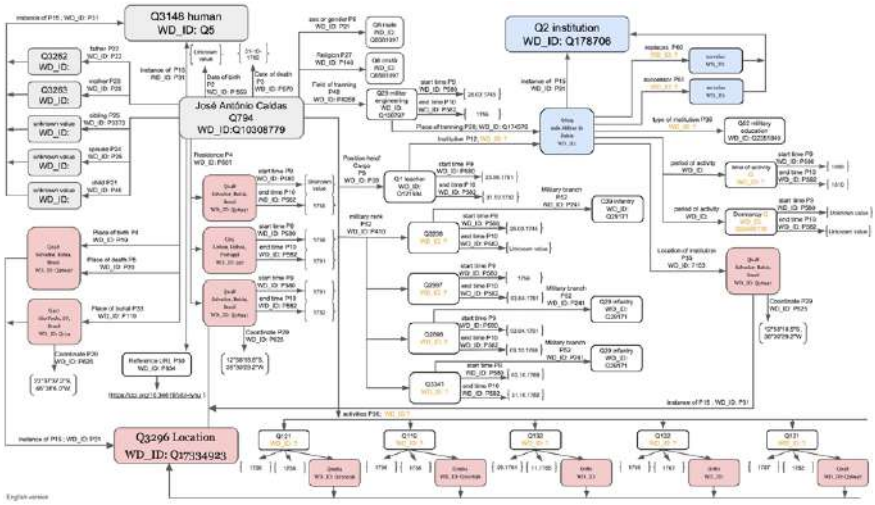


Figure 4 – Diagram showing the proposal for a new ontology using as a case study Q794. Colors indicate different "instance of": grey is person/human, blue is institution, pink is location and white has no instance.

Qualifiers[48] are used to provide additional information or refine a property's value as stated. It consists of a property and a value, which are equivalent to those employed in statements. In *e*Viterbo, the majority of qualifiers are starting or ending dates, but it can also be coordinates or other information, as in the case of military rank. For the moment, there are no restrictive qualifiers—qualifiers

[42] <https://www.wikidata.org/wiki/Property:P21>.
[43] <https://www.wikidata.org/wiki/Q18608871>.
[44] <https://www.wikidata.org/wiki/Q121928698>.
[45] <https://eviterbo.fcsh.unl.pt/wiki/Property:P6>.
[46] <https://eviterbo.fcsh.unl.pt/wiki/Property:P27>.
[47] <https://eviterbo.fcsh.unl.pt/wiki/Property:P36>.
[48] <https://www.wikidata.org/wiki/Help:Qualifiers>.

to be used in property constraints—but those will be considered (see Shimizu et al. 2022, 6–13. Additionally, certain items and properties will be required for WikibaseQualityConstraints[49] to function correctly. This was not considered at this stage, even if this is one of the most important recommendations of the 2022 report, because, first of all, it is necessary to be able to install a quality constraints extension on the Wikibase instance.

One of the most significant problems is the vast number of possibilities when searching for an alignment with Wikidata, as well as the inconsistencies within Wikidata itself (Westerinen 2023). The diagram indicates non-alignments (or the absence of a preferred alignment) or the non-existence of properties and items, which are coloured orange and marked with a question mark.

There are challenges associated with identifying equivalent items in highly specific contexts, such as old military ranks. It was not possible to find alignments from the most basic post, "Praça", to some of the highest and particular items such as "Sargento-mor com exercício de Engenheiro"; this is normal, and these items are easy to create. Another case is properties. It was not possible to find an equivalent of "activity" (P56[50]), the closest found were "field of work" (WD_ID: P101[51]), which in the case study being used, would be "military engineering" or "occupation" (WD_ID: P106[52]), which would be "military engineer" and "teacher." None of these, however, is sufficient to define everyday or specific tasks, such as "construction site supervision", "Cartographic Drawing" or "Border demarcation." Properties[53] have to be proposed[54] and approved by the Wikidata community, so this is also a step that needs to be taken.

There are, however, some properties of which the significant challenge came as a surprise[55]. It is the case of "institution" which exists as an item in Wikidata but not as a property. In *e*Viterbo, it is used as a property which states the organisation in which someone has worked or studied (P12[56]). Also, type of institution[57],

[49] <https://gerrit.wikimedia.org/g/mediawiki/extensions/WikibaseQualityConstraints/>.

[50] <https://eviterbo.fcsh.unl.pt/wiki/Property:P56>.

[51] <https://www.wikidata.org/wiki/Property:P101>.

[52] <https://www.wikidata.org/wiki/Property:P106>.

[53] <https://www.wikidata.org/wiki/Help:Properties>.

[54] <https://www.wikidata.org/wiki/Wikidata:Property_proposal>.

[55] It seems institutions are not being discussed as a whole, but in separate projects. See: Wikiproject Heritage institutions is discussing a data structure, https://www.wikidata.org/wiki/Wikidata:WikiProject_Heritage_institutions/Data_structure. However, there are several other wikiprojects which touch on this, such as https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Organizations or https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Higher_education. In the Heritage institution project, it is possible to observe that a dormancy period, a successor, or a predecessor are not being used in this structure and do not exist as properties in Wikidata, as far as I could tell. Instead, they are described as "follows" and "replaces" or "replaces by". This is probably due to the structure being built for existing institutions rather than historical ones. Using historical examples to develop the data structure can actually help in understanding potential future needs.

[56] <https://eviterbo.fcsh.unl.pt/wiki/Property:P12>.

[57] <https://www.wikidata.org/wiki/Q117023459>.

which was mentioned before as a probable property (P36[58]) to be transformed into a subclass in *e*Viterbo, is an item at Wikidata (Wd_ID: Q117023459[59]) that is used as a metaclass. A metaclass[60] is a class that contains items which are all classes themselves, so probably institution should be a metaclass[61].

Locations are also a big challenge, but surprisingly not on the alignment of places/locations, but in the modelling itself. In *e*Viterbo location have the following rule, which allows the researchers to immediately understand what the item represents by their description: Lisboa, Lisboa, Portugal (Q25[62]) represents the city of Lisbon in Portugal (WD_ID:Q597[63]); Lisboa, Portugal (Q204[64]) the district of Lisbon (WD_ID: Q207199[65]), Portugal (Q310[66]) is the country (WD_ID: Q45[67]). The majority of these administrative locations are not difficult to align. However, *e*Viterbo also has more detailed locations, such as streets or buildings, that rarely exist on Wikidata, such as "Instituto Industrial de Lisboa, Lisboa,-" (Q198[68]), which is the building of the old Lisbon Industrial Instituto, in Lisboa. The "-", means "which follows" that is, located in the city of Lisbon, district of Lisbon, Portugal. When defining the "instance of" for location items in *e*Viterbo "physical location" (WD_ID: Q17334923[69]) was chosen since its descriptions seemed to be the most open "property describing the position of something in space " In Portuguese the description says "*localização de um objeto* (físico ou abstrato)" which means "location of an object (physical or abstract)" not the same thing, but very open nonetheless. However, the use of classes and subclasses should also be considered, as in Wikidata the concept of location is subdivided and in *e*Viterbo that also happens (Place of Birth; Place of Death; Place of Buriel; Location of institution etc.), even if in 2022 the proposal was to replace all these location by one single property and use it as a qualifier.

The debate surrounding the choices and alignments in the ontology could continue, however, the most crucial point for this particular text has already been demonstrated: the modelling of data in Wikibase is not straightforward. If an alignment with Wikidata is wanted, it adds a degree of complexity, even if

[58] <https://eviterbo.fcsh.unl.pt/wiki/Property:P36>.

[59] <https://www.wikidata.org/wiki/Q597>.

[60] <https://www.wikidata.org/wiki/Q19478619>.

[61] The use of metaclasses is under discussion, raising a lot of questions since it is subdocumented. The WikiProject Ontology Group has assumed the task of drafting documentation. See: https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology/Modelling within this scope, I proposed a project discussing institutions.

[62] <https://eviterbo.fcsh.unl.pt/wiki/Item:Q25>.

[63] <https://www.wikidata.org/wiki/Q597>.

[64] <https://eviterbo.fcsh.unl.pt/wiki/Item:Q204>.

[65] <https://www.wikidata.org/wiki/Q207199>.

[66] <https://eviterbo.fcsh.unl.pt/wiki/Item:Q310>.

[67] <https://www.wikidata.org/wiki/Q45>.

[68] <https://eviterbo.fcsh.unl.pt/wiki/Item:Q198>.

[69] <https://www.wikidata.org/wiki/Q17334923>.

aligning some specific items and properties with Wikidata is not that difficult. Moreover, the existing literature is often too complex, making it inaccessible to non-expert readers (for example, Shimizu et al., 2024). The difficulty of constructing a coherent and usable ontology, aligned with Wikidata, is a significant obstacle to the use of Wikibase, particularly in the context of the humanities. Nevertheless, some initiatives in Wikimedia communities, such as WikiProject Ontology [70]can play an essential role in making it more accessible.

3. In conclusion

As previously stated, it was determined that *e*Viterbo will continue to be a self-hosted installation in the future. The new version will be installed from the start on a shared hosting machine at NOVA FCSH. The biggest challenge behind this option will be overcoming the difficulties of installing specific extensions and fundamental modules. However, it will facilitate system maintenance. To address the issue, it was decided to involve the *Digital Development of Research Nucleus* (NDDI) in the development of the second version of *e*Viterbo. The role of this unit at NOVA FCSH, is not to support the development of digital research but to support research units' digital infrastructures, which is quite different, and to define policies for website hosting, and to define research data management policies. Nevertheless, the unit's technical staff has been open to participation[71]. Having a stable team that knows the system and how to maintain it is essential to the existence of a self-hosted Wikibase instance. In this case, it implies the training and transfer of knowledge from the former "team" to a new one. It will involve costs and time, but hopefully, it can result in the use of Wikimedia software as a default option or one of the options for other projects at NOVA FCSH, so that knowledge and resources are not wasted. It can also potentiate the use of federated wikibases[72] between research teams. This means sharing information between Wikibase instances, which has considerable potential for the academic community.

The transformation of the infrastructure, along with the development of a new ontology, is a complex process, as illustrated in the pages of this chapter, which greatly benefits from collaborative input and methodological discussions. This is a role already played by Wikimedia chapters and communities, which are making an effort to bridge the distance between the software and the non-technical user, and to create connections between the Wikimedia universe and Academia. The support of Wikimedia Portugal has been essential to *e*Viterbo development. A deeper commitment from researchers with the Wikimedia movement is also crucial. The goal of *e*Viterbo is to be replicable by other research teams from the

[70] <https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology>.

[71] My acknowledgements to Rui Araújo and Daniel Monteiro from NDDI, as well as to Tiago Mauricio, who has been accompanying this process.

[72] <https://www.mediawiki.org/wiki/Wikibase/Federation>.

humanities. To achieve this, a simplified, comprehensive model needs to be developed. Documenting the difficulties and choices, as done in this chapter, is an essential part of that path and a way to contribute to this collaborative effort.

References

Dobriy, Daniil, and Axel Polleres. 2022. "Analysing and Promoting Ontology Interoperability in Wikibase." *Proceedings of the 3rd Wikidata Workshop 2022*, edited by Lucie-Aimée Kaffee, Simon Razniewski, and Kholoud Saad Alghamdi, 3262:7. CEUR-WS. <http://ceur-ws.org/Vol-3262/>

Faria, Alice Santiago (Coord.), Araujo, Renata Malcher de, Pacheco, Mafalda Batista, and Sandra MG Pinto (eds). 2022. eViterbo. *Encyclopaedic platform online*. CHAM-Centro de Humanidades, Faculdade de Ciências Sociais e Humanas, Universidade NOVA de Lisboa. eISBN: 978-989-8492-77-7. <https://eviterbo.fcsh.unl.pt/>

Faria, Alice Santiago, and Rute Correia. *eViterbo Global Report. Outlines on the Creation of a Research Open Platform Using MediaWiki and Wikibase*. Zenodo, 30 de novembro de 2022. <https://doi.org/10.5281/zenodo.7380115>

Faria, Alice Santiago. 2024. "*e*Viterbo 2.0. Challenges in moving into wikibase cloud." Wikidata days presentation. Lisbon. <https://doi.org/10.5281/zenodo.15375849>

Ferranti, Nicolas, Axel Polleres, Jairo Francisco de Souza, and Shqiponja Ahmetaj. 2022. "Formalizing Property Constraints in Wikidata", CEUR Workshop Proceedings (CEUR-WS.org), October.

LGBTdb. 2025. "property constraint (P5)." 15 April. <https://lgbtdb.wikibase.cloud/wiki/Property:P5> (accessed April 15, 2025).

Lindemann, David. 2025. "FAIR Data on Wiki-Platforms." Presentation, Lisbon, Portugal, January 14 (accessed January 14, 2025). <https://commons.wikimedia.org/entity/M158467443>

MediaWiki. 2024. "Wikibase/DataModel." 1 December (accessed December 1, 2024). <https://www.mediawiki.org/wiki/Wikibase/DataModel>

MediaWiki. 2024. "Wikibase/DataModel/Primer." 8 October (accessed October 8, 2024). <https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer>

MediaWiki. 2024. "Wikibase/Federation." 15 November (accessed November 15, 2024). <https://www.mediawiki.org/wiki/Wikibase/Federation>

MediaWiki. 2025. "Wikibase/Creating and deleting data." 19 April (accessed April 19, 2025). <https://www.mediawiki.org/wiki/Wikibase/Creating_and_deleting_data#Cradle>

Ribeiro, José Silvestre. 1872–1914. *Historia dos estabelecimentos Scientificos Litterarios e Artisticos de Portugal nos Sucessivos Reinados da Monarquia*. 19 vols. Lisboa: Typografia Real da Academia de Sciencias.

s.a. n.d. "WikibaseQualityConstraints." <https://gerrit.wikimedia.org/g/mediawiki/extensions/WikibaseQualityConstraints/#installation> (accessed April 1, 2025).

Shimizu, Cogan, Andrew Eells, Seila Gonzalez, Lu Zhou, Pascal Hitzler, Alicia Sheill, Catherine Foley, and Dean Rehberger. 2024. "Ontology design facilitating Wikibase integration — and a worked example for historical data." Journal of Web Semantics 82 (October):100823. <https://doi.org/10.1016/j.websem.2024.100823>

Varvantakis, Christos. 2025. "The Wikibase Software: Data and Collections in the Linked Open Data Web." January 11. <https://zenodo.org/doi/10.5281/zenodo.14655779> (accessed January 11, 2025).

Viterbo, Francisco Marques de Sousa. 1892. *Artes e artistas em Portugal: contribuições para a sua história das artes e indústrias portuguezas*. Lisboa: Livraria Ferreira.

Viterbo, Francisco Marques de Sousa. 1898. *Arquitétos e engenheiros militares portugueses ou a Serviço de Portugal*. Lisboa: Minerva Commercial.
Viterbo, Francisco Marques de Sousa. 1899. *Diccionario historico e documental dos architectos, Engenheiros e constructores portuguezes ou a...* 3 vols. Lisboa: Impr. Nacional. <http://archive.org/details/diccionariohist01lisbgoog>
Viterbo, Francisco Marques de Sousa. 1903. *Notícia de alguns pintores portugueses e de outros que, sendo estrangeiros, exerceram a sua arte em Portugal*. Lisboa: Typographia da Academia Real das Sciencias de Lisboa.
Viterbo, Francisco Marques de Sousa. 1917. *Artífices Portugueses ou domiciliados em Portugal*. Coimbra: Imprensa da Universidade.
Westerinen, Andrea and Lydia Pintscher. 2023. "Wikidata Challenges in the Semantic Web Community." Presentation Wikidata Modelling Days. December 2. <https://commons.wikimedia.org/w/index.php?title=File:Wikidata_Challenges_in_Semantic_Web_Community.pdf&page=6> (accessed [date]).
Wikidata. 2024. "Help:Basic membership properties." 11 April (accessed April 11, 2024). <https://www.wikidata.org/wiki/Help:Basic_membership_properties>
Wikidata. 2024. "Wikidata:WikiProject Ontology." 12 August (accessed August 12, 2024). <https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology>
Wikidata. 2025. "Álvaro Siza Vieira (Q251365)." 3 April (accessed April 3, 2025). <https://www.wikidata.org/wiki/Q251365>
Wikidata. 2025. "Help:Qualifiers." 15 April (accessed April 15, 2025). <https://www.wikidata.org/wiki/Help:Qualifiers>
Wikimedia. Meta-Wiki. 2012. "Wikidata/Notes/Entities and Snaks." 14 August (accessed August 14, 2012). <https://meta.wikimedia.org/wiki/Wikidata/Notes/Entities_and_Snaks>
Wikimedia. Meta-Wiki. 2024. "Software Collaboration for Wikidata/Call Submissions." 15 July (accessed July 15, 2024). <https://meta.wikimedia.org/wiki/Software_Collaboration_for_Wikidata/Call_Submissions>

Un'ontologia per il patrimonio culturale teatrale su Wikibase.cloud

Donatella Gavrilovich, Giovanni Bergamin, Valeria Paraninfi

Abstract: The paper focuses on the innovative "Hyperstage" knowledge base ontology model for the semantic reconstruction of Italian theatrical productions. Innovation combines (1) a bottom-up ontology on Wikibase.cloud for the organization and valorization of metadata with (2) an integrated PKB taxonomy based on the three different categories of a) conception, b) staging, and c) post-production documentation. This framework allows automatic aggregation of each digital object associated with the theatrical production into one of these phases. The project aims to transcend the traditional limits of theatrical data models through a system of ontological hierarchical relationships to trace the evolutionary path of works and identify influences and connections between different productions.

Keywords: Theatrical productions, Digital objects, Bottom-up ontology, Wikibase data model, Hyperstage.

1. Introduzione

Le basi di dati e di conoscenza, dedicate alle arti dello spettacolo e concepite sui principi dell'*Open Science*, sono oggi ampiamente presenti in rete. Alcuni dei prodotti più significativi sono AusStage (2003), il più grande *gateway* digitale al mondo, il portale francese Les Archives du Spectacle (2007), ECLAP (European Collected Library of Artistic Performance, 2010), Swiss Performing Arts Platform (2012), Frankfurt Performing Arts for theatre and dance studies (2015), il National Portal of Theatre Museums of Russia (2019), Digital Pina Bausch Archive (2020).

AusStage è stato il modello per la creazione della piattaforma norvegese IbsenStage, lanciata nel 2014 (Hanssen 2018, 24), così come ECLAP lo è stato per le successive realizzazioni degli archivi digitali europei, soprattutto per quanto riguarda la struttura ontologica e la scheda digitale dei metadati progettati sulla base di standard internazionali. Ciò nonostante, ECLAP non ha soddisfatto la committenza e il sito è stato disattivato dopo pochi anni. Il limite di ECLAP è dovuto proprio alla sua struttura ontologica, che raccoglie e suddivide, per categorie di appartenenza, gli elementi eterogenei che costituiscono l'evento teatrale. Essa obbliga l'utente ad avviare sempre una nuova ricerca, perché il sistema

Donatella Gavrilovich, University of Rome Tor Vergata, Italy, gavrilovich@lettere.uniroma2.it, 0000-0002-6430-4164
Giovanni Bergamin, Logos RI, Italy, giovanni.bergamin@logos-ri.eu, 0000-0002-2912-5662
Valeria Paraninfi, University of Rome Tor Vergata, Italy, valeria.paraninfi@uniroma2.it, 0009-0001-2738-9947

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Donatella Gavrilovich, Giovanni Bergamin, Valeria Paraninfi, *Un'ontologia per il patrimonio culturale teatrale su Wikibase.cloud*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.18, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 131-146, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

fornisce la connessione dati senza un ordine chiaro e di facile utilizzo. Inoltre, la scheda dei metadati è molto concisa e non esiste alcuna connessione con altri elementi dello stesso evento.

In Italia non esiste né un museo nazionale teatrale, né una piattaforma che colleghi tutti gli archivi teatrali digitali esistenti, pubblici e privati, per conoscere, condividere e valorizzare il patrimonio dei beni teatrali delle singole istituzioni italiane. Studiosi, archivisti e professionisti delle arti dello spettacolo hanno manifestato la necessità di creare una piattaforma comune per esigenze di ricerca e di lavoro (Gavrilovich 2020). Il Ministero della Cultura (MiC) ha accolto questa esigenza a suo modo e nel dicembre 2020 ha pubblicato online "Gli Archivi dello Spettacolo", un database che raccoglie i metadati essenziali del contenuto di archivi, posti sotto la sua tutela, ordinandoli per regioni. Si tratta di una sorta di inventario, che nasce dalla necessità del MiC di conoscere ciò che tutela come «patrimonio di particolare rilevanza storica», ma che non risponde ai diversificati fabbisogni dell'utenza.

2. Il progetto Hyperstage: origini e obiettivi per una nuova base di conoscenza

Catalogare lo spettacolo teatrale significa non solo raccogliere in modo ordinato, secondo gli standard vigenti, un'enorme quantità di dati eterogenei riguardanti documenti e materiali (ad es. copioni, costumi, bozzetti di scena, figurini, spartiti, fotografie, note di regia ecc.), ma anche creare un collegamento logico tra essi, avendo come scopo la ricostruzione filologica dell'unità storico-artistica e socio-culturale di una rappresentazione teatrale che non esiste più.

Per raggiungere questo risultato bisogna partire dalla concezione unitaria dello spettacolo teatrale, che deve essere considerato come un'opera d'arte totale, allo stesso modo di un'opera d'arte figurativa complessa qual è, ad esempio, un polittico. Per questo motivo nella creazione della struttura ontologica di una base di conoscenza dedicata alle arti dello spettacolo è necessario procedere in modo che gli elementi costituenti di quel tutto, che non esiste più, tornino a 'dialogare' come in origine, incrociando i dati (Gavrilovich 2017, 28-39).

Partendo da questi concetti e riflessioni, è stato progettato nel 2014 un prototipo per una base di conoscenza *Performance Knowledge base* (PKB) dedicata alla organizzazione dei dati delle produzioni teatrali (Gavrilovich 2017, 28-39). Dieci anni dopo questa sperimentazione è confluita nel progetto PRIN 2022 *HYPERSTAGE: an Open Knowledge base for the semantic reconstruction of theatrical performances through the harvesting and processing resources from the New Italian Network of theatrical digital archives*[1].

[1] Le unità di ricerca del progetto Hyperstage sono: Unità di Ricerca 1 [Università di Roma Tor Vergata, Associazione Logos-ri di Firenze e Consortium GARR di Roma], Unità di Ricerca 2 [Università di Bologna], Unità di Ricerca 3 [Accademia di Belle Arti di Venezia]. I partner fornitori di contenuti sono: Teatro Stabile di Torino, Teatro alla Scala di Milano, Accademia Olimpica di Vicenza, Fondazione Romaeuropa e Fondazione INDA di Siracusa. Donatella Gavrilovich è il Principal Investigator.

La principale innovazione è la metodologia del progetto Hyperstage, basata su competenze e conoscenze umanistiche. Essa ha permesso di affrontare e risolvere il problema, finora irrisolto da parte degli esperti IT (Diwisch e Thull 2014) (Estermann 2017; 2020) (Beck e Voß 2020) (Weiberg 2020), della contestualizzazione storica, sociale, politica, artistica e culturale delle informazioni inerenti a una certa rappresentazione teatrale, così come dell'organizzazione dell'enorme quantità dei suoi dati eterogenei.

Applicando le metodologie solitamente in uso nell'ambito degli studi teatrali (De Marinis 2008), è stato possibile trovare la giusta soluzione che si basa, da un lato, sull'organizzazione dei dati secondo le tre fasi fondamentali della produzione di un'opera teatrale (concezione, realizzazione e post-produzione); dall'altro, sulla raccolta strutturata per tipologie dei materiali che sono stati ordinati in due raggruppamenti distinti (Gavrilovich 2013; 2017; 2020). Il primo gruppo include tutti gli oggetti digitali caratterizzati da immagini; il secondo gruppo include quelli caratterizzati da testo. Nei casi dubbi, ad esempio una locandina con testo e immagine, il criterio da seguire è la rilevanza dell'uno rispetto all'altro.

Hyperstage introduce un'importante innovazione nel mondo Wiki: la classificazione PKB grazie alla quale è possibile gestire la grande quantità e l'eterogeneità delle risorse digitali e di cui parleremo in seguito.

3. Preservare l'effimero: un'ontologia per le arti teatrali

3.1 Necessità di una ontologia

Per creare una base di conoscenza aperta (Open knowledge base) dedicata alla rappresentazione semantica degli spettacoli teatrali, è necessario strutturare i dati utilizzando un'ontologia di dominio compatibile con il Web semantico. In informatica, un'ontologia definisce le entità, i loro attributi e le relazioni tra le entità, favorendo l'interoperabilità e la condivisione della conoscenza in un dominio specifico[2].

Quando si presenta una ontologia può essere d'aiuto ricordare la fondamentale distinzione tra il livello TBox e il livello ABox (Nardi e Brachman 2010, 16-20). Il livello TBox (dove «T» sta per «terminological») definisce il vocabolario dell'ontologia: le entità di un determinato dominio (classi e proprietà) e i vincoli che regolano i collegamenti tra queste entità. Possiamo dire anche che il livello TBox si occupa dello schema o del modello dei dati che decidiamo di trattare. Il livello ABox (dove «A» sta per «assertional») si occupa invece dei 'dati' che popolano il modello o lo schema definito nel livello TBox. Gli elementi che caratterizzano questo livello sono gli individui (o le specifiche istanze) delle classi definite nel TBox, le caratteristiche e le relazioni specifiche tra questi individui.

Il dominio teatrale manca di un'ontologia consolidata e generalmente adottata. Recenti studi esplorano lo sviluppo di ontologie per le arti dello spettacolo

[2] Questa definizione è liberamente tratta da Guarino, Oberle e Staab 2009.

(performing arts). Ad esempio Mitsopoulou, Kyprianos e Brattis (2024) analizzano modelli come: lo Swiss Performing Arts – SPA – data Model (Estermann e Schneeberger 2017); l'iniziativa Linked digital future della Canadian Arts Presenting Association, CAPACOA (Estermann e Julien 2019); il Wikiproject Performing arts in ambito Wikidata[3]. Un altro studio (Skaug e Aalberg 2024) – dedicato ai modelli concettuali collegati alla documentazione degli spettacoli teatrali (theater performances) e in particolare al concetto di opera (concept of work) – prende in considerazione anche le strutture di metadati che non si basano su un'ontologia formale – come ad esempio IbsenStage.

Per favorire l'interoperabilità, framework generalisti come Wikidata e Schema.org sono importanti, ma non specifici per il teatro. In particolare Schema.org è stato creato per essere un vocabolario minimo comprensibile dai motori di ricerca, mentre Wikidata si propone di essere un nodo centrale (un hub) tra ontologie di dominio. La principale differenza tra Schema.org e Wikidata è che la prima è una ontologia definita a livello di TBox da una comunità basata su un accordo tra i principali motori di ricerca[4], mentre Wikidata – attiva dal 2012 – è sia una base di conoscenza aperta, sia una ontologia basata sulla collaborazione (Samuel 2017): in altre parole sia il livello TBox che il livello ABox vengono sviluppati mettendo in pratica il modello di collaborazione che – dal 2001 – caratterizza Wikipedia.

3.2 Lo sviluppo bottom-up su Wikibase.cloud

Per modellare l'ontologia e gestire i dati, Hyperstage adotta Wikibase.cloud[5], una piattaforma basata su Wikibase, il software di Wikidata. Rispetto ad altri strumenti (es. Protégé), Wikibase.cloud offre semplicità e integrazione con l'ecosistema Wikibase[6], con vantaggi come: facilità d'uso, accessibile anche ai non esperti; trasparenza, con modifiche tracciabili e versioni ripristinabili; gratuità d'uso; efficienza, con costi di sviluppo e manutenzione ridotti per le applicazioni esterne che sfruttano le API di Wikibase. La piattaforma consente di definire l'ontologia mentre si inseriscono dati, caricare metadati tramite API, eseguire query semantiche con SPARQL e creare applicazioni utente (es. portali pubblici) tramite API. Nella pratica, partendo dai metadati forniti dai partner del progetto, le informazioni vengono organizzate e caricate nell'istanza Wikibase.cloud di Hyperstage. Parallelamente si svolge anche un'attività di sviluppo e di manutenzione bottom-up dell'ontologia che integra un mapping semantico a livello di TBox (allineamento dove possibile delle classi e proprietà con Wikidata e Schema.org) e un mapping a livello di ABox (riconciliazione delle istanze tramite identificatori univoci – es. URI Wikidata – per entità che sono già presenti in altre basi di conoscenza).

[3] <https://www.wikidata.org/wiki/Wikidata:WikiProject_Performing_arts>.

[4] <https://schema.org/docs/about.html>.

[5] <https://www.wikibase.cloud/>.

[6] <https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy2021>.

3.3 Le entità principali

Un punto di partenza per il progetto Hyperstage è l'osservazione di una significativa convergenza nella organizzazione dei metadati nelle banche dati che documentano gli spettacoli teatrali:

> The fundamental characteristic that all these databases have in common is that they somehow describe the performing aspect (production, event, performance, listing) and what was performed (play, work, original, original work, script). This can be compared to the way Holden describes the notion that creative activity has two components: the idea and the carrier. In this case: what the performance performs and the performance itself. When the databases represent the performing-aspect, they usually use the production level (Skaug e Aalberg 2024, 8).

Si tratta della distinzione tra «opera» intesa come contenuto intellettuale e le specifiche istanze di «spettacolo» che mettono in scena quell'opera. Il riferimento a una identità comune migliora l'organizzazione dei dati – in quanto crea connessioni di tipo semantico – e favorisce l'interoperabilità. Un punto di confronto chiave per l'ontologia Hyperstage è l'IFLA Library Reference Model (2017)[7] che definisce l'opera come «the intellectual or artistic content of a distinct creation». Questo modello consolida il lavoro iniziato nel 1998 con FRBR (Functional Requirements for Bibliographic Records). Questo approccio è stato adottato anche dal gruppo *Wikidata:WikiProject Performing Arts*[8].

In Figura 1 si possono trovare le principali classi e proprietà che fondano l'ontologia Hyperstage:

- <https://hyperstage.wikibase.cloud/entity/Q1>. Spettacolo (theatrical production): «rappresentazione di un evento o serie di eventi quasi identici inclusa la loro produzione complessiva»,
- <https://hyperstage.wikibase.cloud/entity/Q2>. Opera creativa teatrale (theatrical creative work): «specifica creazione artistica orientata allo spettacolo teatrale» – il riferimento è all'entità LRM-E2 «the intellectual or artistic content of a distinct creation»,
- <https://hyperstage.wikibase.cloud/entity/Q1445>. Agente: superclasse comprendente le sottoclassi Organizzazione (Q5) e Persona (Q6),
- <https://hyperstage.wikibase.cloud/entity/Q1355>. Risorsa documentale (documentary resource),
- <https://hyperstage.wikibase.cloud/entity/Q12>. PKB: Classificazione della risorsa secondo la tassonomia PKB,
- <https://hyperstage.wikibase.cloud/entity/P6>. Messinscena di (staging of): proprietà che collega Q1 a Q2,

[7] <https://tinyurl.com/IFLALRM2017>. Si veda anche LRMoo, versione 1, 2024: <https://tinyurl.com/LRMoo2024>.

[8] <https://www.wikidata.org/wiki/Wikidata:WikiProject_Performing_arts/Data_structure>.

- <https://hyperstage.wikibase.cloud/entity/P165>. Contributi (contribution by): superproprietà che raccoglie le specifiche proprietà indicanti tutti i contributi che un Agente (Q1445) fornisce a Q1 o a Q2,
- <https://hyperstage.wikibase.cloud/entity/P110>. Documentato da (documented by): proprietà che collega Q1 a Q1355,
- <https://hyperstage.wikibase.cloud/entity/P14>. Classificazione PKB (PKB classification): proprietà che collega Q1355 a Q12.

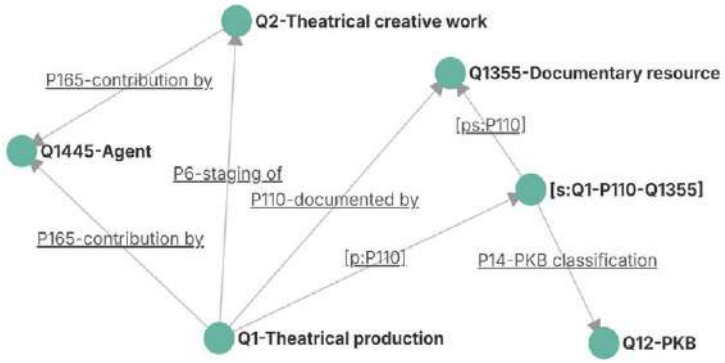


Figura 1 – Grafo delle classi e delle proprietà (relazioni) principali dell'ontologia Hyperstage.

3.4 Conflazione e gestione delle ambiguità

Durante il processo di sviluppo bottom-up dell'ontologia Hyperstage, è emersa una problematica significativa concernente la conflazione, ovvero la fusione in un unico item di entità concettualmente distinte. Sebbene tale pratica possa presentare vantaggi in contesti specifici, si rivela controproducente nella costruzione di ontologie complesse come Hyperstage. La conflazione di istanze, infatti, introduce ambiguità semantiche che compromettono l'usabilità e l'interoperabilità dell'ontologia.

Come argomentato da Guarino e Welty (2002) nell'ambito della valutazione ontologica con OntoClean, la mancata distinzione tra entità concettualmente separate porta a una perdita di precisione semantica e a potenziali incoerenze all'interno del modello. In particolare, abbiamo riscontrato una potenziale conflazione problematica nella distinzione tra l'opera originale e le sue derivazioni. Queste ultime rappresentano le diverse produzioni che si sono succedute nel tempo.

Per illustrare la complessità della gestione dei metadati relativi alla proprietà «coreografo (P42)», si consideri il caso del balletto *Giselle*. La coreografia, elemento intrinseco e fondamentale dell'opera coreutica, riveste un ruolo di primaria importanza nella sua definizione.

Nel caso specifico di *Giselle*, capolavoro del periodo romantico debuttato nel 1841, le coreografie originali di Jean Coralli e Jules Perrot hanno esercitato un'influenza significativa sulla storia della danza. Tuttavia, la persistente popolarità dell'opera ha portato, nel corso del tempo, alla creazione di numerose nuove interpretazioni e riallestimenti. Queste produzioni presentano spesso revisioni coreografiche che riflettono l'evoluzione delle tecniche e delle sensibilità artistiche contemporanee.

Tale evoluzione solleva una questione metodologica: come gestire in modo accurato e disambiguo le informazioni relative ai coreografi delle diverse rappresentazioni all'interno di un sistema informativo o di una piattaforma dedicata?

Una semplice attribuzione sistematica di Jean Coralli e Jules Perrot come coreografi di ogni rappresentazione successiva al 1841 condurrebbe a un problema di conflazione. In questo contesto la conflazione si manifesta come un'attribuzione errata o ambigua di informazioni, in quanto l'associazione esclusiva di Coralli e Perrot oscurerebbe il contributo specifico dei coreografi che hanno apportato naturali rielaborazioni all'opera originale. Tale pratica genererebbe un'informazione distorta e non rappresentativa della reale paternità coreografica delle produzioni successive.

Per garantire una descrizione precisa e accurata della paternità coreografica di *Giselle* nelle sue diverse manifestazioni, è importante fare una chiara distinzione tra la coreografia originale di Coralli e Perrot e i successivi riallestimenti (o adattamenti).

3.5 Gerarchizzazione delle produzioni teatrali: un modello per la gestione dei metadati

Al fine di prevenire fenomeni di conflazione e ambiguità, Hyperstage adotta un modello di gerarchizzazione delle opere. Questa decisione si rivela cruciale per mantenere una distinzione netta tra le diverse opere, prevenendo l'erronea fusione tra entità in un unico item. Infatti, attraverso la definizione di livelli gerarchici, è possibile stabilire relazioni chiare e non ambigue, facilitando l'analisi, l'origine e l'evoluzione delle diverse componenti di una produzione teatrale.

Lo schema in Figura 2 illustra la struttura gerarchica delle produzioni teatrali, evidenziando le relazioni tra le diverse fasi e versioni delle produzioni teatrali.

Al fine di stabilire definizioni univoche prima di procedere con la descrizione dello schema gerarchico, è necessario affrontare la complessa nozione di 'opera'. Come altri modelli di catalogazione delle arti performative, Hyperstage si confronta con questa sfida, per la quale il concetto di *Gesamtkunstwerk* (opera d'arte totale) offre un utile parallelismo. Analogamente a Wagner, che teorizzava l'unione di musica, teatro e scenografia in un'unica opera, Hyperstage mira a fornire agli utenti un'esperienza di consultazione unitaria. Tuttavia, la definizione di 'opera' in Hyperstage richiede ulteriori chiarimenti, data la sua intrinseca complessità nelle arti performative, come evidenziato da Skaug e Aalberg (2024). Un altro interessante punto di riferimento nel panorama scientifico è fornito da Mitsopoulou, Kyprianos e Brattis (2024). In questo lavoro, gli auto-

ri sottolineano l'importanza delle strutture gerarchiche per l'organizzazione della conoscenza, un aspetto cruciale anche per Hyperstage. Come evidenziano Mitsopoulou, Kyprianos e Brattis (2024, 6), la tassonomia, uno strumento fondamentale per la classificazione, è intrinsecamente una gerarchia di termini. L'articolo riporta infatti la definizione di tassonomia come «taxonomy as a hierarchy of terms that denote object types», il che implica una struttura teorica di termini organizzata in un grafo, con nodi rappresentati dai termini e archi che ne esprimono le relazioni. Questa struttura si sviluppa a partire da un singolo termine, la radice della gerarchia, evidenziando come la classificazione gerarchica sia essenziale per ordinare e strutturare le informazioni in modo chiaro e logico, un principio che è stato tenuto presente nello sviluppo di Hyperstage. Per garantire la precisione nella gerarchizzazione di Hyperstage, è necessario chiarire la terminologia fondamentale:

- Opera creativa. Si tratta di un concetto astratto che si riferisce a una specifica opera teatrale 'generatrice', dalla quale si sono sviluppate tutte le altre. Questa 'opera generatrice' rappresenta l'ideazione originale, comprensiva di elementi fondamentali quali il soggetto, il compositore, l'autore, il coreografo. Ad esempio, il balletto *Giselle*, la cui 'opera creativa' è ricondotta al libretto originale di Jules-Henri Vernoy de Saint-Georges, Théophile Gautier e Jean Coralli, con le musiche di Adolphe Adam, coreografia originale di Jean Coralli e Jules Perrot (1841). In Hyperstage, l'opera creativa di *Giselle* funge da punto di origine per tutte le successive messe in scena, produzioni e adattamenti del balletto.
- Produzione. Si intende l'insieme degli spettacoli che condividono un allestimento scenico, costumi, impostazione musicale e impianto coreografico sostanzialmente identici. Questo concetto si riferisce alla collaborazione creativa di un gruppo di artisti e tecnici (regista, scenografo, costumista, musicista, coreografo, interpreti, illuminotecnica ecc.) per realizzare una specifica interpretazione di un'opera.
- Rappresentazione. Con questo termine si intende una specifica messinscena, ovvero la rappresentazione di un'opera teatrale o performativa in un determinato luogo e momento. Ogni spettacolo è un evento unico e irripetibile, anche se fa parte di una stessa produzione.

Alla luce di quanto esposto, è possibile individuare tre livelli di gerarchizzazione:

1) Primo livello. «Primary work | Creazione Artistica» (in blu scuro): rappresenta la concezione originale (generatrice) della produzione teatrale. È il punto di partenza da cui si diramano tutte le successive elaborazioni e adattamenti.
2) Secondo livello «Adapted Work | Adattamento» (in viola): si configura come una nuova versione della Creazione Artistica [Primary work], in cui uno o più elementi costitutivi della stessa (come la regia, la scenografia, la coreografia, i costumi) vengono rielaborati da altri artisti o da altri professionisti. L'Adattamento si configura come un'opera autonoma, distinta dalla Creazione Artistica [Primary work] proprio per le modifiche apportate.

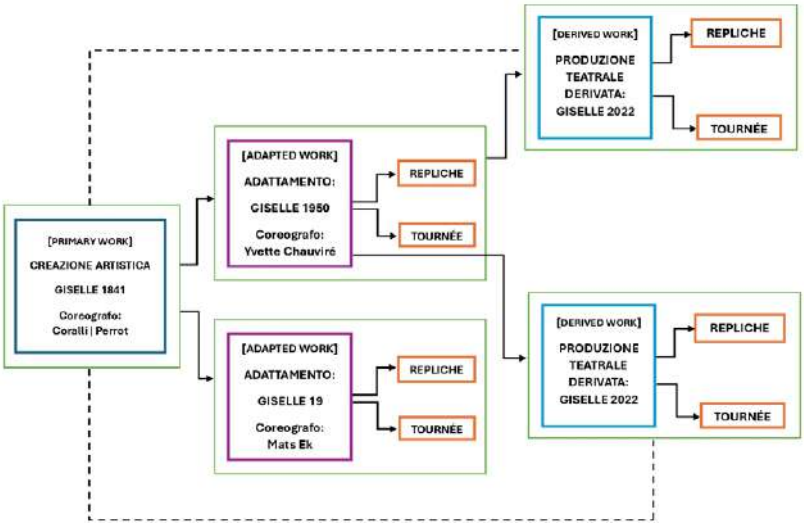


Figura 2 – Schema gerarchico delle produzioni teatrali.

3) Terzo livello «Derived Work | Derivazione» (in celeste): identifica una produzione teatrale che non deriva direttamente dalla Creazione Artistica originale, bensì da un suo Adattamento, pur mantenendo un legame genetico con l'opera originale – nella Fig. 2 si può notare una linea tratteggiata che collega la Primary work con la Derived Work. La Derivazione presenta ulteriori trasformazioni e rielaborazioni rispetto all'Adattamento da cui trae origine.

Le repliche e le tournée (in arancione) rappresentano le esecuzioni successive alla prima messinscena di una determinata produzione. Le repliche sono le rappresentazioni che si svolgono nella stessa sede, mentre le tournée indicano una serie di rappresentazioni della stessa produzione in diverse località. Sebbene possa sembrare una semplice riproposizione, ogni replica presenta una sua unicità, con variazioni che possono riguardare il cast, gli allestimenti e, soprattutto nel nostro contesto digitale, i materiali ad essa associati (immagini e documenti).

Hyperstage si è confrontato con la necessità di rappresentare digitalmente questa ricchezza di sfumature, portandoci a esplorare due approcci differenti su wikibase.cloud:

1) La creazione di un item specifico per ogni replica: nonostante la chiarezza e la precisione garantite, questo approccio si è dimostrato insostenibile nel lungo periodo, generando una quantità di dati difficilmente gestibile. Unica eccezione è rappresentata dalle produzioni interamente realizzate all'estero, per le quali è stato creato un item specifico per la produzione in tournée.
2) L'inserimento delle proprietà «repliche (P22)» e «tournée (P112)» all'interno dell'entità relativa ad una produzione specifica (QXXX): questo secon-

do approccio si è rivelato più flessibile e sostenibile. Consente di associare le informazioni dettagliate di ogni replica direttamente alla produzione teatrale principale, senza moltiplicare inutilmente gli elementi. Tuttavia, richiede una progettazione accurata delle proprietà e delle relazioni tra i dati, per garantire che tutte le informazioni rilevanti siano catturate e rese accessibili.

Grazie ai qualificatori e ai riferimenti di una specifica dichiarazione[9], è possibile creare una rete di connessioni semantiche che arricchisce significativamente la rappresentazione delle informazioni e facilita la ricerca e l'analisi dei dati.

In sintesi, la gerarchizzazione delle produzioni teatrali, unita a una gestione uniforme di repliche e tournée, si rivela fondamentale per tracciare con precisione l'evoluzione di un'opera. Questo approccio permette di distinguere chiaramente la forma originale, le sue reinterpretazioni, superando le ambiguità nella gestione dei metadati.

3.6 Implementazione del modello gerarchico delle produzioni teatrali in Hyperstage

Per i dettagli della gerarchizzazione precedentemente descritta si rinvia alla pagina di hyperstage.wikibase.cloud. Si rinvia anche alla Fig. 3 dove è rappresentato il grafo che illustra il modello implementato in Hyperstage per strutturare la gerarchia delle produzioni teatrali di Giselle[10].

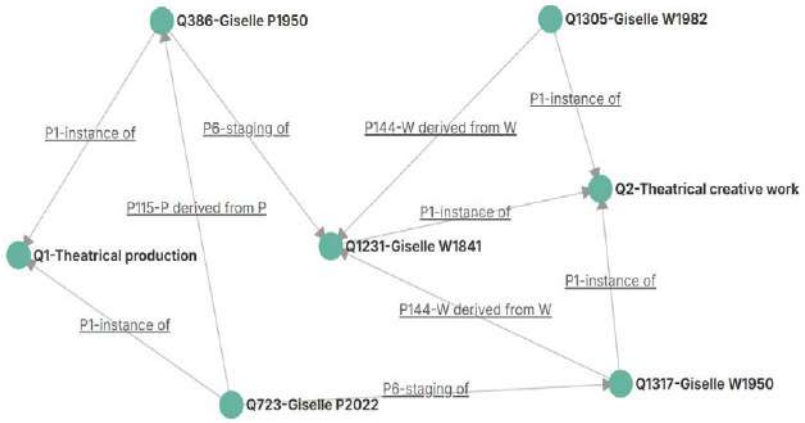


Figura 3 – Esempio di relazioni di derivazione: Q1 da Q1 e Q2 da Q2.

[9] Per il Wikibase data model: <https://www.mediawiki.org/wiki/Wikibase/DataModel>.

[10] <https://tinyurl.com/P113-P115-P116>.

4. Accesso alle risorse digitali: il sistema di classificazione PKB

4.1 Gestione delle risorse digitali in Hyperstage (wikibase.cloud)

Una singola produzione teatrale può generare e aggregare una quantità considerevole di risorse digitali. In wikibase.cloud abbiamo sperimentato diverse metodologie per la gestione delle risorse digitali, optando infine per un approccio basato sulla categorizzazione delle risorse in base alle loro caratteristiche intrinseche. Questo approccio prevede la suddivisione in due grandi categorie:

1) foto di scena, foto delle prove, foto del backstage (con interpreti) e documenti dello spettacolo (locandine, informative, brochure, programma di sala ecc.),
2) bozzetti di costume, bozzetti di scena, foto del costume realizzato, foto della scena realizzata.

Tale distinzione è motivata dalla diversa natura delle risorse documentali. La prima categoria, infatti, raccoglie elementi non corredati da dati tecnici dettagliati, mentre la seconda, comprende risorse con informazioni tecniche associate.

Hyperstage ha l'obiettivo di massimizzare il valore delle risorse digitali, creando connessioni significative tra di esse. Per questo obiettivo è stata sfruttata la capacità di Wikibase di precisare le dichiarazioni con qualificatori e riferimenti, consentendo a Hyperstage di arricchire la comprensione e la reperibilità delle sue risorse digitali. Un esempio chiave di questa capacità si trova nell'utilizzo della proprietà P31 «documentato da (con codominio[11] stringa)». Anche se superficialmente potrebbe sembrare una semplice stringa, Hyperstage sfrutta appieno il suo potenziale attraverso l'impiego di numerosi qualificatori. Questi qualificatori trasformano una semplice indicazione in un ricco insieme di metadati contestuali. Essi permettono di collocare la risorsa digitale in modo preciso nel tempo e nello spazio, fornendo dettagli fondamentali come: informazioni sul fotografo (identificando l'autore dell'immagine); dettagli sul materiale utilizzato (specificando la tecnica o il supporto impiegato es. pellicola, digitale, tipo di carta); contesto temporale (EDTF); localizzazione precisa; persone ritratte (specificando e collegando il nome dell'interprete alla risorsa digitale); informazioni sullo spettacolo (come l'atto, la scena o il momento specifico dello spettacolo documentato) ecc.

Per gestire in modo coerente le informazioni associate a risorse digitali particolarmente complesse, abbiamo creato la proprietà P110 «documentato da (con codominio elemento)». Queste risorse includono: bozzetti di costume, bozzetti di scena, fotografie del costume realizzato e fotografie della scena realizzata. Per semplificare la struttura dei dati, le abbiamo raggruppate nella proprietà «sottoclasse di (P2)»: «risorsa documentale (Q1355)».

Per ognuna di queste risorse è stato creato un elemento specifico. Ad esempio, per il bozzetto di scena intitolato *La clinica*, realizzato per lo spettacolo di

[11] Codominio o rdfs:range è l'insieme di valori che una proprietà può assumere: <https://www.w3.org/TR/rdf-schema/#ch_range>.

Dino Buzzati al Piccolo di Milano (1953), abbiamo creato l'elemento (Q1333)[12]. A questo elemento possiamo associare tutte le dichiarazioni e i relativi qualificatori, senza rischio di ambiguità. Questo sistema ci permette di raccogliere tutte le informazioni necessarie, incluse quelle utili per confrontare e collegare le diverse risorse tra loro, generando così una sorta di scheda catalografica.

Le proprietà utilizzate come qualificatori in «documentato da (P31)» e come dichiarazioni in «documentato da (risorsa) (P110)» sono elencate in questa pagina su hyperstage.wikibase.cloud: https://tinyurl.com/hyperstageprd[13]; in questa altra pagina il dettaglio delle differenze: https://tinyurl.com/DRP31-P110[14].

4.2 Un sistema interconnesso

Le proprietà mostrate nella tabella al link precedente sono fondamentali per trasformare le risorse documentali in un sistema realmente interconnesso. Queste proprietà permettono di creare legami significativi tra i diversi elementi, andando oltre la semplice archiviazione. Ad esempio, è possibile collegare una foto di un bozzetto di costume alla fotografia del costume realizzato e, successivamente, all'immagine dell'attore che indossa quel costume in scena. Questa connessione non solo documenta le singole risorse, ma ricostruisce l'intero processo creativo legato a uno specifico elemento dello spettacolo.

Le potenzialità di questa interconnessione sono molteplici. Un utente può facilmente risalire a tutti gli interpreti che hanno indossato uno specifico costume nel corso delle rappresentazioni, oppure confrontare un modellino di scena con le fotografie della scena effettivamente realizzata.

Il valore aggiunto del nostro sistema risiede proprio in questa capacità di mettere in relazione le risorse digitali. Un ricercatore o un appassionato può esplorare una specifica produzione teatrale e osservare l'evoluzione delle risorse ad essa associate nel tempo, tracciando per esempio le modifiche apportate ai costumi tra le diverse repliche o le variazioni scenografiche.

Le proprietà P31 «documentato da (con codominio stringa)» e P110 «documentato da (con codominio risorsa)» arricchite rispettivamente dai qualificatori e dalle dichiarazioni, sono un potente strumento in questo processo, permettendo di contestualizzare profondamente ogni elemento e di svelare relazioni altrimenti nascoste tra le diverse risorse.

4.3 Ordinare la «scena digitale»: l'efficacia della classificazione PKB

Nell'ambito della documentazione delle produzioni teatrali, un elemento cruciale per garantire un'organizzazione ottimale e una facile reperibilità delle risorse documentali è rappresentato dal sistema di classificazione PKB. La clas-

[12] <https://hyperstage.wikibase.cloud/wiki/Item:Q1333>.

[13] <https://tinyurl.com/hyperstageprd>.

[14] <https://tinyurl.com/DRP31-P110>.

sificazione PKB è stata mappata in un modello ontologico espresso in formato RDF e reso interrogabile tramite query SPARQL. Ciò significa che a ogni tipo specifico di risorsa documentale è attribuito un codice alfanumerico grazie al quale è possibile identificare in maniera univoca la tipologia della risorsa. La sua architettura si fonda su due macro-categorie principali: «pictures» (immagini) e «documents» (documenti). All'interno di queste, come descritto da Gavrilovich (2020), la classificazione PKB struttura le risorse in base a tre fasi fondamentali che ne determinano il processo creativo e i riscontri, permettendo di collegare materiali eterogenei all'interno di un flusso produttivo:

1) concezione,
2) realizzazione,
3) post-produzione.

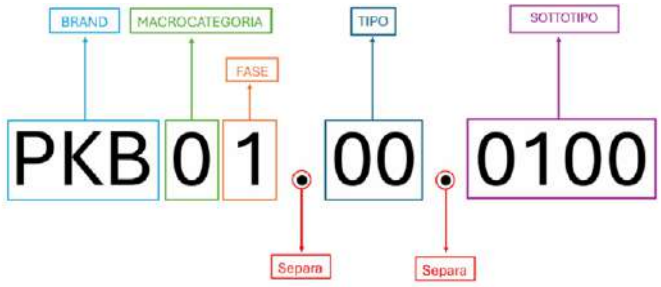


Figura 4 – Esempio di codice PKB: Immagini. Concezione e Progettazione. Scene. Bozzetti di scena.

Questa tripartizione temporale permette di contestualizzare accuratamente ogni risorsa. Un codice PKB (Fig. 4) è costituito da:

- «PKB» è un acronimo che serve a facilitare l'identificazione del brand (in questo caso, il sistema di classificazione stesso) e a ridurre l'ambiguità.
- Le due cifre «01» successive identificano – con la prima cifra, che può assumere valore 0 (immagini) oppure 1 (documenti) le due macro-categorie e – con la seconda cifra, che può assumere i valori 1,2,3 – le rispettive 3 fasi del processo creativo:
- 01 (IMMAGINI, concezione),
- 02 (IMMAGINI, realizzazione),
- 03 (IMMAGINI, post-produzione),
- 11 (DOCUMENTAZIONE, concezione),
- 12 (DOCUMENTAZIONE, realizzazione),
- 13 (DOCUMENTAZIONE, post-produzione).
- Il punto separa la macrocategoria/fase dal tipo.
- Le due cifre a seguire «00», delimitate anch'esse dal punto, identificano il tipo di risorsa ossia: scene, personaggi attrezzi di scena, registrazioni video, foto, immagini, manifesti ecc.

- Il secondo punto separa il tipo dal sottotipo.
- Le quattro cifre successive indicano:
- la prima e la seconda, il sottotipo 1 (le due cifre permettono di ospitare valori che vanno da 00 a 99). Questo primo livello definisce in modo più preciso la risorsa rispetto alla categoria superiore (tipo),
- la terza, il sottotipo 2, aggiunge un ulteriore livello di dettaglio alla specificazione fornita dal sottotipo 2,
- la quarta, il sottotipo 3, definisce in modo ancora più granulare e dettagliato la risorsa.

Queste quattro cifre lavorano in sequenza per restringere progressivamente il campo di definizione della risorsa, passando da una classificazione più ampia a una sempre più dettagliata.

La classificazione completa si trova in questa pagina di hyperstage.wikibase.cloud: https://tinyurl.com/classpkb.

5. Conclusioni

I risultati intermedi del progetto Hyperstage confermano l'adeguatezza della scelta tecnologica di Wikibase.cloud per la gestione della ricostruzione semantica delle produzioni teatrali attraverso la raccolta e l'elaborazione di risorse (metadati e risorse digitali documentali) provenienti dagli archivi dei partner del progetto. In particolare il fatto che Wikibase consenta uno sviluppo bottom-up dell'ontologia ha permesso di sperimentare rapidamente – con dati reali – la sostenibilità di soluzioni che provengono sia da database che documentano spettacoli teatrali (come ad esempio IbsenStage) sia da proposte che provengono anche dal mondo dei dati bibliografici (IFLA LRM).

Il modello si struttura a partire da un nucleo essenziale di entità e dalle loro relazioni. Le entità principali: Q1445 «Agente», Q1 «Spettacolo (Theatrical production)», Q2 «Opera creativa teatrale (Theatrical creative work)», Q1355 «Risorsa documentale (Documentary resource)», Q12 «Classificazione PKB». Oltre alla relazione P6 «messinscena di (staging of)» tra Q1 «Spettacolo» e Q2 «Opera creativa teatrale» che consente il riferimento ad una identità comune al di là delle specifiche istanze di Spettacolo, sono state prese in considerazione le relazioni di tipo gerarchico: P115 «spettacolo derivato da spettacolo», una relazione tra entità Q1 «Spettacolo»; P113 «opera teatrale derivata da opera teatrale», una relazione tra entità Q2 «Opera creativa teatrale».

Le funzionalità offerte dai qualificatori (qualifiers) e dai riferimenti (references) previsti nel Wikibase data model sono state sfruttate da Hyperstage in molti contesti, tra questi: la gestione delle repliche e la relazione tra Spettacolo e Risorsa documentale. Senza i qualificatori anche esprimere correttamente con la sintassi RDF[15] che

[15] <https://www.w3.org/RDF/>.

ad esempio «Carla Fracci ha interpretato Giselle nello Spettacolo Giselle al Teatro alla Scala nel 1971» risulterebbe estremamente complesso[16].

Infine la classificazione PKB che organizza la documentazione in formato digitale dello Spettacolo ha trovato il suo posto all'interno dell'ontologia: ogni simbolo di classificazione è definito come istanza della classe «classificazione PKB».

Le prospettive di lavoro future comprendono l'introduzione di meccanismi di controllo dell'inserimento dei dati basati su ShEx[17] e il potenziamento del mapping sia a livello di ABox che di TBox. Un obiettivo critico è definire un framework per la manutenzione collaborativa dell'ontologia anche dopo il termine del progetto.

Riferimenti bibliografici

Beck, Julia, e Franziska Voß. 2020. "Aggregating Performing Arts: Representing different Collections in one Database." *Arti dello Spettacolo/Performing Arts*, Special Issue, Giugno: 23-30.

De Marinis, Marco. 2008. *Capire il teatro. Lineamenti di una nuova teatrologia.* Roma, Bulzoni.

Diwisch, Kerstin, e Bernhard Thull. 2014. "Modeling and Managing the Digital Archive of the Pina Bausch Foundation." In *Metadata and Semantics Research, 8th Research Conference, MTSR 2014*, Karlsruhe, Germany, November 27-29, Proceedings. Cham, Springer.

Estermann, Beat, e Christian Schneeberger. 2017. "Data Model for the Swiss Performing Arts Platform, Version 0.51." Bern University of Applied Sciences, E-Government Institute.

Estermann, Beat, e Frédéric Hulien. 2019. *Linked digital future for the performing arts*, CAPACOA, <https://capacoa.ca/documents/research/LDF_Report_2019.pdf>

Estermann, Beat. 2020. "Creating a Linked Open Data Ecosystem for the Performing Arts (LODEPA)." *Arti dello Spettacolo/Performing Arts*, Special Issue, giugno: 31- 49.

Gavrilovich, Donatella. 2013. "How to Catalogue the Cultural Heritage 'Spectacle'." In *Information Technologies for Performing Arts, Media Access, and Entertainment*, Eds. Paolo Nesi e Raffaella Santucci, vol. 7990. Berlin, Springer, 39-49.

Gavrilovich, Donatella. 2017. "Performing Arts Archives. Dal Database al Knowledge base: stato dell'arte e nuove frontiere di ricerca." *Arti dello Spettacolo/Performing Arts* III, settembre: 28-39.

Gavrilovich, Donatella. 2020. "Open Data for an International Performance Knowledge base (PKb) and a persistent identifier (ASPA Code)." *Arti dello Spettacolo/Performing Arts*, Special Issue, giugno: 57-70.

Guarino, Nicola, e Cristopher A. Welty. 2009. "An Overview of Ontoclean." In *Handbook on Ontologies*, a cura di Steffen Staab, e Rudi Studer. Dordrecht, Springer, 201-220. <https://doi.org/10.1007/978-3-540-92673-3>

Guarino, Nicola, Oberle, Daniel, e Steffen Staab. 2009. "What Is an Ontology?" In *Handbook on Ontologies*, a cura di Steffen Staab, e Rudi Studer. Dordrecht, Springer, 1-17. <https://doi.org/10.1007/978-3-540-92673-3>

[16] <https://hyperstage.wikibase.cloud/wiki/Item:Q777>.

[17] <https://shex.io/>.

Hanssen, Jens-Morten. 2018. *Ibsen on the German Stage 1876–1918: A Quantitative Study*, Tübingen: Narr Francke Attempto Verlag.

Holledge, Julie, Bollen, Jonathan, Helland, Frode, e Joanne Tompkins. 2016. *A Global Doll's House. Ibsen and Distant Visions.* London: Palgrave Macmillan.

Mitsopoulou, Evie, Kyprianos, Konstantinos, e Pantelis Brattis. 2024. "Documenting the ephemeral: an ontology for the performing arts." *Journal of Information Science*, <https://doi.org/10.1177/01655515241271052>

Nardi, Daniele, e Ronald J. Brachman. 2010. "An Introduction to Description Logics." In *The Description Logic Handbook: Theory, Implementation and Applications*, a cura di Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, Peter F. Patel-Schneider. Cambridge: Cambridge University Press, 1-44.

Samuel, John. 2017. "Collaborative approach to developing a multilingual ontology: a case study of Wikidata." In *Metadata and Semantic Research. MTSR 2017* a cura di Emmanouel Garoufallou, Sirje Virkus, Rania Siatri, Damiana Koutsomiha. *Communications in Computer and Information Science*, vol 755. Cham, Springer, 167-72. <https://doi.org/10.1007/978-3-319-70863-8_16>

Skaug, Johanna, e Trond Aalberg. 2024. "The Concept of Work in Theater Performances." *Cataloging & Classification Quarterly*, 62 (3–4), 280–301. <https://doi.org/10.1080/01639374.2024.2357208>

Weiberg, Birk. 2020. "Modeling Performing Arts: On the Representations of Agency." *Arti dello Spettacolo/Performing Arts*, Special Issue, Giugno: 50-56.

Leveraging Wikidata for Amateur Theatre Research: The WikiFAIR Approach of the P-CITIZENS Amateur Theatre Wiki

Ioanna Papazoglou, Meike Wagner

Abstract: The ERC-funded project P-CITIZENS (Performing Citizenship) studies social and political roles of amateur theatre in Europe (1780–1850). In this effort we built the Amateur Theatre Wiki: a low-overhead platform whose structured data lives in Wikidata while narrative and media content remain local. Using a WikiFAIR approach we harvest troupe data (using OpenRefine and BeautifulSoup), reconcile entities, and render remote statements in infoboxes via UnlinkedWikibase. This separation improves reusability, interoperability, copyright clarity, and accessibility, and demonstrates a sustainable pattern for small to mid-sized humanities projects.

Keywords: Wikidata, Amateur Theatre, WikiFAIR, Digital Humanities, Structured Data.

1. Introduction

As part of the ERC-funded P-CITIZENS (Performing Citizenship) project, hosted at Ludwig Maximilian University Munich, we have developed a collaborative platform to document both historical and contemporary amateur theatre. The project investigates the role of non-professional theatre in shaping concepts of citizenship across Europe between 1780 and 1850. The Amateur Theatre Wiki[1] launched in November 2022, hosts detailed entries on the individuals, organizations, and venues responsible for amateur productions and it welcomes contributions from volunteers who enrich the site with information about active troupes in their regions.

To jump-start community participation, we employ "red links" a well-established engagement strategy that invites users to create missing articles directly within the Wiki. Implementing this feature required assembling an initial roster of amateur-theatre troupes—yet existing public datasets proved sparse, and no specialized registry was available.

[1] <https://www.amateur-theatre-wiki.gwi.uni-muenchen.de/index.php/>.

Ioanna Papazoglou, University of Munich Ludwig Maximilian, Germany, Ioanna.Papazoglou@itg.uni-muenchen.de, 0009-0000-6926-0685
Meike Wagner, University of Munich Ludwig Maximilian, Germany, meike.wagner@lmu.de, 0000-0002-1003-3979

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Ioanna Papazoglou, Meike Wagner, *Leveraging Wikidata for Amateur Theatre Research: The WikiFAIR Approach of the P-CITIZENS Amateur Theatre Wiki,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.19, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 147-153, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

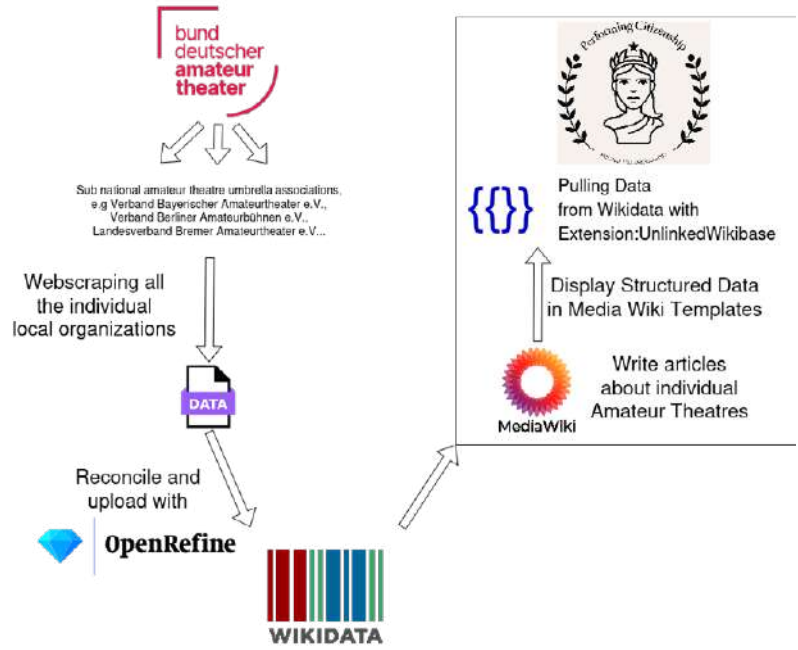


Figure 1 – Data extraction and ingestion process from regional federation catalogues to Wikidata and MediaWiki.

In this paper, we first describe our workflow for locating, extracting, and structuring data from diverse web sources, reconciling records against Wikidata and creating new items for previously undocumented troupes. We then report on our application of the WikiFAIR guidelines to simplify our infrastructure: rather than self-hosting a separate structured database, we integrate Wikidata into our MediaWiki instance, rendering remote statements—such as addresses, founding dates and images—directly within local infoboxes alongside the full-text articles.

2. Extraction of amateur theatre data from regional catalogues

Regional theatre federations (or Dachverbände) often maintain online catalogues of their member associations (see Figure 1). In Germany, most states provide publicly accessible member lists; however, exceptions such as Baden-Württemberg and Hesse lack suitable formats for automated extraction. For example, Hesse's directory appears only as a dropdown menu on a questionnaire subpage, listing organization names without accompanying metadata. Effective scraping therefore requires at minimum a linked web presence or an associated city name.

To harvest these records, we use Python's BeautifulSoup and the "Instant Data Scraper" Chrome extension. This pipeline locates and parses HTML tables, lists and form elements, extracting troupe names, locations and contact details into a structured CSV output suitable for subsequent reconciliation with Wikidata.

3. Motivation for implementing WikiFAIR recommendations

Having assembled a heterogeneous collection of structured data on amateur theatre associations in Germany, the project faced the challenge of deciding how best to allocate limited resources to present this data within MediaWiki beyond the use of red links. WikiFAIR[2] offers guidelines, developed in collaboration with the WikiKULT Community of Practice[3], for implementing FAIR principles in research projects with constrained budgets and manpower.

As a catalyst for projects with FAIR[4] targets of Findability, Accessibility, Interoperability and Reusability, WikiFAIR focuses on integration of research data with different free knowledge platforms, so that their technical infrastructure and services fulfill most of the FAIR requirements automatically. Following these guidelines, we focused on implementing a bridge between Wikimedia projects (mainly Wikidata, Wikipedia and Wikimedia Commons) and our MediaWiki instance, in order to reduce the complexity of the project's technical stack. Although we deployed a simple self-hosted MediaWiki instance via Docker (a widespread containerization solution for software) and managed with Ansible (a configuration management system)[5], we realised early on that running a separate Wikibase deployment to manage the newly acquired structured data would require substantial additional effort.

Consequently, we followed the WikiFAIR recommendation to integrate our dataset into Wikidata, not into an self-hosted alternative. This approach conserved considerable project resources that would otherwise have been spent developing a custom ontology and maintaining multiple components of the Wikibase ecosystem. Moreover, by leveraging Wikidata's various API interfaces, long-term hosting and backup infrastructure, our data automatically inherits FAIR compliance through an established, sustainable platform.

4. Upload process

To upload data to Wikidata, we first check for existing amateur theatre troupes by running a reconciliation request through OpenRefine an open-source tool for cleaning, exploring, and reconciling structured data against external data-

[2] WikiFAIR Project Page <https://meta.wikimedia.org/wiki/WikiFAIR>.

[3] <https://meta.wikimedia.org/wiki/WikiKult–OffeneKulturdaten>.

[4] <https://www.go-fair.org/fair-principles/>.

[5] <https://ansible.com>.

bases such as Wikidata[6]. This step prevents the creation of duplicate items for well-known associations that already appear in Wikidata due to existing Wikipedia articles. During reconciliation, we apply an import schema based on a predefined set of properties , listed in Table 1.

For long-term citability, we link to Archive.org snapshots of the original member lists whenever possible as the Internet Archive ensures persistent and open access in line with FAIR principles (see Figure 3). Finally, we visualize the newly uploaded data on a map, displaying the coordinates of each troupe's nearest city (Figure 2).

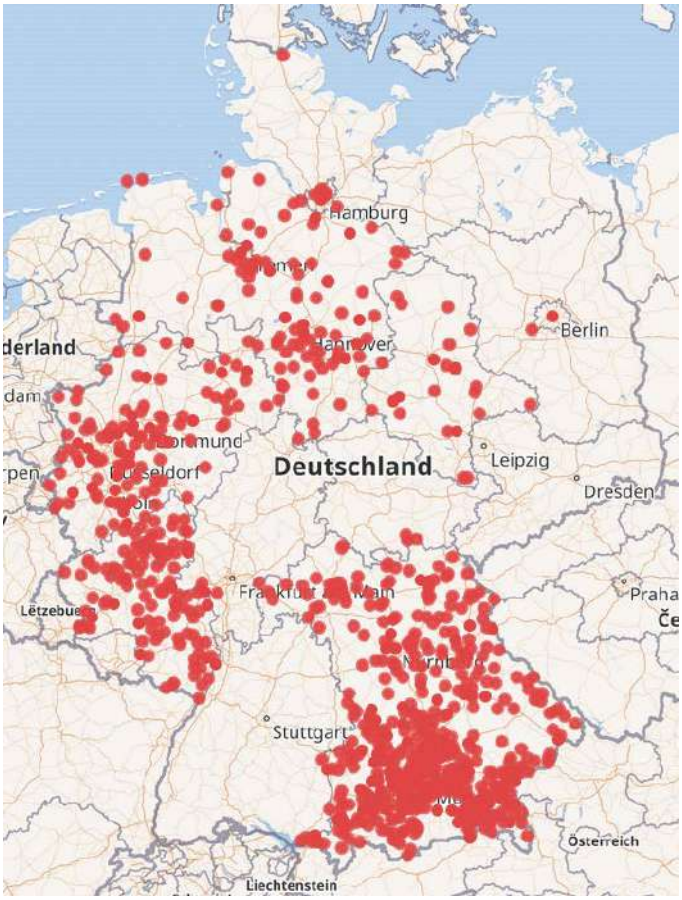


Figure 2 – Result of <https://w.wiki/86S$> SPARQL query visualizing sample amateur theatre associations in Germany.

[6] <https://openrefine.org/>.

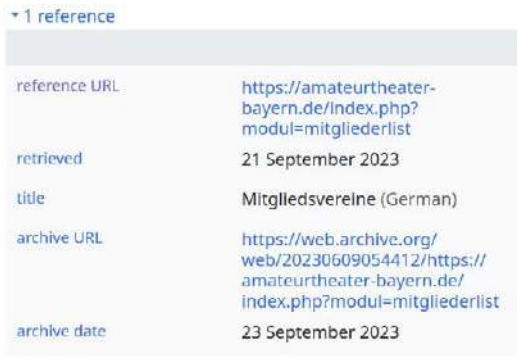


Figure 3 – Archival references attached to uploaded Wikidata statements (Archive.org snapshots) related to Wikidata item Q2513550.

Property	Description	Example Value
P31	instance of	Q2416217 (theatre troupe)
P571	inception date	date of founding
P1448	official name	full legal name of the association
P101	field of work	Q455695 (amateur theatre)
P17	country	Q183 (Germany)
P463	member of	e.g. Q2513550 (Verband Bayerischer Amateurtheater)
P159	headquarters location	nearest city (from contact address)
P1454	legal form	Q9299236 (registered association)
P856	official website	URL of the troupe or association

Table 1: Added properties to Wikidata

5. Automated infobox rendering via UnlinkedWikibase

Because we elected not to deploy our own Wikibase instance, the standard Wikibase Client and Repository extensions—which require direct database access—cannot render structured data in MediaWiki pages. Consequently, traditional Wikidata infobox templates, which depend on parser functions tied to a locally linked Wikibase, are inoperable on our platform.

Several alternatives exist, including embedding live SPARQL query results from the Wikidata Query Service via an <iframe>, which can display maps or tables but cannot integrate with template logic; adopting a federation extension such as LinkedWiki, which introduces its own parser functions and configuration overhead; or using a Lua-based solution that fetches and caches remote entities , provided by the UnlinkedWikibase extension.

We selected the UnlinkedWikibase extension for its minimal footprint and compatibility with existing infobox templates.[7]

7 <https://www.mediawiki.org/wiki/Extension:UnlinkedWikibase>.

Once installed, the extension adds parser functions—such as #entity, #props, and #statements—that retrieve labels, claims and sitelinks from Wikidata over HTTP and cache them locally for a configurable interval. No database schema changes are required, and Lua modules[8] need only specify a Wikidata Q-ID to populate an infobox dynamically. For an example about an amateur theatre actor see Figure 4. Since the templating system in MediaWiki is complex and a hard development target, we swichted the Infobox rendering to the Capiunto extension, that allows users to create fully functional boxes directly in Lua code.

This approach preserves the familiar MediaWiki templating workflow, reduces maintenance overhead, and ensures that all structured data remain up-to-date, fully FAIR-compliant and seamlessly integrated alongside our full-text articles.

Figure 4 – Example of an infobox for a person.

Credits

We thank regional amateur theatre federations for publishing member lists that enabled structured data creation, and the WikiKULT Community of Practice for articulating WikiFAIR guidelines.

[8] <https://www.amateur-theatre-wiki.gwi.uni-muenchen.de/index.php/Special:AllPages?from=&to=&namespace=828>

References

Meta-Wiki contributors. 2026. "WikiFAIR." *Meta-Wiki* <https://meta.wikimedia.org/w/index.php?title=WikiFAIR&oldid=28987358>.

Reagle, Joseph M. 2010. *Good Faith Collaboration: The Culture of Wikipedia*. Cambridge, MA: MIT Press.

Reconciliation Community Group. 2023. *Reconciliation Community Group Final Specifications, Version 0.2*. W3C Community Group Report (CG-FINAL-specs-0.2). World Wide Web Consortium (W3C). <https://www.w3.org/community/reports/reconciliation/CG-FINAL-specs-0.2-20230410/>.

Spinellis, Diomidis, and Panagiotis Louridas. 2008. "The Collaborative Organization of Knowledge." *Communications of the ACM* 51, no. 8 (August): 68–73. <https://doi.org/10.1145/1378704.1378720>.

Committenze semantiche: Wikidata per la ricerca storico-artistica[1]

Alessio Ionna

Abstract: The research presents the application of Wikidata for the semantic modelling of the patronage of the Buonaccorsi family of Macerata, a prominent dynasty of the Papal State during the 17th and 18th centuries. Through a systematic analysis of the sources, a dataset of over 400 family-related entities was structured and annotated with ontologically consistent properties. The interconnected data allow for a diachronic and relational reading of art-historical dynamics. Precise referencing of statements increases the scientific reliability of the model, facilitating data reuse. The study shows how Wikidata and Linked Open Data provide an effective infrastructure for the integration, dissemination and analysis of complex knowledge in the humanities.

Keywords: Wikidata; Art-historical research; Artistic patronage.

1. La ricerca storico-artistica nel contesto del Web 3.0

Diversi ambiti di ricerca, compreso quello umanistico, hanno visto negli ultimi anni una sempre maggiore attenzione alle opportunità aperte dall'impiego delle tecnologie informatiche e digitali come nuovi approcci verso la ricerca. La concettualizzazione di una figura che unisca competenze umanistiche e informatico-digitali come quella dell'umanista digitale diventa pertanto centrale in un contesto storico-sociale fortemente improntato al digitale, in quanto spesso la gestione delle infrastrutture della circolazione della conoscenza è in mano a figure tecniche che non condividono gli stessi valori culturali (Numerico et al. 2010, 7).

Tale approccio è fondamentale soprattutto per il settore del patrimonio culturale. L'ambito dei beni culturali è infatti ancora legato ad un concetto neoidealista di cultura (Montella 2016a, 18), per cui la qualità culturale è spesso legata alla bellezza e a tutto ciò che è artistico, e che qualsiasi tentativo di messa in pratica tecnologica comporterebbe un tradimento di questi valori (D'Angelo 2016). Pertanto, l'applicazione delle tecnologie digitali al settore culturale è stata per lungo tempo osteggiata e ridotta quasi esclusivamente alla mera applicazione strumentale, non riconoscendone alcuna funzione metodologica (Roncaglia 2002).

[1] Ultima consultazione delle risorse online in data 13 giugno 2026.

Alessio Ionna, University of Macerata, Italy, a.ionna1@unimc.it, 0000-0002-1880-1581

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

 DOI 10.36253/979-12-215-1010-2.20, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 153-164, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

Tale refrattarietà ad ogni possibile contaminazione tecnologica in ambito storiografico permane dominante fino agli anni '90, quando lo sviluppo del World Wide Web di Tim Berners-Lee determinò un cambio di approccio alla ricerca attraverso un'informatizzazione di base delle competenze umanistiche (Noiret 2008, p. 197; Paci 2021, 24-25). Il cambio di mentalità successivo agli ultimi anni del XX secolo ha riguardato anche l'ambito dei beni culturali, stimolando lo sviluppo di un approccio più "antropologico" e svuotato da ogni retorica crociana (Montella 2016b, 19-20).

La continua evoluzione che il web sta vivendo nell'ultimo decennio con il passaggio dal web 2.0 a quello 3.0, definito anche web semantico, ha posto nuove problematiche agli umanisti digitali, in particolare per quanto riguarda i processi di creazione e gestione di contenuti a carattere culturale. Il web 3.0 si basa sulla tecnologia dei Linked Data, strumenti lungamente impiegati dalle scienze dure che negli ultimi anni stanno prendendo piede anche nelle scienze umane. Le capacità relazionali dei Linked Data possono essere molto utili per favorire la circolazione della conoscenza culturale, permettendo di poter accedere a informazioni strutturate in un formato comprensibile anche ad utenti macchina. Inoltre, la possibilità di creare relazioni può essere impiegata per sviluppare un approccio innovativo allo studio di fenomeni culturali complessi come quello delle committenze artistiche.

Qualora uno storico dell'arte intendesse approcciarsi allo studio del mecenatismo di una famiglia aristocratica si troverebbe a dover districare una fitta ragnatela di legami che non coinvolgono solo l'aspetto artistico, ma anche quello socio-culturale in cui la casata si muoveva. Per fare ciò, gli storici sono costretti ad una lunga attività di ricerca su più fronti, consultando un vasto un numero di fonti disponibili in diversi formati e spesso poco o per nulla accessibili. Tale dispersione informativa rende il lavoro dello storico complicato e non sempre affidabile, allungando notevolmente i tempi di ricerca e rendendo i risultati difficili da consultare. Un modo per agevolare questo tipo di ricerche è far sì che tutte le informazioni riguardanti committente, artista, opera e contesto siano raccolte in un unico spazio e rese facilmente accessibili con la possibilità di mettere in relazione tra loro i dati e le informazioni, restituendo così un quadro il più ampio possibile del contesto della committenza. Le tecnologie digitali, e in particolar modo quelle del web semantico possono fornire una valida risposta a questa esigenza.

2. Il caso studio: la famiglia Buonaccorsi di Macerata e la piattaforma Wikidata

Come caso studio su cui testare questo innovativo approccio alla ricerca storico-artistica è stata individuata la committenza artistica della famiglia Buonaccorsi di Macerata. I Buonaccorsi sono stati una delle dinastie più rilevanti dello Stato Pontificio a cavallo tra XVIII e XIX secolo e nel corso della loro storia hanno ricoperto svariati incarichi politici e religiosi sia nella Marca di Ancona che nel resto dello stato (Melatini 1998, 13-14). A partire dal XVI secolo la dinastia maceratese iniziò ad incrementare il proprio prestigio, arrivando a vantare tra i suoi membri ben due cardinali: Bonaccorso Buonaccorsi, ordinato cardinale nel 1669 (Cardella 1793, 202) e Simone Buonaccorsi nel 1763 (Pignatelli 1969). Quest'ultimo è una figu-

ra centrale per la storia della famiglia, in quanto tende ad essere considerato dalla storiografia un "papa mancato": egli infatti partecipò a ben due conclavi, quello del 1769 e quello del biennio 1774-75, ed in entrambi i casi la sua elezione venne negata dal veto postogli dal re di Francia (Pignatelli 1969; Melatini 1998, 25).

L'importanza dei Buonaccorsi è da ricercarsi anche nel fatto che furono un'importante famiglia di committenti artistici. Il loro gusto collezionistico li portò a stringere legami con le più importanti scuole pittoriche dell'epoca. Tali legami erano esplicitati dalle opere d'arte che adornavano le pareti dei loro palazzi di Macerata, Roma e Potenza Picena.

Apice della loro committenza fu la realizzazione della maestosa Galleria dell'Eneide di Palazzo Buonaccorsi a Macerata, commissionata dal conte Raimondo Buonaccorsi, uno dei più importanti mecenati artistici della sua epoca (Capriotti 2018; Coltrinari 2018; Haskell 2020, 316).

Il patrimonio artistico della famiglia Buonaccorsi ha avuto la fortuna di giungere intatto fino al secondo dopoguerra. Questa continuità era stata garantita attraverso un fidecommesso redatto nella seconda metà del Cinquecento in cui era indicato, tra le diverse regole e obblighi, che l'intero patrimonio poteva essere ereditato esclusivamente dal figlio primogenito (Melatini 1998, 17; Coltrinari 2018, 20). Con l'Unità d'Italia e l'emanazione del Regio Decreto n. 71 questo documento cessa di essere vincolante e i beni iniziano ad essere spartiti tra i diversi membri della casata, per poi essere definitivamente alienato negli anni '60.

Ad oggi, gran parte del patrimonio della famiglia Buonaccorsi è di proprietà pubblica. Il palazzo cittadino venne acquistato dal comune di Macerata nel 1967 ed oggi è la sede dei musei civici (Melatini 1998, 35). La biblioteca privata venne fatta oggetto di prelazione negli anni '70 ed è oggi il nucleo principale della Biblioteca Statale di Macerata (Monti 2013, 187), mentre l'archivio di famiglia venne venduto dagli eredi Buonaccorsi direttamente allo Stato italiano ed oggi è conservato presso l'Archivio di Stato di Macerata (Coltrinari 2018, 16; Domenichini 2019, 137). La villa di Potenza Picena è stata proprietà di diversi soggetti privati, finché nel 2022 è entrata a far parte del patrimonio culturale statale a seguito dell'attività di prelazione da parte del Ministero della Cultura.

Per quanto riguarda invece i beni storico-artistici della famiglia Buonaccorsi, solo alcuni di questi sono stati recuperati dal mercato antiquario (Barucca, Sfrappini 2001), mentre il resto della quadreria privata risulta ancora disperso se non per poche opere conservate presso collezioni private (Costanzi 2005, 272).

Data la grande importanza che ha ricoperto a livello locale e nazionale, la famiglia Buonaccorsi è stata fin da subito oggetto di ricerche e studi di livello internazionale, che hanno permesso di inserire il fenomeno mecenatesco maceratese nel più ampio quadro degli studi sulla cultura artistica europea del XVII secolo. Da menzionare un importante convegno tenutosi nel 2017 presso il dipartimento di Scienze della formazione, beni culturali e turismo dell'Università degli studi di Macerata proprio sul tema della committenza Buonaccorsi nel contesto culturale dell'Europa settecentesca.

Poter rappresentare efficacemente un complesso fenomeno come la committenza di una famiglia nobile dipende dalla disponibilità di un dataset *open*

access che permetta di descrivere tutte le informazioni utili alla restituzione del contesto mecenatesco e l'ambiente online che risponde in maniera idonea a tale esigenza è l'ecosistema della Wikimedia Foundation. La Wikimedia Foundation è uno dei player più importanti nell'attuale panorama web e il fatto di portare avanti una serie di progetti, popolarmente definiti Wiki, incentrati sulla libera circolazione della conoscenza la ha resa particolarmente apprezzata in ambito umanistico per quanto riguarda la circolazione e la diffusione della cultura in ambiente digitale (Bertacchini et al. 2022).

Il progetto della fondazione che si concentra maggiormente su questo aspetto della libera circolazione della conoscenza online è Wikipedia: enciclopedia online concepita con l'obiettivo di diventare universale e capace di raccogliere l'intero sapere umano. È stata ideata con approccio *bottom-up*, in contrapposizione al rigido modello enciclopedico dominante fino alla fine del XX secolo, secondo cui solo gli esperti hanno le competenze per poter realizzare contenuti enciclopedici affidabili (Codogno 2009). Nell'attuale ottica web, però, tale impostazione infrastrutturale si dimostra non più efficace, in quanto pur essendo garanzia di affidabilità, non permette di mantenere aggiornate le informazioni contenute. Il modello Wikipedia di contro permette una continua revisione e aggiornamento dei contenuti enciclopedici, resi affidabili dal continuo controllo effettuato dalla comunità di utenti (Roncaglia 2021, 78).

Gli ambienti Wiki impiegati per la ricerca che si presenta in questa sede sono stati 3: la già citata enciclopedia online Wikipedia, il *repository* di contenuti multimediali Wikimedia Commons e la base di conoscenza Wikidata.

Wikidata è uno dei progetti più recenti della Wikimedia Foundation ed è stata sviluppata con l'intento di proporsi come "base di conoscenza universale" libera e collaborativa. Inizialmente ideata esclusivamente come archivio di dati centralizzato per tutti i progetti Wikimedia, con il tempo ha acquisito una propria autonomia (Vrandečić 2013, 92). Al suo interno è possibile rappresentare ogni tipo di entità – sia essa una persona fisica, evento storico, oggetto reale, concetto astratto ecc. – attraverso Linked Open Data (LOD). Le entità sono descritte all'interno del database attraverso strutture semantiche definite "elementi", a loro volta descritti tramite dichiarazioni (*statements*) che qualificano l'entità attraverso l'attribuzione di proprietà, rappresentandola e con la possibilità di collegarle l'una con l'altra. Esattamente come gli altri progetti Wiki, anche Wikidata è stata sviluppata con un approccio *bottom up*, in cui sono gli utenti che creano i dati all'interno della piattaforma. I dati caricati sono utilizzabili attraverso la licenza di libero dominio (CC0), rendendoli così liberamente riusabili da chiunque per qualsiasi scopo anche al di fuori dell'ambiente Wikidata, permettendo così l'interoperabilità tra la piattaforma e il resto del cloud dei LOD.

La peculiarità che ha determinato la scelta della base di conoscenza Wikidata come piattaforma dove poter caricare i dati necessari alla rappresentazione semantica del fenomeno della committenza artistica è la possibilità di citare le fonti impiegate per il recupero delle informazioni, esattamente come accade per le pubblicazioni scientifiche. Questa possibilità di citare le fonti permette di valutare l'affidabilità dei dati caricati, rendendo più sicuro il loro riutilizzo per scopi di ricerca.

Questa capacità descrittiva ha determinato il vasto impiego della piattaforma soprattutto in ambito culturale, dimostrandosi un valido strumento per la conservazione e l'arricchimento dei dati. Il settore bibliotecario è fra i primi ambiti ad utilizzare Wikidata con lo scopo di creare un linguaggio il più standardizzato possibile per agevolare il dialogo tra i vari soggetti che operano nel contesto biblioteconomico (Van Veen 2019; Bianchini, Sardo 2022, 292).

Mentre l'impiego di Wikidata in ambito biblioteconomico è ampiamente diffuso e radicato, solo da pochi anni la piattaforma ha visto la sua applicazione anche negli altri settori MAB, quali archivi e musei. Eccezion fatta per il caso degli Archivi Nazionali di Francia (Clavaud 2019), in ambito archivistico non vi è stato ancora un fattivo impiego della piattaforma, nonostante il vivace dibattito sviluppatosi negli ultimi anni (Kapsalis 2019; Colla et al. 2021, Feliciati 2022a; 2022b)[2]. In ambito museale e storico artistico Wikidata viene impiegata come strumento per descrivere semanticamente i beni conservati presso un'istituzione culturale e rappresentarne in maniera diacronica l'intero contesto di riferimento (Alexiev et al. 2020; Bianchini, Spinelli 2020; Ionna 2023). Il più delle volte, i progetti culturali basati sulla piattaforma Wiki si limitano ad impiegarla come *repository* di dati e come strumento di riconciliazione tra le entità conservate su diversi database (Evenstein Sigalov, Nachmias 2023; Fagerving 2023).

3. Il Wikiproject Buonaccorsi

Operativamente, le azioni volte alla restituzione in ambiente semantico del fenomeno oggetto della ricerca sono state suddivise in tre fasi: spoglio, ricognizione e caricamento. La fase di spoglio ha riguardato la raccolta di tutte le informazioni più rilevanti recuperate da fonti bibliografiche e da nuove ricerche d'archivio, necessarie ad una corretta restituzione del fenomeno mecenatesco.

La fase di ricognizione è stata svolta all'interno di Wikipedia, Wikimedia Commons e Wikidata, i tre progetti Wikimedia scelti come ambienti di ricerca, con lo scopo di individuare la presenza di possibili risorse riconducibili alla committenza Buonaccorsi.

La terza fase, quella di caricamento ha riguardato invece la creazione di nuovi elementi Wikidata partendo dalle informazioni emerse durante la fase di spoglio.

La fase di spoglio ha permesso di vagliare criticamente la bibliografia esistente dedicata alla famiglia Buonaccorsi. A questa analisi bibliografica hanno poi fatto seguito ulteriori ricerche archivistiche al fine di recuperare le informazioni necessarie a coprire alcune lacune informative e approfondire il fenomeno mecenatesco.

A seguito della fase di ricognizione, all'interno di Wikipedia è emersa la presenza di 8 voci enciclopediche intitolate a "Buonaccorsi", di cui però solo 4 erano riconducibili al contesto maceratese. All'interno di Wikimedia Commons

[2] Si segnala in merito il gruppo di lavoro *WikiProject Archival Description* sulla rappresentazione delle strutture archivistiche all'interno di Wikidata. <https://www.wikidata.org/wiki/Wikidata:WikiProject_Archival_Description>.

erano presenti poco meno di 460 contenuti multimediali, la maggior parte dei quali non riconducibili ai Buonaccorsi di Macerata. All'interno di Wikidata erano invece presenti solo gli elementi delle pagine presenti nelle diverse versioni linguistiche dell'enciclopedia Wikipedia.

La fase di caricamento ha riguardato la creazione delle risorse necessarie alla rappresentazione del contesto mecenatesco nell'ecosistema Wikimedia. All'interno di Wikipedia sono state migliorate le voci preesistenti e creata la pagina dedicata alla famiglia Buonaccorsi. All'interno di Wikimedia Commons si è provveduto all'organizzazione dei contenuti multimediali attraverso la creazione di una categoria Commons relativa alla famiglia Buonaccorsi di Macerata. All'interno di Wikidata le attività hanno riguardato il miglioramento degli elementi già presenti per il raggiungimento di buoni livelli descrittivi e la creazione *ex novo* degli elementi relativi ai membri della famiglia, le opere commissionate, gli artisti chiamati, il contesto storico di riferimento e le fonti impiegate ecc.

Nella fase di caricamento si è cercato di utilizzare le proprietà più idonee alla descrizione di ogni entità che contribuisce alla complessità del fenomeno mecenatesco. Le principali proprietà usate per la descrizione della committenza artistica della famiglia Buonaccorsi in ambiente Wikidata sono state: "Commissionato da" (P88), "Proprietario di" (P127) e "Luogo" (P276).

La prima, "Commissionato da" (P88), è stata impiegata per riconciliare gli elementi delle istanze culturali commissionate da specifici membri della famiglia Buonaccorsi: in particolare tale proprietà, con valore "Raimondo Buonaccorsi" (Q109805360) è stata largamente impiegata per la descrizione delle opere presenti nella Galleria dell'Eneide. La proprietà "Proprietario di" (P127) è stata impiegata per relazionare al grafo semantico della famiglia Buonaccorsi tutti gli elementi riguardanti quelle entità culturali e non per cui non è possibile ricondurre ad uno specifico committente, ma che risulta dalle fonti consultate essere appartenuto alla casata maceratese nel corso della sua storia. La proprietà "Luogo" (P276) è stata invece usata per geolocalizzare gli elementi, permettendo in questo modo di determinare spazialmente l'estensione del fenomeno artistico.

Al termine del caricamento è stato possibile realizzare un dataset di oltre 400 elementi Wikidata riconducibili al fenomeno mecenatesco della famiglia Buonaccorsi. Questi elementi sono stati tutti adeguatamente referenziati e correttamente relazionati tra loro, al fine di restituire efficacemente l'intero contesto di riferimento.

Dal punto di vista dell'impatto dei dati al di fuori di Wikidata, si è tentato di valutare il numero di visualizzazioni degli elementi del dataset in esame. Tramite il *tool PageViewAnalysis* è stato possibile effettuare un'analisi all'interno dell'intero dataset. I risultati hanno permesso di individuare quali fossero gli elementi più visitati dagli utenti Wikidata: "Famiglia Buonaccorsi" (Q109805886) con 320 visualizzazioni, "Raimondo Buonaccorsi" (Q109805360) con 264 e "Palazzo Buonaccorsi" (Q262019) con 208. L'elemento con il minor numero di visualizzazioni è risultato essere invece "Mazzagalli" (Q116234274) con solo una visualizzazione.

I dati emersi dalla analisi attraverso PageView hanno permesso quindi di determinare l'esistenza di "nodi semantici" all'interno del grafo della famiglia Buonaccorsi: difatti, i tre elementi maggiormente visitati sono stati inseriti co-

me valore per gli *statements* di diversi elementi e risultano pertanto centrali. Di contro, "Mazzagalli" è relativo solo al cognome della coniuge di uno dei membri della casata ed è considerabile un nodo "terminale", relegato al margine del grafo proprio per la sua debole capacità relazionale.

L'intero dataset della famiglia Buonaccorsi è analizzato attraverso delle *query* realizzate attraverso il Wikidata Query Service, per poi essere restituito graficamente attraverso dei *tools*, definiti visualizzatori, messi a disposizione dalla piattaforma e dalla comunità degli utenti Wikidata. Tale visualizzazione a grafo ha permesso di valorizzare i dati conservati all'interno di Wikidata attraverso la manifestazione di tutte le relazioni che intercorrono tra i diversi elementi. Questa visione diacronica della committenza Buonaccorsi può permettere di comprendere efficacemente le relazioni tra i dati e di conseguenze favorirne il loro riutilizzo da parte degli utenti (Fig. 1).

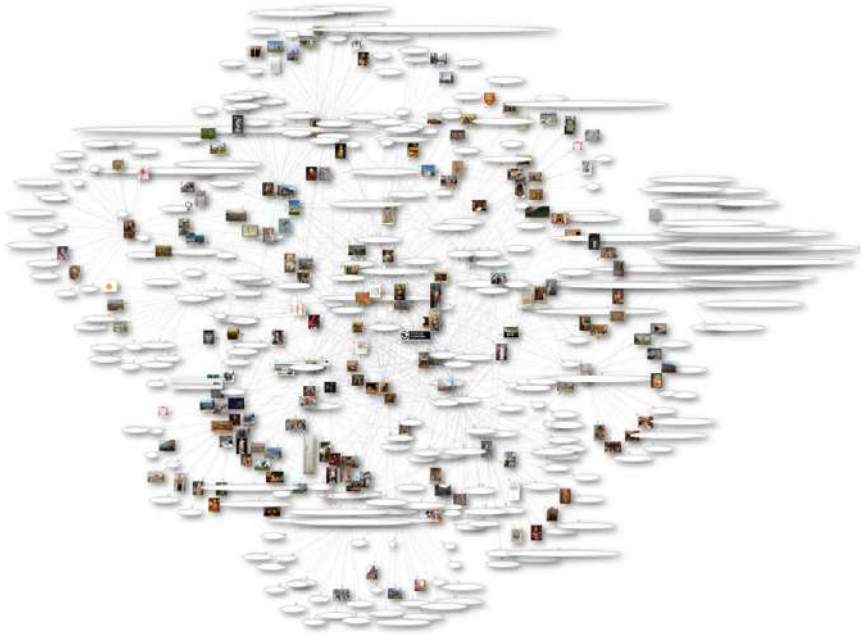


Figura 1 – Grafo del fenomeno mecenatesco della famiglia Buonaccorsi di Macerata, realizzato attraverso il *tool* Wikidata Query Service.

4. Conclusioni

In conclusione, questo approccio alla ricerca storico-artistica basato sulle tecnologie del web semantico e dei Linked Open Data può essere efficacemente applicato allo studio delle committenze artistiche delle famiglie aristocratiche, permettendo di restituire molti aspetti del complesso fenomeno mecenatesco

attraverso le capacità descrittive delle tecnologie semantiche. Ciò non implica che questa metodologia non possa essere applicata ad altri contesti storici complessi, come ad esempio la possibilità di descrivere in Wikidata le articolazioni di un archivio storico e le sue vicissitudini conservative, permettendo un'eventuale riconciliazione tra più tronconi di uno stesso fondo archivistico disperso tra diversi istituti, o la possibilità di analizzare, partendo da una fonte artistica, l'evoluzione nel tempo di un determinato soggetto iconografico, potendolo confrontare in tempo reale con l'originale e le diverse manifestazioni che lo raffigurano, facilitando di gran lunga il lavoro dello storico delle immagini. Inoltre, la possibilità offerta da *repository* online come Wikidata di accedere a un vasto patrimonio informativo, liberamente accessibile e riutilizzabile, può dare modo agli storici di qualsiasi estrazione di poter descrivere attraverso i dati la totalità dei contesti relativi ai loro campi di studio e di poterli così interrogare sotto diverse prospettive. Ciò permette di manifestare diacronicamente collegamenti prima inespressi e completamente ignorati, aprendo in questo modo nuovi punti di vista sul contesto studiato e dando la possibilità di sviluppare nuove prospettive di ricerca fino a questo momento inesplorate.

Riferimenti bibliografici

Alexiev, Vladimir, Plamen Tarkala, Nikola Georviev, e Lilia Pavlova. 2020. "Bulgarian Icons in Wikidata and EDM." *Digital Presentation and Preservation of Cultural and Scientific Heritage* 10, 2: 45-64. <https://doi.org/10.55630/dipp.2020.10.2>.

Barucca, Gabriele, e Alessandra Sfrappini, a cura di. 2001. *"Tutta per ordine dipinta." La Galleria dell'Eneide di Palazzo Buonaccorsi a Macerata*. Urbino: Quattroventi.

Bertacchini, Enrico, Enrico Ferraris, e Alice Fontana. 2022. "La fruizione delle collezioni digitali di beni archeologici: un'esplorazione delle immagini su Wikimedia Commons," DigitCult - Scientific Journal on Digital Cultures 7 (2): 99-106. <https://doi.org/10.36158/97888929562237>.

Bianchini, Carlo, e Lucia Sardo. 2022. "Wikidata: A New Perspective towards Universal Bibliographic Control." *JLIS.it. Italian Journal of Library, Archives and Information Science* 13, 1: 291-311. <https://doi.org/10.4403/jlis.it-12725>.

Bianchini, Carlo, e Pasquale Spinelli. 2020. "Wikidata at Fondazione Levi (Venice, Italy). A Case Study for the Publication of Data about Fondo Gambara, a Collection of 202 Musicians' Portraits" *JLIS.it. Italian Journal of Library, Archives and Information Science* 11, 3: 16-38. <https://doi.org/10.4403/jlis.it-12648>.

Capriotti, Giuseppe. 2018. "Prima dell'Eneide. Bacco ed Ercole nella decorazione delle stanze di Palazzo Buonaccorsi." in *La Galleria dell'Eneide di palazzo Buonaccorsi a Macerata. Nuove letture e prospettive di ricerca per il Settecento europeo*, a cura di Giuseppe Capriotti, Francesca Coltrinari, Patrizia Dragoni, Susanne A. Meyer, e Massimiliano Rossi, *IL CAPITALE CULTURALE. Studies on the Value of Cultural Heritage* 8: 335-67. <https://doi.org/10.13138/2039-2362/1975>.

Cardella, Lorenzo. 1793. *Memorie storiche de cardinali della santa romana chiesa scritte da Lorenzo Cardella parroco de' ss. Vincenzo ed Anastasio alla regola* VII. Roma: Stamperia Pagliarini.

Clavoud, Florence. 2019. "Transformer les métadonnées des Archives nationales en graphe de données: enjeux et premières réalisations." *La Gazette des archives* 54: 59-88.

Codogno, Maurizio. 2009. "Wikipedia, i Pokémon e la teoria della complessità nei sistemi emergenti." Convegno From Diderot to Wikipedia: an Epistemological Revolution, Padova, 16 ottobre 2009. <https://wiki.wikimedia.it/wiki/File:Diderot-codogno.odt>.

Colla, Davide, Goy Annamaria, Leontino Marco, e Diego Magro. 2021. "Wikidata Support in the Creation of Rich Semantic Metadata for Historical Archives." *Applied Sciences* 11, 10. <https://doi.org/10.3390/app11104378>.

Coltrinari, Francesca. 2018. "I Buonaccorsi committenti d'arte fra Macerata e l'Italia: geografia, dinamiche e relazioni." In *La Galleria dell'Eneide di palazzo Buonaccorsi a Macerata. Nuove letture e prospettive di ricerca per il Settecento europeo*, a cura di Giuseppe Capriotti, Francesca Coltrinari, Patrizia Dragoni, Susanne A. Meyer, e Massimiliano Rossi, *IL CAPITALE CULTURALE. Studies on the Value of Cultural Heritage* 8: 15-55, <https://doi.org/10.13138/2039-2362/1988>.

Costanzi, Costanza. 2005. "Le vie di fuga: principali percorsi nelle dispersioni delle opere d'arte dalle Marche." In *Le Marche disperse. Repertorio di opere d'arte dalle Marche al mondo*, a cura di Costanza Costanzi. Milano: Silvana Editoriale.

D'Angelo, Paolo. 2016, "Estetica come scienza dell'espressione e linguistica generale," in *Croce e Gentile*. Roma: Treccani. <https://www.treccani.it/enciclopedia/estetica-come-scienza-dell-espressione-e-linguistica-generale_(Croce-e-Gentile)/>.

Domenichini, Roberto. 2019. "La confraternita della morte ed orazione di Potenza Picena (Monte Santo), il suo archivio e la famiglia Compagnoni Marefoschi (secc. XVI-XIX)." In *Nuove indagini storico-archivistiche su Macerata e il suo territorio: dedicato a Pio Cartechini (1928-2016)*, Atti del convegno di studi maceratesi, Tolentino, Abbadia di Fiastra, 25-26 novembre 2017, 137-64. Macerata: Centro studi storici maceratesi.

Evenstein Sigalov, Shani, e Rafi Nachmias. 2023. "Investigating the potential of semantic web for education: Exploring Wikidata as a learning platform." *Education and Information Technologies* 28: 12565-614. <https://doi.org/10.1007/s10639-023-11664>.

Fagerwing, Alicia. 2023. "Wikidata for authority control: sharing museum knowledge with the world." *Digital Humanities in the Nordic and Baltic Countries Publications* 5, 1: 222-39. <https://doi.org/10.5617/dhnbpub.10665>.

Feliciati, Pierluigi. 2022a. "Call me by your name: towards an authority data control shared between archives and libraries." *JLIS.it. Italian Journal of Library, Archives and Information Science* 13,1: 203-14. <https://doi.org/10.4403/jlis.it-12733>.

Feliciati, Pierluigi. 2022b. "Names, things, places: towards a semantic, sustainable, usable integration?" *JLIS.it. Italian Journal of Library, Archives and Information Science* 13, 3: 145-53. <https://doi.org/10.36253/jlis.it-480>.

Haskell, Francis. 2020. *Mecenati e pittori. L'arte e la società italiana nell'epoca barocca*. Torino: Einaudi.

Ionna, Alessio. 2023. "Rappresentazione digitale semantica del contesto storico-culturale dell'isola di San Giorgio Maggiore e delle collezioni artistiche della Fondazione Giorgio Cini di Venezia." *IL CAPITALE CULTURALE. Studies on the Value of Cultural Heritage* 28: 557-73. <https://doi.org/10.13138/2039-2362/3288>.

Kapsalis, Effie. 2019. "Wikidata: Recruiting the Crowd to Power Access to Digital Archives." *Journal of Radio & Audio Media* 26, 1: 134-42. <https://doi.org/10.1080/19376529.2019.155952>.

Melatini, Antonella. 1998. *I Bonaccorsi. Storia di un grande casato*. Civitanova Marche: Communication words.

Montella, Massimo. 2016a. "Cultura: nozione umanistico-neoidealista." in *Economia e gestione dell'eredità culturale. Dizionario metodico essenziale*, a cura di Massimo Montella. Milano: Wolters Kluwer- CEDAM, 18-19.

Montella, Massimo. 2016b. "Cultura: nozione antropologica." in *Economia e gestione dell'eredità culturale. Dizionario metodico essenziale*, a cura di Massimo Montella, 19-20. Milano: Wolters Kluwer- CEDAM.

Monti, Ornella. 2013. "Il fondo Buonaccorsi della Biblioteca statale di Macerata." *RiMarcando* 8: 187-92.

Noiret, Serge. 2008. "Informatica, storia e storiografia: la storia si fa digitale." *Memoria e Ricerca* 28: 189-201.

Numerico, Teresa, Fioromonte, Domenico, e Francesca Tomasi. 2010. Introduzione a *L'umanista digitale* di Teresa Numerico, Domenico Fioromonte e Francesca Tomasi. Bologna: il Mulino, 7-15.

Paci, Deborah. 2021. "Una storia di incroci: scienze storiche e tecnologie informatiche." *Rivista di Ricerca e Didattica Digitale* 1: 13-30. <https://doi.org/10.53256/RRDD_210101>.

Pignatelli, Giuseppe. 1969. "BONACCORSI, Simone." *Dizionario Biografico degli Italiani*, 11. < https://www.treccani.it/enciclopedia/simone-bonaccorsi_(Dizionario-Biografico)/ >.

Roncaglia, Gino. 2002. "Informatica umanistica: le ragioni di una disciplina." *Intersezioni* 12, 3: 353-76. <https://dspace.unitus.it/server/api/core/bitstreams/64b7d55a-fc45-4cc7-a56a-eeb5c7d12bd0/content>.

Roncaglia, Gino. 2021. "Encyclopedias and encyclopedism in the era of the Web." *JLIS.it. Italian Journal of Library, Archives and Information Science* 12, 3: 69-90. <https://doi.org/10.4403/jlis.it-12757>.

van Veen, Theo. 2019. "Wikidata: From 'an' Identifier to "the" Identifier." *Information Technology and Libraries* 38, 2: 72-81. <https://doi.org/10.6017/ital.v38i2.10886>.

Vrandečić, Denny. 2013. "The rise of Wikidata." *IEEE Intelligent Systems* 28, 5: 90-95. <https://doi.org/10.1109/MIS.2013.119>.

Modeling Architectural Heritage in Wikidata: A Case Study of Early Modern European Hospitals

Frieder Leipold, Maximilian Kristen, Lily Marie Baumeister, Isabella Limmer, Chiara Franceschini

Abstract: The ERC project ARCHIATER (2024–29) uses Wikidata as a sustainable Virtual Research Environment to study the visual and architectural culture of premodern European hospitals. Through the WikiProject Spital, it develops shared standards for modeling, querying, and visualizing data. Building on the WikiFAIR initiative, the project advances FAIR principles and demonstrates how Wikidata fosters long-term accessibility and interdisciplinary collaboration.

Keywords: Wikidata, VRE, hospitals, WikiFAIR, data management, ARCHIATER.

1. Introduction

The storage devices in this image (Fig. 1) trace the evolution from vinyl records and punched cards to floppy disks, hard drives, CDs/DVDs, USB sticks, and memory cards. Now, ask yourself: which of these storage devices could I still read in my office today? While digital media enabled storage on an unprecedented scale, each shift risks data loss if information is not migrated[1].

This example highlights a pressing issue in the digital humanities: the fragility of research data. As digital methods become central to academia, long-term sustainability is threatened by software updates, changing formats, obsolete media, and the short lifespan of project-based databases that often vanish once funding ends.

[1] Cf. Jeff Rothenberg, "Ensuring the Longevity of Digital Information." *Scientific American* 272, no. 1 (January 1995): 42-47.

Frieder Leipold, University of Munich, Germany, F.Leipold@kunstgeschichte.uni-muenchen.de, 0000-0001-8848-9186
Maximilian Kristen, University of Munich, Germany, max@kristenonline.de, 0009-0008-3929-337X
Lily Marie Baumeister, University of Munich, Germany, L.Baumeister@campus.lmu.de
Isabella Limmer, University of Munich, Germany, isabella.Limmer@campus.lmu.de
Chiara Franceschini, University of Munich, Germany, Chiara.Franceschini@kunstgeschichte.uni-muenchen.de, 0000-0002-4836-7544

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Frieder Leipold, Maximilian Kristen, Lily Marie Baumeister, Isabella Limmer, Chiara Franceschini, *Modeling Architectural Heritage in Wikidata: A Case Study of Early Modern European Hospitals*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.21, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 165-174, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

The ARCHIATER project, funded by the European Research Council, responds to this crisis by treating Wikidata not as a passive archive but as a collaborative, sustainable Virtual Research Environment (VRE). It integrates Wikidata to document and analyse the visual and architectural culture of premodern European hospitals, showing how the platform fosters long-term collaboration, transparency, and innovation through both qualitative and quantitative approaches.

Drawing on lessons learned from earlier digital projects—such as the 3D reconstruction of Schloss Weikersheim and the DeJongeWiki on Schloss Arenberg—the ERC project ARCHIATER (2024–29) joins the broader WikiFAIR initiative. This initiative advocates for academic data to be hosted on Wikimedia platforms, ensuring compliance with the FAIR principles: Findable, Accessible, Interoperable, and Reusable[2]. We argue that Wikidata provides a robust infrastructure for the humanities and can serve as a model for future digital research projects.

Figure 1 – Evolution of Storage Media: Records, Punched Cards, Floppies, Hard Disk, DVD/CD, USB Flash Drives, Memory Cards; Diptangshudatta on Wikimedia Commons[3], licensed under CC-BY-SA-4.0.

[2] Cf. FAIR Principles, <https://www.go-fair.org/fair-principles/>.

[3] <https://commons.wikimedia.org/wiki/File:Evolution_of_Data_Storage.jpg>.

2. The ARCHIATER Project: goals and framework

ARCHIATER, based at Ludwig-Maximilians-Universität München explores the architectural, artistic, and religious dimensions of hospital buildings from the 14th to 18th centuries. These institutions were far more than centres of medical care. Focusing on the countries of pre-modern Europe, ARCHIATER explores how hospitals and their visual culture reflected and shaped their urban environments. While differing in regional contexts, these institutions share common traits as complex cultural nodes.

So, we are dealing with a research subject that is vast both temporally and geographically, a subject that developed very complex structures. Additionally, ARCHIATER is a collaboration between scholars at the universities of Munich in Germany and Lucca in Italy. This raised the question of how the ongoing research progress could be documented and how the emerging data connections could be stored in the medium term.

Although there are highly advanced tools for data processing, even in a scientific context, these offerings are often associated with paid licenses. If you no longer pay, you lose access. Furthermore, you are dependent on the financial stability of the respective service provider.

3. Data fragility in historical research

Despite the growing sophistication of digital tools, many research datasets have a short shelf life. Projects often develop bespoke databases or digital platforms tailored to their research questions. These platforms may offer short-term functionality, but they are rarely maintained after funding ends. Without continued investment, they fall victim to outdated software, broken links, and incompatible data structures.

These problems are compounded by data silos. Custom platforms are often poorly integrated with one another. Their data cannot be easily shared, reused, or linked to broader knowledge networks. The result is a pattern of digital ephemerality: valuable research outputs vanish into obscurity.

To counteract this, ARCHIATER chose to forgo a custom-built infrastructure. Instead, the project embraced Wikidata—a free, open, multilingual, and structured knowledge base maintained by the Wikimedia Foundation. This decision was not made lightly but built upon previous experiences that revealed both the potentials and the limitations of traditional digital infrastructures.

4. Learning from previous projects: toward WikiFAIR

The decision to use Wikidata was shaped by prior experiences in collaborative digital heritage projects. One of these was the digital 3D reconstruction of Schloss Weikersheim, part of the initiative "Virtuelle Rekonstruktion: Kulturliegenschaften gestern und heute," a joint effort by the Staatliche Schlösser und Gärten Baden-Württemberg and the Corpus der barocken Deckenmalerei

at LMU München[4]. This project used archival documents and architectural analysis to recreate lost historical interiors. While it succeeded in producing compelling scholarly reconstructions, the issue of long-term and open access to the data hosted in a WissKI environment remained challenging.

Another influential project was DeJongeWiki, a digital platform developed to document the architectural history of the Kasteel van Arenberg in Heverlee, Belgium[5]. The DeJongeWiki was a side initiative of the Horizon 2020 project PALAMUSTO and the Raymond Lemaire International Centre for Conservation at KU Leuven[6]. It aimed to collate building records, historical images, and scholarly commentary into a comprehensive resource. However, questions of long-term sustainability in a hybrid MediaWiki–Wikibase environment with a complementary semantic database loomed large.

These experiences lead to the development of WikiFAIR[7], a grassroots initiative that advocates for academic data to be hosted within Wikimedia platforms. WikiFAIR emphasizes the value of Wikidata, Wikimedia Commons, and related projects as environments that fulfil the FAIR data principles. By aligning with open-access infrastructure maintained by a global community, scholars can ensure that their data lives on.

ARCHIATER embraced this logic. Rather than investing in another standalone database, the project opted to contribute to the global knowledge graph of Wikidata—an ecosystem that is both robust and sustainable.

5. Wikidata as research infrastructure

Wikidata offers a unique set of features that make it highly suitable as a Virtual Research Environment (VRE). Its openness, with all structured data available under the CC0 license, ensures free and unrestricted access to data, encouraging widespread use and integration. This openness fosters transparency and collaboration across the global research community. Sustainability is another key advantage. Hosted by the Wikimedia Foundation, Wikidata benefits from long-term support, making it more reliable than many other digital platforms that rely on project-specific funding. Researchers can be confident that their data will remain accessible for the long term, addressing a significant challenge in digital humanities—data preservation.

Wikidata also excels in version control, as every edit is tracked, allowing scholars to monitor changes and maintain data integrity over time. This feature supports academic rigor by providing a clear record of revisions and facil-

[4] Virtuelle Rekonstruktion: Kulturliegenschaften gestern und heute, <https://www.vr-ssg.hs-mainz.de/>.

[5] DeJongeWiki, mainpage, <https://set.kuleuven.be/rlicc/dejongewiki/w/index.php/Main_Page>.

[6] Webpage of the PALAMUSTO project, <https://www.palamusto.eu/blog>; webpage of the Raymond Lemaire International Centre for Conservation, <https://set.kuleuven.be/rlicc>.

[7] WikiFAIR, project page on Meta-Wiki, <https://meta.wikimedia.org/wiki/WikiFAIR>.

itating collaborative work. Its interoperability as part of the linked open data (LOD) cloud allows seamless connections with other datasets, enabling data sharing, comparison and integration. For interdisciplinary projects like ARCHIATER, this feature is invaluable for connecting research across different fields and regions.

Additionally, Wikidata's multilingualism enables researchers from diverse linguistic backgrounds to collaborate effectively. Each entity can have labels and descriptions in multiple languages, promoting international cooperation. Finally, community curation ensures that Wikidata's data remains up-to-date and of high quality. A global network of volunteers and researchers contributes to the continuous improvement and development of entries, enhancing the platform's reliability and accuracy, with more than 23.000 active users per month (a total of 660.000 users since its inception) and over 400.000 edits per day (by users and bots)[8].

For projects like ARCHIATER, which involve collaboration between scholars from different countries and disciplines, these features make Wikidata not just a data repository but a dynamic platform for collaborative knowledge production. Working with Wikidata began by identifying datasets on premodern hospitals, a challenge since they were modeled from diverse perspectives. These institutions also varied greatly in purpose and structure, so the project defined several Wikidata entities representing distinct hospital subtypes such as poorhouse (Q1929784), orphanage (Q160645), leper colony (Q1406569), lazaretto (Q859247) or Bimarestan (Q1293141).

However, it should also be noted that institutions have had various uses over the centuries, sometimes even simultaneously. To create a prototype whose modeling style can be referenced for other hospitals, the Heilig-Geist-Spital in Munich (Q1594901) was chosen. This hospital has been used as a poorhouse (Q1929784), a pilgrims' hostel (Q2094954), a poor hospital (Q48689779), or an orphanage (Q160645).

6. The WikiProject Spital: structure and goals

In the next step, a model of the data for the Heilig-Geist-Spital in Munich was designed that was as logical and understandable as possible. The relevant entities for the criteria and the properties for the attributes were applied not only in the Wikidata dataset of the Heilig-Geist-Spital in Munich. To effectively coordinate the contributions to Wikidata, ARCHIATER launched the WikiProject Spital[9].

Essential information such as the founding year, religious affiliation, architectural features, and their desired coding are listed there. The fundamental decision was made to view hospitals primarily as buildings (Q41176) or building complexes (Q1497364), which then have a specific use (P366). If new entries are created by the project, this will be done according to these guidelines. If in-

[8] Cf. <https://wikidata.wikiscan.org/>.

[9] Webpage, <https://www.wikidata.org/wiki/Wikidata:WikiProject_Spital>.

formation is added to existing datasets, existing models should remain as untouched as possible, and data should be entered additionally according to the WikiProject Hospital.

A search of Wikidata reveals that over 2.000 entries on Wikidata deal with pre-modern hospitals or facilities belonging to a subcategory. However, a geographical analysis also reveals a blatant research bias, focused on the Cold War Western Europe. Given the extent of the missing data, the main focus of the ARCHIATER project is to map existing research and supplement it with our own findings. A comprehensive, thorough survey of the entire European region from the 14th to the 18th century cannot be achieved within the scope of the ARCHIATER project—considering that cities like Paris, Venice, and Florence alone had hundreds of hospitals over time.

The WikiProject Spital also focuses on building a repository of SPARQL queries, which will support data analysis and visualization. These queries allow researchers to extract specific data from the Wikidata platform, enabling them to generate visual representations and identify patterns and trends within the historical hospital data. These visualizations help to uncover insights that would otherwise be difficult to see in raw data form.

7. Visualizing data with SPARQL and GIS

Figure 2 – Distribution of the «Lazzaretto» and «Leprosarium» datasets around the Mediterranean, Interactive GIS map as a visualization of the SPARQL query[10].

Wikidata's support for SPARQL enables ARCHIATER to conduct quantitative analysis on its dataset. Through custom queries, researchers can generate maps and timelines that reveal historical patterns. For instance:

- Mapping the spread of lazarettos in response to the Black Death and connecting them with researched locations of leper colonies. This approach provides insight into how the sea was not only a conduit for trade and com-

[10] <https://w.wiki/DnUm>.

munication between societies, but also served as a barrier to separate and isolate the ill (Fig. 2).
- Visualizing the density of documented Holy Spirit Hospitals (Q1594896) in Southern Germany compared to Northern Italy[11].
- A search for the Late Antique and Byzantine type of xenodocheion (Q1517445) suggests that these may no longer be locatable in the eastern Mediterranean region. Instead, pilgrim accommodations along the Way of St. James in Spain are categorized as such[12].

These visualizations, often presented as GIS-based maps, allow complex patterns to emerge from what would otherwise remain isolated data points. They are not only valuable for academic research but also serve as powerful tools for teaching and public outreach.

Each of these hospital subtypes is represented in Wikidata with clearly defined properties and relationships, allowing for a detailed and nuanced understanding of each type's unique role in society. By modelling these differences explicitly, the project enables richer analysis and meaningful comparisons across time and space, fostering a deeper understanding of how medical care was organized and provided in various historical contexts. This classification also contributes to a more structured approach to historical research, where each institution can be examined within its specific context and compared to similar institutions in different regions or periods.

8. Modeling artistic and social networks

Quantitative methods in ARCHIATER are significantly enriched by qualitative data modeling, particularly when it comes to representing complex cultural artifacts such as artworks. Wikidata provides the infrastructure to build detailed knowledge graphs that link hospitals to artworks, their patrons, historical events, and associated figures. This is especially valuable in documenting the often-fragmented histories of artworks that have undergone changes in location, context, or condition.

Here the Heilig-Geist-Spital in Munich (Q1594901) serves as an example again. Using Wikidata, this institution can be linked to the ceiling painting by Cosmas Damian Asam (Q715693), completed in 1727, which was later destroyed during World War II. A subsequent reconstruction of the artwork was carried out by Karl Manninger (Q1365373) from 1971 to 1975. Additional links include the religious order responsible for managing the hospital and its artistic program (Fig. 3).

Such modeling is not without challenges. Artworks are not static entities—they may be relocated, altered, restored, or recontextualized in different institu-

[11] Link to SPARQL query, <https://w.wiki/DrwA>.

[12] Link to SPARQL query, <https://w.wiki/DrwR>.

tional or liturgical settings. Their physical condition may change over time, and documentation is often incomplete or dispersed. Wikidata allows these multiple, sometimes conflicting, layers of information to coexist through qualifiers, temporal markers, and references. Each element—be it an artist, a patron, a specific date, or a war-related destruction—becomes a node in a network of semantic relationships.

This approach allows researchers to trace the life cycle of an artwork across time and space, linking it not only to the institution that originally housed it but also to broader historical and social developments. The resulting knowledge graph enables a level of complexity and nuance that traditional textual scholarship alone often cannot achieve, making Wikidata a powerful tool for humanities research.

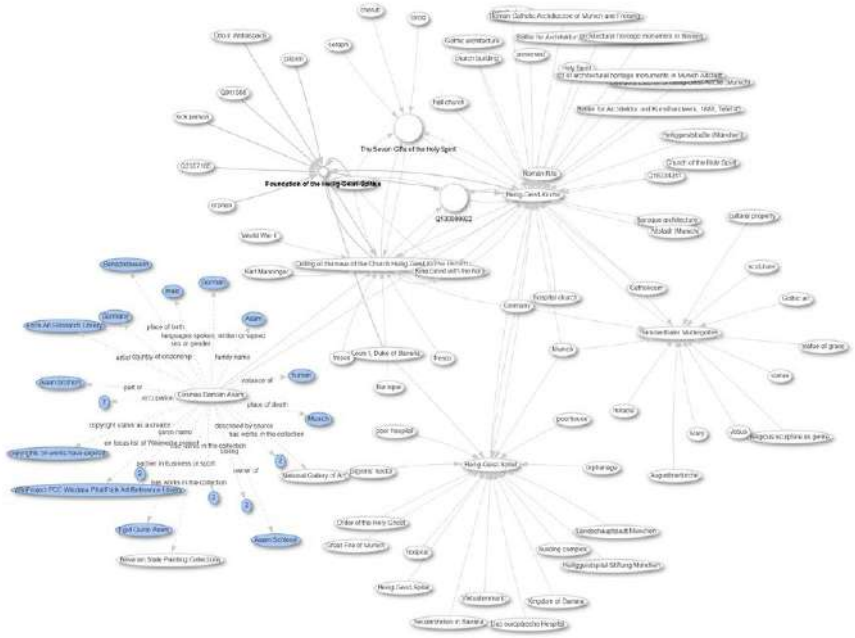


Figure 3 – Relationship network of the painter Cosmas Damian Asam in connection with the Heilig-Geist-Spital in Munich, presented as an interactive graph generated through a SPARQL query[13].

9. International and interdisciplinary collaboration

WikiProject Spital and its guidelines also address one of the main challenges of the ARCHIATER project as a collective of international researchers

[13] <https://w.wiki/DnUs>.

in Germany and Italy from diverse disciplinary backgrounds. Wikidata provides a shared platform that bridges methodological and linguistic divides. Italian researchers, for example, contribute detailed archival knowledge and expertise in iconography, while German partners bring skills in digital modeling and architectural analysis. The multilingual structure of Wikidata enables each team to work in their preferred language while contributing to a shared data ecosystem.

This collaboration also fosters epistemic diversity. Rather than imposing a single interpretive framework, Wikidata allows for multiple, coexisting perspectives. Wikidata's editorial infrastructure also makes it possible to specifically track which changes were made by which user. In this way, in the spirit of a Virtual Research Environment (VRE), it is possible to track which colleague in the project is working on which objects and with which focus.

In this sense, ARCHIATER also aims to lay the groundwork for similar projects. It aims to facilitate methodological dialogue among researchers from various academic traditions. To explore Wikidata a platform where scholars with different approaches to studying historical hospitals—whether in the fields of art history, architecture, religious studies, or digital humanities—can collaborate and share their findings. By creating this shared space, ARCHIATER fosters a collaborative environment where individual insights contribute to a growing corpus of structured, linked data. The project empowers scholars to contribute locally while thinking globally.

10. Conclusion

ARCHIATER shows how Wikidata can serve as a sustainable Virtual Research Environment, avoiding the obsolescence of isolated databases and fostering international collaboration. Building on the WikiFAIR initiative and through the WikiProject Spital, the project develops shared standards for modeling, querying, and visualizing premodern hospitals. In this way, Wikidata becomes not just a repository but an active partner in knowledge creation, advancing FAIR principles and contributing to a more open, interconnected research landscape.

References

Albers, Laura. 2020. "Die semantische Datenmodellierung im Corpus der barocken Deckenmalerei in Deutschland." In *Digitale Raumdarstellung: Barocke Deckenmalerei und Virtual Reality*, edited by Stephan Hoppe, Hubert Locher, and Matteo Burioni, 224–55. Heidelberg: arthistoricum.net (Computing in Art and Architecture; 4). <https://doi.org/10.11588/arthistoricum.774.c10131>.

Leipold, Frieder, and Jan-Eric Lutteroth. 2020. "Schloss Weikersheim in 3D. Neue Zugänge zur Architekturgeschichte der nordalpinen Spätrenaissance." In *Digitale Raumdarstellung: Barocke Deckenmalerei und Virtual Reality*, edited by Stephan Hoppe, Hubert Locher, and Matteo Burioni, 132–57. Heidelberg: arthistoricum.net (Computing in Art and Architecture; 4). <https://doi.org/10.11588/arthistoricum.774.c10127>.

Leipold, Frieder, Max Kristen, and Krista De Jonge. 2023. "Knowledge for Men and Machines: The De Jonge Wiki as an Example of a Scientific Research Database." *Zeitschrift für digitale Geisteswissenschaften* 8. <https://zfdg.de/sites/default/files/pdf_uploads/2023_007_leipold_et_al_dejonge_v1_1.pdf>.

Leipold, Frieder. 2022. "Futuristische software voor een oud kasteel." *Geniaal* 56: 24–25.

Lutteroth, Jan-Eric, and Stephan Hoppe. 2018. "Schloss Friedrichstein 2.0 — Von digitalen 3D-Modellen und dem Spinnen eines semantischen Graphen." In *Computing Art Reader: Einführung in die digitale Kunstgeschichte*, edited by Piotr Kuroczyński, Peter Bell, and Lisa Dieckmann, 184–98. Heidelberg: arthistoricum.net (Computing in Art and Architecture; 1). <https://books.ub.uni-heidelberg.de//arthistoricum/catalog/book/413/chapter/5822>.

Schelbert, Georg. 2017. "... warum nicht gleich Wikidata?!" Abstract for the *Digital Humanities im deutschsprachigen Raum* conference.

Wintergrün, Dirk. 2019. "Netzwerkanalysen und semantische Datenmodellierung als heuristische Instrumente für die historische Forschung." PhD diss., Friedrich-Alexander-Universität Erlangen-Nürnberg. <https://open.fau.de/items/c2f6a877-f011-444f-be3e-8456b7ea5ba9>.

A Data Visualization Tool for Heritage Buildings in India and Bangladesh: The Project of a MediaWiki-based Platform for ID-SCAPES

Giuseppe Resta, Sidh Losa Mendiratta, Tiago Trindade Cruz

Abstract: The ID-SCAPES research project studies medieval and early modern religious architecture of Christian minorities in India and Bangladesh, often representing contested heritage. A key output will be *Herichurch*, a MediaWiki-based platform designed to present historical entries on significant churches, making information accessible to a broader audience. The Wiki will integrate texts, photographs, 3D visualizations, and drawings for effective cultural heritage dissemination. Drawing lessons from successful case studies, *Herichurch* Wiki intends to explore the connection between a MediaWiki encyclopaedia and a structured database to produce accurate and accessible content about heritage buildings.

Keywords: Digital Humanities, heritage buildings, India, churches, preservation by record.

1. Introduction

In late 2023, the European Research Council approved funding for the ID-SCAPES project, which aims to study and document the medieval and early modern religious architecture of Christian minorities in India and Bangladesh (Fig. 1). These religious sites are often multi-layered and contested heritage, and some buildings have suffered from increasing neglect, erasure and even effacement during recent decades. Furthermore, following the end of colonial rule in those two countries, many within the Church hierarchies sought to distance themselves from colonial legacies, and many of the older churches were radically transformed or completely rebuilt with contemporary designs. Taking into consideration the risks arising from these processes and the progressive erasure of cultural heritage, ID-SCAPES aims to produce a social history of the built environment of India and Bangladesh's medieval and early modern churches and sacral landscapes (built before ca. 1800), including functioning, ruined, and disappeared buildings.

ID-SCAPES challenges the European-centric historiographical framework that has traditionally been applied to these topics. Through extensive

Giuseppe Resta, University of Porto, Portugal, gresta@arq.up.pt, 0000-0001-8489-5291
Sidh Losa Mendiratta, University of Porto, Portugal, smendiratta@arq.up.pt, 0000-0003-2960-8100
Tiago Trindade Cruz, University of Porto, Portugal, tcruz@arq.up.pt, 0000-0001-8792-5142

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giuseppe Resta, Sidh Losa Mendiratta, Tiago Trindade Cruz, *A Data Visualization Tool for Heritage Buildings in India and Bangladesh: The Project of a MediaWiki-based Platform for ID-SCAPES*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.22, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 175-183, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

fieldwork and the analysis of previously unexplored visual and written documents, this research project promotes a new methodological approach that acknowledges the complex histories of these buildings. It addresses issues such as cultural encounters, caste, religious accommodation, indigenous agency, and local spatial and artistic traditions, revealing how these factors have influenced church architecture.

One of the principal outputs of the project will be *Herichurch.org*, a MediaWiki-based platform intended to provide entries with historical information on a selection of significant churches, accessible to a wider audience. Following the idea of "preservation by record", the project's Wiki will combine texts, photographs, 3D visualizations and CAD drawings as fundamental assets for cultural heritage dissemination, conservation and research.

In this paper, we highlight the typologies of data that our project will address. Additionally, we review several comparable projects in the field of digital humanities that have utilized the Wikidata database. Furthermore, we examine the common challenges faced by these projects in relation to IDSCAPES. Finally, we outline a potential roadmap for integrating the project outputs into an indexed collection of church architectures, organized in a MediaWiki-based platform, with a focus on how they can be published to reach a broader audience.

Figure 1 – Church of Our Lady of Divine Providence, Goa, India (credits: Sidh Losa Mendiratta, 2020).

2. ID-SCAPES structure of knowledge: from photo archives to CAD drawings

The ID-SCAPES project will create a comprehensive collection of datasets focused on historical, architectural, and geographical research gathered from archival sources and fieldwork (Fig. 2). These datasets will facilitate analysis and interpretation of various subjects. A significant portion of the data generated regarding the religious architecture and sacred landscapes of Christian minorities in India and Bangladesh will be synthesized and made accessible to a broad audience through the *Herichurch.org* Wiki. The online resource will serve as a semi-structured georeferenced database in open access that can be collaboratively edited, ensuring interoperability and access to the information sourced from all the project's datasets. *Herichurch.org* and other digital tools created by ID-SCAPES can impact heritage management interventions, especially in the case of endangered or contested sites.

During the project's timeline (2024–29), ID-SCAPES will curate six datasets that will contribute to the *Herichurch.org* entries:

1) Photographs (~12000 project photographs taken during previous and forthcoming fieldwork in India and Bangladesh and ~18500 historical photographs acquired from private collections in Europe, South Asia and the USA),
2) Cartographic / Iconographic documents (~250 manuscripts and ~100 printed documents acquired from private collections in Europe, South Asia and the USA),
3) CAD drawings (~120 two-dimensional visuals and ~30 three-dimensional visuals generated by the project),
4) Project map (1 map of georeferenced sites linked with the individual entries on *Herichurch.org*),
5) Bibliographic resources (~500 manuscript sources, ~500 printed sources, and ~50 published outputs collected and generated by the project, organized with a reference management software),
6) Oral Sources (~50 field work interviews).

Datasets 1 and 2 contain image files in PDF, JPEG and TIFF formats; dataset 3 vector files in DWG and DWF formats; dataset 4 Geopackage files for geospatial information; dataset 5 both ENL library format and PDFs; dataset 6 audio files in MP4 format.

Published outputs and *Herichurch.org* entries will be assigned persistent identifiers (DOIs). Data gathered from archival and private sources will have restricted access, while the Creative Commons Attribution 4.0 International (CC BY 4.0) license will apply exclusively to the project photography, allowing for broad reuse. Datasets 1 and 2 are provided with ~15 metadata fields to describe each item.

For the functioning of the *Herichurch.org* Wiki, each entry is associated with one ID-SCAPES inventory number; geographic information (coordinate location, country, state, district, sub-division, city, locator map image); religious

significance (ethnolinguistic group, parish, past parish, diocese, dedicated to); architectural quality (church in use, conservation status, original typology, present typology, built area, heritage designation); historical facts (past dedication, inception date, rebuilding date, Christian liturgical rite, original religious order); contact information (official website, present religious order, current priest); and licensing information. All these fields will use standardized Wikidata property IDs. Also, the web layout will take into consideration that most internet connections in India and Bangladesh are accessed via mobile devices (Waghmare 2025).

Figure 2 – Sample photo from a fieldwork at the church of Our Lady of Mount Carmel, Goa, India (credits: Sidh Losa Mendiratta, 2017).

To illustrate the *Herichurch.org* entries, we will use photographs, 3D visualizations and CAD drawings generated by ID-SCAPES, and visual documents already existing in Wikimedia Commons. Regarding visual documents from institutions and private collections, the licensing will be negotiated on a case-by-case basis, image by image. This process is coherent with the activities postulated for standard data curation: description, annotation, collection/aggregation, storage, and migration (Sabharwal 2015, 20).

ID-SCAPES intends to organise this knowledge structure through Wikibase, the software for managing structured data developed by Wikimedia Deutschland. The Wikibase Data Model is particularly significant because it allows for the enrichment of the basic semantic triple with references (sources) and qualifiers, which better represent ranges or contrasting perspectives on historical

facts (Bacchi and Bergamin 2018). This applies, for example, when the foundation date of a church can only be defined within a specific interval, as reported by certain authors. Additionally, embedded maps allow for Wikibase projects to be searched both in georeferenced and alphanumeric methods. These methods are currently employed in digital humanities in order to make data as interoperable as possible and knowledge dissemination as wide as possible (Zhao 2023).

3. Wikidata and georeferenced history

The preservation by record approach is preferable where "preservation *in situ* (keeping a site in an unchanged state) is impossible or impractical and involves making a detailed and comprehensive record before the material is lost" (Janik 2014, 112). The research theme of ID-SCAPES is in a geography different from that of the host institution, focusing on a body of cultural heritage that is often associated with colonialism in South Asia. There is a sense of urgency in documenting these sites, given that many of these structures have disappeared, and that present government policies in South Asia, generally speaking, don't prioritize the research and preservation of minority heritage (Sahgal et al. 2021; Thames and Chaudhry 2014). Thus, studying early modern religious architecture using an innovative digital platform that captures and engages with the complex histories of these significant cultural sites is more effective and fitting.

The project faces typical challenges associated with research practices in digital humanities. Sabharwal (2015, 31–35) reviews the criticism that was voiced after the publication of the first *Digital Humanities Manifesto* in 2008, addressing the renewed importance of archives for the humanities. As shown with ID-SCAPES's structure of knowledge, the project's archive spans across multiple formats and modes of systematization.

Flanders and Muñoz (2011) outline three methods for data treatment in the humanities: interpretive layering, data capture and preparation, and capturing scholarly agency, each reflecting different levels of data curation. These methods emphasize the importance of documentation, curatorial decisions, and the contextualization of original works to encourage new interpretations of sources. Focusing on historical phenomena, access to sources has evolved from being limited to a few scholarly publications to now offering ubiquitous access to extensively digitized collections. However, such digital ephemera cause "implications for archives and libraries" that warn "about data loss due to crises, lack of control, or emerging political landscapes intolerant of open knowledge" (Sabharwal 2015, 51). The cyberattacks on the *Internet Archive* that rendered the digital library unavailable between May and October 2024 (Wikipedia contributors 2025), and on the British Library in October 2023 are two examples of such crises.

In the archive sector, Wikidata has proven to be an essential tool for expanding the infrastructure of linked archived data. Cultural heritage institutions, such as the Museum of Modern Art, the Library of Congress, and the Smithsonian, have successfully established a two-way integration of their authority data with Wikidata. This integration benefits both their institutional websites and allows

for open access to a portion of their collections through Wikipedia (Hawkins 2022). The same drive is observed in recent research projects.

As shown in the analysis by Zhao (2023), Wikidata for Digital Humanities "is evolving as a linking hub of data" in a direction that can integrate original datasets with existing ones. For instance, the COURAGE project faced the issue of matching pre-existing entities of people, groups, and events pertaining to political opposition in the Socialist era, with the possibility to link project-specific individuals or organizations to a wider network of authority IDs (Faraj and Micsik 2019). Reducing data redundancy was one of the main issues that led to the release of Wikidata in 2012. ID-SCAPES will similarly face the problem of integrating some existing data on Indian and Bangladeshi churches with more granular information found in private and public collections. Rossenova, Duchesne, and Blümel (2022) analyze two ways of linking Wikidata and independent Wikibases in cultural heritage research projects. One can prefer the networked, standardized option of Wikidata, as in the case of the *DigAMus goes Wikidata* project (von Heyl, Thiel, and Schlott 2025), where items or properties are already present in Wikidata and only need curation. Alternatively, a private Wikibase containing customized historical information can be interlinked with Wikidata entries, as in the *Corpus der barocken Deckenmalerei in Deutschland* project (Vogel and Moises 2021), in which archival material, 2D drawings, photos, and 3D models datasets complement standard Wikidata items such as places and historical figures. This second approach better describes how the ID-SCAPES structure of knowledge could be ideally rendered.

Another relevant nucleus of reflection for the project is the mapping of georeferenced information linked to the project's Wiki. Many of the buildings under research have disappeared, are in ruins, or have been deeply transformed, and there is a sense of urgency in documenting whatever structures remain. This leads to four possibilities:

- The contemporary building, in use or ruins, is the same as described in sources from Datasets 2 and 5, and its location is certain,
- The contemporary building represents a significant alteration of the original structure that can be found in sources from Datasets 2 and 5, with its location also being certain,
- The original building left no traces, but Datasets 1, 2, and 5 allow geographical cross-examination to pin an approximate location,
- The original building left no material traces and only generic positional information in Dataset 5, i.e. the name of the parish. The location can only be identified with an area and referenced to literary or oral sources.

The work involving spatial approximations and uncertainties aims to balance scientific reliability with the need to provide consistent information across the set of ID-SCAPES architectural example. The authors have previously experimented with methods from Historical GIS to map buildings that no longer exist (Resta 2025). This was done in parallel with the cross-examination of historical visual documents to support fragmented information from written sources

(Mendiratta, Resta, and Cruz 2025; Mendiratta, Resta, and Reis Leite 2024). The integration of GIS with the semantic web in a web GIS suite interlinks archival documents in a mapping environment. The *Ghent Mapped* project clearly illustrates the challenges mentioned earlier. It provides "a public history project that spatially aggregates and maps cultural heritage records—predominantly movable heritage—from seven Ghentian heritage organizations using a custom web GIS" (Ducatteeuw, Danniau, and Verbruggen 2025). Diverse datasets, depicting both national landmarks and everyday places with material provided by the communities, are organized in a web map featuring a timeline that spans from 1550 to the present day. The authors have represented those places that have disappeared using an approximate polygon shape derived from historical documents. Boundaries are defined as the spatial domains that interact with each specific building (Ducatteeuw, Danniau, and Verbruggen 2025). In ID-SCAPES, one church without a specific location can be coupled with the corresponding religious jurisdiction that is usually available in registries. Alternatively, this spatial domain can correspond to an area whose offset reflects the degree of certainty that sources allow in each case, constituting one additional parameter to be associated with the geocoordinates.

4. Herichurch roadmap

At present, the development of the *Herichurch.org* Wiki is in its initial stage. The research team is considering the best strategy for making the project's findings easily accessible. The goal is to create a practical tool that can be used on-site and visualized on a smartphone. As previously discussed, two main concerns for the forthcoming Wiki are how to link the project's data with existing Wikidata entries and how to georeference historical records. These decisions pass through a thorough study of best practices in the sector of DH that ID-SCAPES initiated in March 2025. All the datasets presented in this paper are being consolidated in the first years of the project, while the elaboration of the Wiki will start in 2026. This allows time to define the appropriate metadata structure that better integrates with the tools to be employed.

Some degree of connection with Wikidata entities is necessary to ensure interoperability and the reusability of the Wiki output by third parties such as research institutions, administrations, and associations.

The analysis of *eViterbo,* which is the MediaWiki-based research output of the *TechNetEMPIRE* project, provides significant insights. This project focuses on the network of individuals—such as engineers, architects, cartographers, carpenters, and masons—who contributed to the formation of the Portuguese colonial empire. For this reason, it is adjacent to ID-SCAPES both in content and methodology. The limitations and challenges addressed in *eViterbo*'s global report (Faria and Correia 2022) could be a starting point to set the *Herichurch* Wiki in the right direction. We believe that several issues should be addressed in the initial phases of the project. These include the project's infobox template script, the alignment of the project's properties with Wikidata, redundant prop-

erties in data ontology, the implementation of *WikibaseQualityConstraints*, and proper data modeling.

Entries will likely link to existing entries in Wikipedia, *eViterbo*, and *Heritage of Portuguese Influence* (HPIP). HPIP is an online encyclopedic resource that maps and describes the architectural and urban heritage of Portuguese influence in various regions around the world. This project was initiated in 2007 by the Calouste Gulbenkian Foundation, and published as a four-volume work in 2011.

In summary, *Herichurch.org* Wiki intends to explore further this connection between a MediaWiki-based encyclopedia and a structured database to produce accurate and accessible content about India and Bangladesh's medieval and early modern churches and sacral landscapes. In this initial phase, significant efforts are devoted to designing three aspects of the authoring and publishing platform: integrating a selection of existing data on Indian churches with more granular information found in private and public collections; setting a privately hosted Wikibase containing customized historical information interlinked with Wikidata entries; mapping georeferenced historical records linked to the project's findings. To align with the spirit of Wikimedia projects, analyzing successful case studies and setbacks from similar experiences was crucial for taking a more informed and strategic approach to develop the *Herichurch.org* project.

References

Bacchi, Cristian, and Giovanni Bergamin. 2018. "New Ways of Creating and Sharing Bibliographic Information: An Experiment of Using the Wikibase Data Model for UNIMARC Data." *JLIS: Italian Journal of Library, Archives and Information Science* 9 (3): 35–74. <https://doi.org/10.4403/jlis.it-12458>.

Ducatteeuw, Vincent, Fien Danniau, and Christophe Verbruggen. 2025. "Mapping Ghent's Cultural Heritage: A Place-Based Approach with Web GIS." *International Journal of Digital Humanities*. <https://doi.org/10.1007/s42803-025-00099-4>.

Faraj, Ghazal, and András Micsik. 2019. "Enriching Wikidata with Cultural Heritage Data from the COURAGE Project." In *Metadata and Semantic Research*, edited by Emmanouel Garoufallou, Francesca Fallucchi, and Ernesto William De Luca, 407–18. Cham: Springer International Publishing.

Faria, Alice Santiago, and Rute Correia. 2022. *eViterbo Global Report. Outlines on the Creation of a Research Open Platform Using MediaWiki and Wikibase*. Zenodo. https://doi.org/10.5281/zenodo.7380114.

Flanders, Julia, and Trevor Muñoz. 2011. "An Introduction to Humanities Data Curation." *DH Curation Guide: A Community Resource Guide to Data Curation in the Digital Humanities*. <https://archive.mith.umd.edu/dhcuration-guide/guide.dhcuration.org/glossary/intro/index.html>.

Hawkins, Ashleigh. 2022. "Archives, Linked Data and the Digital Humanities: Increasing Access to Digitised and Born-Digital Archives via the Semantic Web." *Archival Science* 22 (3): 319–44. <https://doi.org/10.1007/s10502-021-09381-0>.

Janik, Liliana. 2014. "'Preservation by Record': The Case from Eastern Scandinavia." In *Open-Air Rock-Art Conservation and Management: State of the Art and Future Perspectives*, edited by Timothy Darvill and Antonio Batarda Fernandes, 112–24. New York-London: Taylor and Francis.

Mendiratta, Sidh Losa, Giuseppe Resta, and Antonieta Reis Leite. 2024. *Rediscovered Visual Documents of the Portuguese Estado da Índia: The Filippe Nery Xavier Collection*. Coimbra: Universidade de Coimbra, Centro de Estudos Sociais.

Mendiratta, Sidh Losa, Giuseppe Resta, and Tiago Cruz. 2025. "The Fall of Portuguese Jaffna in a Rediscovered Visual Document." *The Ceylon Journal* (3).

Resta, Giuseppe. 2025. "Mapping the Invisible: Problems of Interpretation and Representation of Hormuz Island, Iran." *Athens Journal of Architecture* 11 (2): 201-32. <https://doi.org/10.30958/aja.11-2-5>.

Rossenova, Lozana, Paul Duchesne, and Ina Blümel. 2022. "Wikidata and Wikibase as Complementary Research Data Management Services for Cultural Heritage Data." *CEUR Workshop Proceedings*.

Sabharwal, Arjun. 2015. *Digital Curation in the Digital Humanities: Preserving and Promoting Archival and Special Collections*. Waltham, MA: Chandos Publishing.

Sahgal, Neha, Jonathan Evans, Ariana Monique Salazar, Kelsey Jo Starr, and Manolo Corichi. 2021. *Religion in India: Tolerance and Segregation*. Washington, DC: Pew Research Center.

Thames, Knox, and Sahar Chaudhry. 2014. "Shrinking Religious Freedom in South Asia." *Foreign Policy*, 30 April, 2014. <https://foreignpolicy.com/2014/04/30/shrinking-religious-freedom-in-south-asia/>.

Vogel, Evelyn, and Jürgen Moises. 2021. "Blick nach oben." *Süddeutsche Zeitung*. < https://www.sueddeutsche.de/muenchen/bayern-virtuelle-sehenswuerdigkeiten-dekkengemaelde-1.5184036 >.

von Heyl, Anke, Sonja Thiel, and Zahia Schlott. 2025. "Datenlage gut – Kooperation zwischen NFDI4Culture und dem DigAMus Award." NFDI4Culture. <https://nfdi4culture.de/de/nachrichten/good-data-situation-cooperation-between-nfdi4culture-and-digamus-award.html>.

Waghmare, Abhishek. 2025. "Access to Phones and the Internet." *Data for India*. <https://www.dataforindia.com/comm-tech/>.

Wikipedia contributors. 2025. "Internet Archive." *Wikipedia, The Free Encyclopedia*. <https://en.wikipedia.org/w/index.php?title=Internet_Archive&oldid=1282670714>.

Zhao, Fudie. 2023. "A Systematic Review of Wikidata in Digital Humanities Projects." *Digital Scholarship in the Humanities* 38 (2): 852–74. <https://doi.org/10.1093/llc/fqac083>.

Integrating Projects Working Around an Open Database of Published Music Recordings: a call for collaboration

Toni Sant

Abstract: While Wikipedia's WikiProject Music offers broad encyclopedic coverage, several initiatives within the Wikimedia ecosystem are advancing the structured representation of music metadata through Wikidata, Wikibase, and related platforms. These include AfroSounds, led by Oreoluwa (User:ReoMartins) in Nigeria since 2022; a Slovak music metadata proposal presented at the 2024 CEE Meeting; and the Malta Music Memory Project, developed by the author with the M3P Foundation since 2009. This article calls for collaboration within the digital humanities and data science to develop an interoperable data model and workflow for music metadata, incorporating automation where appropriate and addressing legal constraints that may require interim staging in Wikibase before integration with Wikidata.

Keywords: digital curation, music, databases, collaboration, Wikidata.

1. Introduction

I'm very grateful to the Wikidata and Research conference organizers for allocating me some time and space to share ideas about a topic that holds deep personal and professional significance to me: the role of Wikidata as an open database in bringing together information about published musical recordings.

Music is deeply rooted in the history of humanity, and its preservation is crucial for future generations to understand the diverse cultures and expressions that have shaped our world. However, the systems for cataloging and accessing music data often remain fragmented. Whether it's a jazz composition in Paris or a traditional song from Kinshasa, without systematic preservation infrastructures these musical works and their histories can become lost.

The variety of musical styles and recording histories is vast and complex. This is where open platforms like Wikidata come into play. In this lightning talk I explore some of the ways in which Wikidata and other open databases are transforming the way we archive and share music data, particularly focusing on three examples from across Europe and Africa, and why collaboration through open, global systems can empower more inclusive and accessible global music heritage.

Toni Sant, University of Salford, United Kingdom, A.Sant2@salford.ac.uk, 0000-0001-6393-0502

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Toni Sant, *Integrating Projects Working Around an Open Database of Published Music Recordings: a call for collaboration,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.23, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 185-192, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

2. The importance of open databases for music

To understand the value of open databases like Wikidata for published music recordings, a general overview is useful, even if the primary audience of this conference is familiar with it in various ways. As an open, collaborative, and structured database, Wikidata is very well suited for capturing information about recorded music: tracks, albums, artists, genres, and other related metadata[1]. This information often includes historical, cultural, and technical aspects that can illustrate how a specific element of music has evolved over time.

Wikidata's openness is critical in the context of music archiving. For musicians, researchers, and institutions in Europe, Africa, or anywhere around the world, Wikidata provides a platform for contributing information about music, artists, releases, and historical recordings in a standardized way. Data stored in Wikidata is accessible not only to individuals but also to systems and tools that can process and analyze that data. This interconnectedness creates a framework for building a global catalog of music, which can be as comprehensive as possible.

These are four ways Wikidata enhances published music recordings:

1) Preservation of Cultural Heritage: Many musical recordings are at risk of being lost to time. Wikidata enables notably valuable cultural artifacts, especially from underrepresented regions, to be preserved for future generations.
2) Facilitation of Access and Discovery: Wikidata allows easy access to a wealth of information that might otherwise be difficult to discover. In the context of this presentation, it democratizes information about music by making it accessible to researchers and broader audiences around the world.
3) Fostering Collaboration and Innovation: Wikidata promotes collaboration between researchers, archivists, artists, and developers. By making data available to all, innovation is encouraged. This is especially evident in the creation of new tools that can be used to analyze and explore data about published music in new ways.
4) Supporting the Global Music Industry: For the music industry to thrive on a global scale, there needs to be efficient ways of organizing and distributing music across borders. Open databases play a crucial role in ensuring that artists, labels, and fans can connect regardless of where they are located, and Wikidata is at the forefront of such facilitation.

I am familiar with three examples of engagement with Wikimedia platforms to capture information about published music recordings, which resonate well with these four points:

The AfroSounds Project, led by Oreoluwa Ajayi (User:ReoMartins) from Nigeria since 2022. The project was presented at WikiIndaba 2023 in Morocco by Precious Obiorah (User:Prithee_P), where she explained that the aim of

[1] Wikidata:WikiProject Music provides an excellent overview into all the properties of music that relate to appropriate item structures for recorded releases. <https://www.wikidata.org/wiki/Wikidata:WikiProject_Music>.

their project was to address the huge knowledge gap regarding music in Africa in Wikimedia's open knowledge platforms, with particular attention to Wikipedia[2]. Over sixty Wikimedians contributed to this project between 2022 and 2023 from Nigeria, Ghana, Rwanda, Uganda, Guinea, and Algeria, creating over 250 Wikipedia articles and improving about 1,000 existing articles in English and various African languages. Engagement with Wikidata through this project has been negligible. More importantly, by the end of 2024 this project seemed to have stalled. It's not clear why this has happened. The project's goal remains very pertinent, of course.

Open Music Dataspace, led by Daniel Antal, was presented at the 2024 CEE Meeting in Turkey, where it was described as a music data sharing space with Wikibase starting with music published in Slovakia, inspired by the Luxembourg Shared Authority File project. Antal is not a Wikimedian and has therefore engaged a Wikimedian in Residence to help develop this project appropriately; this is Matej Grochal (User:Jetam2). The project is piloted in Slovakia and its music-focused database held by the Music Centre, and the rights-management organisation, SOZA. As this project is working with GDPR-protected private data, it requires novel technical and organisational solutions than what is customary in Wikimedia's public facing open projects. Although the European Music Observatory has established an independent, stand-alone Wikibase for the Slovak Music Observatory to curate and coordinate its data, future plans include integration with Wikidata, with the aim of providing useful data for Wikipedia[3].

The Malta Music Memory Project, developed by the author with the M3P Foundation since 2009, using a MediaWiki site since 2010 and integrating with Wikidata and the Maltese-language Wikipedia since 2022. This collaborative project involves a series of long-term experiments on how to best document the music sector in the small island nation of Malta[4]. Working primarily with a stand-alone MediaWiki site has shifted after about 15 years, in favour of working directly and openly with Wikidata and Wikipedia, wherever possible, often reinterpreting the concepts of notability and verifiability in this specific context. Although the project has been sustained for almost two decades, this is still a very small-scale dataset. While certainly not insignificant, its greatest strength is potentially in the opportunity it provides to serve as a prototype dataset for

[2] Video of Obiorah's 2023 presentation available at <https://www.youtube.com/watch?v=6E9m6B1wynk>.

[3] Detailsat<https://music.dataobservatory.eu/documentation/background.html#sec-dataspace-definition>.

[4] Some of my published writings about the earlier iterations of this project are available in these papers: "Addressing the need for a collaborative multimedia database of Maltese music." *Journal of Music, Technology & Education* 2, 2-3 (2009): 89-96; "Initial Work on the Malta Music Memory Project – and its connections with Oral History" *Journal of Maltese History*, 12 (2011); "Creating a comprehensive archive of Maltese music on CD." In *Preserving Popular Music Heritage*, ed. by Sarah Baker, 114-125. London: Routledge, 2015.

proof of concept on broader integration across significantly larger music industry players across Europe, Africa, and elsewhere around the world.

3. Related initiatives in Europe

Europe is home to a rich and diverse musical tradition that spans centuries. From classical to rock, electronic to folk, the range of musical styles and genres is very broad. Cataloging this wealth of musical recordings presents challenges, some of which are complicated further by a focus on archiving recordings of music before considering the interoperability of data across catalogs. While European countries have made significant strides in the preservation of their musical history, there is still a considerable gap when it comes to interoperability and shared resources. Several European countries have their own initiatives to preserve and share musical recordings, such as the British Library Sounds Archive in the UK and the Archives of Traditional Music in Hungary. Such efforts have also begun to address the need for comprehensive and open music archives. Although primarily focused on preserving recordings, Europeana Sounds was the foremost initiative in this domain. Europeana is a significantly broader initiative that aggregates digitized cultural heritage materials across Europe, including music recordings, from institutions like libraries, museums, and archives. Europeana Sounds ran from February 2014 to January 2017, and remains housed at the British Library, where it continues to be an aggregator for audio and audio-related material for Europeana. It focuses specifically on music and sound, providing an open-access database of sound recordings, including folk songs, historical sound recordings, and contemporary works[5].

Much of the data in European music documentation initiatives has traditionally been siloed in national or institutional archives, and this is likely to produce fragmented archiving systems and inconsistent metadata. Outside Europeana, music archiving projects tend to operate in isolation, and the absence of a common framework or standardized protocols makes it difficult to integrate them across borders. While there is substantial data being archived by both large institutions and small-scale enthusiasts (most often along with the recordings) this work remains fragmented, limiting access and the overall impact of these initiatives.

Despite the vast array of music archiving projects across Europe, there is no unified database where all this data can converge. This fragmentation limits the ability to search across European music and discover less well-known works. Moreover, when musical works are stored in different archives, they are often tagged with varying metadata. The same artist might be referred to by different names, individual recordings might be categorized differently, and the same release might have multiple catalog numbers, which are not always appropriately classified in terms of relation. This inconsistency makes it difficult to connect recordings across borders.

[5] Information about Europeana Sounds is available at <https://www.eusounds.eu>.

4. The state of music archiving in Africa

In light of the inclusion of the Afrosounds Project in this presentation, the challenges of music archiving in the African continent cannot be overlooked. Africa is home to some of the most diverse and culturally rich musical traditions in the world, yet music archiving is often underfunded and neglected. Many African countries have faced challenges around the preservation of their musical heritage due to historical, economic, and technological barriers.

Some major challenges can be identified when considering how music from Africa can be integrated into the proposed Wikidata approach to published music recordings:

- Lack of Infrastructure: Many African countries have limited infrastructure for archiving and cataloging music, which can make it difficult to establish national or regional databases,
- Limited Access to Music Archives: Many African music archives are not digitized or shared publicly, which hinders access to rich musical traditions, especially from rural or less economically developed regions,
- Loss of Recorded History: During the colonial era and post-colonial times, many musical recordings were lost or destroyed, resulting in gaps in the historical record of African music,
- Underrepresentation of African Music in Global Databases: While some African artists and recordings are represented in Western commercial databases, much African music, including modern and contemporary recordings, remains undocumented or misrepresented,
- Fragmentation of Efforts: Unlike Europe, where centralized projects like Europeana are helping to unify data, many African music archives operate in isolation, with little collaboration between countries or organizations.

While Wikimedia projects may provide a potential solution against this latter point on fragmentation, there have been efforts in recent years to address these challenges, most of which originate outside Africa itself:

- The African Music Archives: AMA is one example of how regional efforts are beginning to converge, coordinated by the University of Mainz in Germany since 1991[6].
- The Pan African Music Project: This project aims to create a centralized database of African music, which includes traditional, contemporary, and diasporic forms of music. Through collaboration with institutions and artists, the initiative is helping to digitize and preserve African musical heritage[7].
- International Library of African Music: Founded in 1954 by Hugh Tracey, I.L.A.M. is the largest repository of (mostly traditional) African music in the world, outside of Africa. A research institution affiliated with the Smithso-

[6] More information available at <https://www.blogs.uni-mainz.de/fb07-ifeas-ama-eng/>.

[7] Official website available at <https://pan-african-music.com/en/>.

nian Institute's Folkways, it is devoted to the study of music and oral arts in Africa, it preserves thousands of historical recordings going back to 1929 and supports contemporary fieldwork[8].

By contributing their data about music in Africa to Wikidata, these initiatives can expand their efforts to preserve local traditions and ensure that African music is included in a wider global conversation about music history. What would actually be required to turn this idea into a viable project?

5. Wikidata's role in bridging gaps

How can we move toward integrating music archives and databases? The answer lies in collaboration through Wikidata. The integration of data from music archives into Wikidata creates a unified platform where researchers, music fans, and artists can access, contribute to, and collaborate on music data.

Wikidata can potentially solve many of the challenges identified here by acting as a centralized, interoperable platform that links together musical works from various cultural heritage institutions, standardizing data, and ensuring that diverse music traditions can be discovered in a single place.

Numerous artists, composers, and musical works have already been added to Wikidata, including metadata about recordings, performances, and historical facts. Some European initiatives, such as the aforementioned Europeana and the British Library's Sound Archive, have started exploring ways to incorporate Wikidata to enhance their metadata. This makes the data easier to search, discover, and connect with other global resources, improving access and interoperability.

Suggestions for key steps that can be taken to address not only the major challenges identified in Africa but also the broader goal proposed by this presentation include:

1) Standardization of Metadata and Protocols: One of the first steps toward integration is pushing for standardized metadata formats and protocols for the cataloging of music recordings. The music industry has long struggled with inconsistent metadata, which hampers the ability to share and connect data. Standardizing the way music is cataloged and tagged can make it easier to integrate databases across different countries and regions through Wikidata. In this context, authority control plays a crucial role, enabling disambiguation between performers, composers, and musical works with similar names, while also consolidating related works by the same individual or ones known by different names[9].

[8] Official website available at <https://ilam.africamediaonline.com>.

[9] Répertoire International des Sources Musicales (RISM) and Répertoire International de Littérature Musicale (RILM) are among the most relevant examples of authority files in the field of music. Their respective websites are available at <https://rism.info> and <https://www.rilm.org>.

2) Collaborative Projects: Initiatives like the African Music Archives and Europeana Sound can serve as models for smaller collaborative projects. By encouraging partnerships between institutions and Wikimedians, we can pool resources and share knowledge and expertise. Jointly curated collections, cross-cultural research, and co-hosted digital platforms can be powerful tools in building more cohesive music archives.
3) Open Access and Licensing: Open access and the use of predefined Creative Commons licenses can ensure that data about music is freely available to anyone who wants to access it. Open data policies continue to be adopted in various regions and this can lay the foundation for open music catalogs that everyone can contribute to and benefit from. It should be noted that only adopting CC0 guarantees the possibility of mass imports of data to Wikidata.
4) Collaborating with Technologists: Leveraging tools and technologies that support Wikidata, such as bots for automated data entry, OpenRefine[10], AI-powered music identification, or APIs that connect Wikidata to music distribution platforms, will help scale the integration of music data, while also addressing issues of sustainability over time. By adopting these types of technologies, processes like metadata generation, cataloging, and copyright protection, can be automated to streamline the process.
5) Engagement with Local Communities: It is essential that any integration efforts be driven by the people whose music is being documented. Along with artists, musicians, and their entrepreneurial associates (such as record labels or distributors) collaboration with local ethnomusicologists and cultural custodians is critical in ensuring that music is represented authentically and respectfully.

6. Call for collaboration

To conclude, the integration of information about published music recordings into Wikidata is not just a technical challenge. It is also a cultural and social opportunity. This represents a unique possibility to create an open, accessible, and dynamic global database of information about music. By leveraging the power of structured, collaborative, and open data, we can ensure that music is preserved, celebrated, and made accessible for future generations, while still respecting the right of musicians and performers to exploit their own intellectual property as they see fit, within existing copyright regimes.

We must move beyond isolated efforts and work toward creating a global network of open music databases. The potential of Wikidata to unify and amplify music archives from around the world is evident as soon as information about published music recordings is separated from recordings themselves. By joining forces, we can create a rich, accessible, and sustainable database that

[10] OpenRefine is an open source tool developed to enable data reconciliation on Wikidata, among other web services and external data. It is available at <https://openrefine.org>.

celebrates the depth and breadth of music from around the world. This can be achieved by harnessing the power of collaboration, standardization, and innovation to empower the music archives of the future. The music is out there, waiting to be discovered, shared, and celebrated, but people need to know about it before they can access it.

References

Sant, Toni. 2009. "Addressing the Need for a Collaborative Multimedia Database of Maltese Music." *Journal of Music, Technology & Education* 2 (2-3): 89–96. <https://doi.org/10.1386/jmte.2.2-3.89_1>.

Sant, Toni. 2011. "Initial Work on the Malta Music Memory Project – and Its Connections with Oral History." *Journal of Maltese History* 2 (2): 42–50.

Sant, Toni. 2015. "Creating a Comprehensive Archive of Maltese Music on CD." In *Preserving Popular Music Heritage*, edited by Sarah Baker, 114–25. London: Routledge. <https://doi.org/10.4324/9781315769882-10>.

SEZIONE 4

Scienze / Science

Opening Up and Linking Type Catalogues in Wikidata – Increasing The Accessibility of Natural History Collections

Sabine von Mering

Abstract: Type specimens are fundamental for taxonomy and irreplaceable components of natural history collections. Type catalogues are scholarly publications listing type material of certain taxonomic groups held in specific collections. The open, collaborative knowledge base Wikidata was used to create a multilingual dataset of published type catalogues, initially focussing on articles related to the *Museum für Naturkunde Berlin*. Wikidata items were enriched with additional metadata, enabling users to filter the dataset by institution, collection, taxon, collector, or expedition. This article aims to improve access to relevant biodiversity literature and to highlight the potential of Wikidata for research and knowledge contextualisation by providing a dynamically growing dataset that facilitates the discovery and reuse of published type catalogues, including for AI training purposes.

Keywords: catalogue of type specimens, Linked Open Data, scholarly publications, type specimen, Wikidata.

1. Introduction

Natural history collections around the world house several billion preserved specimens encompassing the natural world (Johnson et al. 2023). These specimens are used to answer questions about the biodiversity and geodiversity on Earth (Miller et al. 2020 and references within). Collections are increasingly digitised, opened up and made accessible to the wider scientific community and beyond. While access to and quantity of collection information has rapidly increased, there is still a gap related to connecting data. However, this linking of data will help to increase discoverability, transparency and usability – and thus also the visibility – of such collections in natural history museums, herbaria or other scientific institutions. Better data integration will also improve contextualisation, e.g. of historical contexts. Open Science principles and the application of Linked Open Data help to unlock valuable information archived in natural history collections.

Type specimens are the most important objects in natural history collections because they are associated with the names of new taxa, serving as a permanent reference to confirm their identity (see for example Daston 2004). As

Sabine von Mering, Museum für Naturkunde – Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany, sabine.vonmering@mfn.berlin, 0000-0003-2982-7792

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Sabine von Mering, *Opening Up and Linking Type Catalogues in Wikidata – Increasing The Accessibility of Natural History Collections,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.25, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 195-209, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

name-bearing specimens, the type material is regularly examined by scientists as decisive objects for resolving taxonomic issues and clarifying species delimitation. Having easy access to type material is, therefore, an important prerequisite for facilitating research. Specific rules and recommendations on naming organisms regarding type specimens are regulated by the international codes of nomenclature, the International Code of Nomenclature for algae, fungi, and plants (ICNafp; Editorial Committee of the Madrid Code 2025) and the International Code of Zoological Nomenclature (ICZN; ICZN 1999), respectively.

As scientists involved in taxonomic revisions are well aware, locating type material is a difficult and time-consuming activity. Collections may have been transferred to other institutions or dispersed to several places, war-related damages may have led to the (partial) loss of type collections. Type catalogues are compendia for information on the locality of type specimens and their status. Therefore, making these historical and recent publications more accessible helps to improve access to information about type specimens.

Catalogues compiling type material have been published since the 19th century (Kabat and Boss 1992; Uetz et al. 2019) with a large increase in the second half of the 20th century. These publications catalogue taxon names and synonyms, type localities, collectors, they comprise information about the institution(s) housing type specimens including duplicate material, sometimes also measurements and notes. Type catalogues list present and sometimes lost or missing type material of specific collections. Catalogues of type specimens exist for specific (1) taxonomic groups (a family or genus, sometimes a whole order or class of organisms), (2) historical and contemporary institutions or collections, (3) groups of collections or museums, e.g. from a country, (4) collectors, (5) research expeditions, and even (6) scientific journals.

There have been attempts to create a global bibliography of type specimens in zoological and paleontological collections (e.g. Wiktor and Rydzewski 1991). For some disciplines, overview catalogues provide bibliographic compilations of type catalogues, e.g. for malacology (Kabat and Boss 1992) or herpetology (Uetz et al. 2019), respectively. However, to my knowledge there is not yet a global registry of published type catalogues or other scientific literature that uses the advantages of semantic technologies and Linked Open Data (LOD).

The open, multilingual and collaboratively curated knowledge base Wikidata provides structured data that are human and machine readable. Wikidata increases the discoverability, transparency and accessibility of research data and here more specifically, information about natural history collections and the type specimens held by them. Wikidata items are pages compiling referenced statements referring to entities such as scholarly articles, people, places but also scientific institutions and organisations, research expeditions, and much more (von Mering et al. 2025 and references within).

In this study, new Wikidata items are created for scholarly articles of type catalogues published in different languages and academic journals, and existing items are enriched by adding different external identifiers linking to scientific databases. In addition, further entities such as the type specimen

holding institution(s) or collection agents connected to the material or collection are linked. The project started with type catalogues from the *Museum für Naturkunde Berlin,* and was then expanded to compile additional catalogues from around the world.

The objectives of the article are to (1) investigate the potential of Wikidata for research institutions by connecting published type catalogues to different entities such as specific collections, taxa, authors etc. and (2) create a dataset of published type catalogues. The dynamically growing open dataset in Wikidata can be (re)used for research in different fields such as taxonomy and systematics, history of collections, digital humanities or provenance research. Thus, the article aims at highlighting the impact LOD and tools such as Wikidata can have for research and knowledge contextualisation, and eventually to bringing the corpus closer to being machine readable and AI-ready.

2. Material and methods

2.1 Definition and delimitation of type catalogues

Type catalogues or catalogues of type specimens are scholarly publications listing type specimens related to a common topic, most commonly for a taxon, a collection or institution's taxonomic type specimens (or a combination of both). They might be called annotated, critical or illustrated type catalogues if the articles contain annotations or illustrations.

Generally, species catalogues and checklists were not included in this dataset. However, there is an overlap between some of these publication types. For example, annotated checklists may include information on type specimens, sometimes even images of them (Q22681019[1]). Within this transition zone, there are also other articles like taxonomic revisions that could be included if information on type specimens is presented.

2.2 Literature search

The starting point for this project was to get an overview of type material held at the *Museum für Naturkunde Berlin*. The search for published type catalogues focussing on this institution was then expanded to compile additional catalogues from around the world and different fields of natural history, especially zoology, palaeontology and botany. A broad online search was complemented by a focussed search in critical digital libraries such as the Biodiversity Heritage Library (BHL[2]) to find publications providing information on type material. To help find such articles in BHL, the tool BioStor[3] that works as an interface to articles available in BHL, was used (Page 2011). Some collections

[1] <http://www.wikidata.org/entity/Q22681019>.

[2] <https://www.biodiversitylibrary.org/>.

[3] <https://biostor.org/>.

provide lists of type catalogues on their websites. However, these quickly become outdated and are often incomplete. In addition, Wikidata was searched for terms like 'type specimen(s)', 'catalogue/catalog of types', 'primary types of', also exemplary in a few other languages than English.

2.2 Compiling a dataset in Wikidata

A multilingual dataset of type catalogues was created in the open knowledge base Wikidata[4]. Wikidata is community-curated, serves as a hub for external identifiers and provides structured, human and machine-readable data. Wikidata already comprises data for a huge number of publications including type catalogues, many are accessible via the BHL or other digital libraries. However, they are currently not easily searchable as a specific type of scholarly work. Existing Wikidata items were enriched by adding different external identifiers such as Digital Object Identifier (DOI), identifiers of digital libraries like BHL or other reference databases. New Wikidata items were created for scientific articles of type catalogues published in different languages. In addition, further entities such as the type specimen holding institution(s) or collection agents connected to the material or collection are linked.

Throughout this article, names (labels) of Wikidata properties or items are given in *italics*, followed by the property numbers (starting with P) or item numbers (starting with Q) with the respective links to the concept URIs in footnotes. The Wikidata item for a *type catalogue* (Q121763407[5]), was updated and enriched with labels and descriptions in a few languages (English, German, Spanish). However, many other languages lack a label and description for catalogues of type specimens. By adding the language of the scholarly articles using the property *language of work or name* (P407[6]), the study profits from Wikidata's strength as a multilingual tool. For all published type catalogues of the dataset, a statement was added to their Wikidata items, indicating that the *publication type of scholarly work* (P13046[7]) is *type catalogue* (Q121763407[8]). In some cases *catalogue* (Q2352616[9]) was also added when the scope of the article was broader.

2.3 Linking of type catalogues to other entities

The property *main subject* (P921[10]) was used to link to the primary topics of the works. More specifically, for the type catalogues these are the taxa, col-

4 <https://www.wikidata.org>.
5 <http://www.wikidata.org/entity/Q121763407>.
6 <http://www.wikidata.org/entity/P407>.
7 <http://www.wikidata.org/entity/P13046>.
8 <http://www.wikidata.org/entity/Q121763407>.
9 <http://www.wikidata.org/entity/Q2352616>.
10 <http://www.wikidata.org/entity/P921>.

lections, or collectors the articles focus on, in addition also to the terms *type specimen* (Q51255340[11]) and *type catalogue* (Q121763407[12]). By tagging different hierarchical ranks of a taxon, for example not only a certain beetle family but also Coleoptera (beetles) and Insecta (insects), dedicated searching and filtering is enabled.

One focus was also to disambiguate the author or authors of the type catalogues (see Groom et al. 2020 and 2022 for the importance of person identifiers). By using the property *author* (P50[13]) instead of *author name string* (P2093[14]), the scholarly articles are connected directly to the items of the specialists. This requires, however, the existence or creation of a Wikidata item for all of these authors. Whenever possible, external identifiers were added to the Wikidata items of the scholarly articles, including among others the *DOI* (P356[15]), *Handle ID* (P1184[16]), *BHL page ID* (P687[17]) or *BHL part ID* (P6535[18]), *BioStor work ID* (P5315[19]), *JSTOR article ID* (P888[20]), *ResearchGate publication ID* (P5875[21]), and the *ZOBODAT publication ID* (P10931[22]) for the database ZOBODAT providing mostly German-language works in the field of zoology and botany. Whenever available, links to PDFs or other openly accessible versions of the publications were added via the property *work available at URL* (P953[23]).

Type catalogues are often published as part of a series. To make these parts easier to find, they were linked via the property *follows* (P155[24]) and its inverse property *followed by* (P156[25]). In addition, the parts can be connected via the property *part of the series* (P179[26]) and linked to a separate item for the series.

The website Scholia[27] was used to visualise some results of the study. By using bibliographic and other information from Wikidata, this service creates visual scholarly profiles for topics[28] and other entities.

[11] <http://www.wikidata.org/entity/Q51255340>.
[12] <http://www.wikidata.org/entity/Q121763407>.
[13] <http://www.wikidata.org/entity/P50>.
[14] <http://www.wikidata.org/entity/P2093>.
[15] <http://www.wikidata.org/entity/P356>.
[16] <http://www.wikidata.org/entity/P1184>.
[17] <http://www.wikidata.org/entity/P687>.
[18] <http://www.wikidata.org/entity/P6535>.
[19] <http://www.wikidata.org/entity/P5315>.
[20] <http://www.wikidata.org/entity/P888>.
[21] <http://www.wikidata.org/entity/P5875>.
[22] <http://www.wikidata.org/entity/P10931>.
[23] <http://www.wikidata.org/entity/P953>.
[24] <http://www.wikidata.org/entity/P155>.
[25] <http://www.wikidata.org/entity/P156>.
[26] <http://www.wikidata.org/entity/P179>.
[27] <https://scholia.toolforge.org/>.
[28] <https://scholia.toolforge.org/topic/>.

3. Results

3.1 The dataset in Wikidata

In total, a dataset of 561 published type catalogues (as of May 4, 2025) was created to highlight the potential of using Wikidata. A static snapshot of the dataset was deposited on Zenodo (von Mering 2025), while a dynamically growing dataset of these type catalogues can be queried in Wikidata using the following SPARQL query[29]. Within the dataset of this study, only five type catalogues were published in the 19th century, 26 stem from the first half of the 20th century. A strong increase is visible from the 1950s, with 277 published type catalogues enriched in Wikidata so far. In the range between the years 2000 and 2025, up to now, 253 catalogues of type specimens were compiled. Of the 561 articles, unsurprisingly the majority was written in English (370, ca. 66%), followed by German (125, ca. 22%) and French (53, 9.4%). Only a few more examples were added for other languages, five articles in Italian, three in Portuguese and Spanish each plus one Chinese and one Russian type catalogue.

The dataset comprises publications dealing with type specimens from a wide range of taxonomic groups, from zoology, palaeontology and botany. The tagging with some higher ranks of the studied taxa (via the property *main subject*) allows searching for type catalogues related to a certain taxon. For example, SPARQL queries for Mollusca[30] and Amphibia[31] retrieved 46 and 29 articles, respectively. A Wikidata knowledge graph is visualising 400 different topics (mostly taxa and institutions) connected to the type catalogue articles via the Wikidata property *main subject* (Fig. 1).

In botany, catalogues seem to be less common than in zoology. However, there are notable works dealing with the life and collections (including type material) of influential botanists such as Carl Linnaeus[32] (Jarvis 2007) and Robert Brown[33] (Mabberley and Moore 2022). Other botanical articles focus mostly on certain taxa (often families) in a specific herbarium (Q100709098[34], Q118145794[35]) or again on the herbarium of a specific botanist (Q106884222[36]).

[29] <https://w.wiki/RJZB>.
[30] <https://w.wiki/RJZR>.
[31] <https://w.wiki/RJZd>.
[32] <http://www.wikidata.org/entity/Q1043>.
[33] <http://www.wikidata.org/entity/Q155764>.
[34] <http://www.wikidata.org/entity/Q100709098>.
[35] <http://www.wikidata.org/entity/Q118145794>.
[36] <http://www.wikidata.org/entity/Q106884222>.

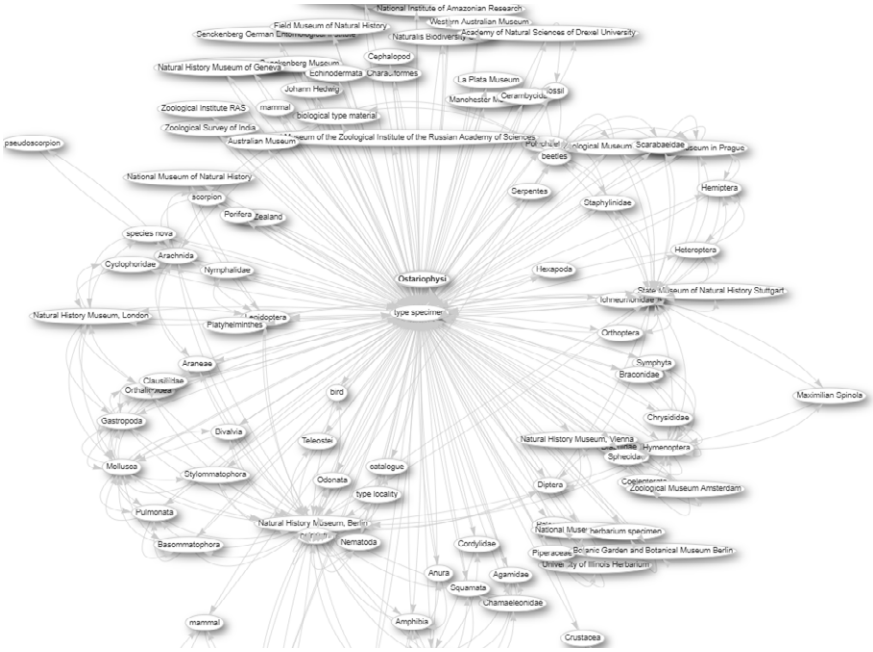


Figure 1 – Screenshot of a Wikidata knowledge graph visualising 400 different topics (mostly taxa and institutions) connected to type catalogue articles via the Wikidata property *main subject* (Wikidata query[37]).

Most "classical" type catalogues compile type material housed in specific – historical and contemporary – collections or institutions. This may include type catalogues of museums now closed (Q109624285[38]) and those listing type specimens formerly held in collections destroyed during wartime (Neumann 2006). These publications also report about exchange of duplicate specimens and acquisition from other museums, thus showing connections between collections. Revised and annotated catalogues offer updates and critical evaluations. A map displays those collections and institutions featured by type catalogues in the dataset that have coordinates associated in Wikidata (Fig. 2).

37 <https://w.wiki/FtJc>.

38 <http://www.wikidata.org/entity/Q109624285>.

Figure 2 – Map displaying collections and institutions featured by type catalogues in the dataset (only those with coordinates associated in Wikidata are shown as red dots), created with Scholia.

A large subset of this dataset are articles featuring type specimens held at the *Museum für Naturkunde Berlin*. The query[39] retrieves about 150 published type catalogues related to this institution. Sometimes, type catalogues also cover the type material housed in a group of collections, for example in several Australian museums. There are a number of type catalogue series focussing mostly on a taxon and the relevant type material in a specific collection. For two exemplary series, all parts were linked via *follows* and *followed by* as well as connected as part of a series. While the series "Die Typen und Typoide der Mollusken-Sammlung des Zoologischen Museums in Berlin" (Q134312479[40]; Fig. 3) comprises twelve parts about type specimens in the mollusc collection of the *Museum für Naturkunde Berlin* (all authored or co-authored by Rudolf Kilias[41]), seven articles belong to the series "Kritischer Katalog der Typen der Fischsammlung des Zoologischen Museums Berlin" (Q134313544[42]) and provide information about the type material housed in the fish collection of the same museum. Several other series are part of the dataset; however, not all are complete, and not all parts are connected to one another.

[39] <https://w.wiki/Esyf>.
[40] <http://www.wikidata.org/entity/Q134312479>.
[41] <http://www.wikidata.org/entity/Q59532054>.
[42] <http://www.wikidata.org/entity/Q134313544>.

Figure 3 – Knowledge graph for a series of published type catalogues, in this case, a series comprising twelve articles on type specimens held in the mollusc collection of the *Museum für Naturkunde Berlin*. Visualisation created with Scholia[43].

The dataset comprises several examples of type catalogues covering type specimens collected by (Q124215534[44], Q86998973[45]) or described by certain scientists (Q131422372[46], Q23069147[47]). Similarly, such catalogues also deal with type specimens gathered during research expeditions (Q134006487[48], Q97499425[49]). One of the rare examples of a female collector being highlighted by a publication is Aimée Fournier de Horrack[50], French entomologist and collector of butterflies. In the title and article she is mentioned under her husband's name, as Madame G. [Gaston] Fournier de Horrack (Q108825313[51]). Her outstanding collection of Lepidoptera is housed at the *Muséum National d'Histoire Naturelle* in Paris. Finally, some scientific journals provide overview articles on typifications of names and type material of those described within the journal. For example, a catalogue by Caballer Gutiérrez et al. (2011) compiled all type material of new taxa described in the journal *Avicennia. Revista de Biodiversidad Tropical* in the first 15 years of its existence. Since 2008,

43 <https://scholia.toolforge.org/topic/Q134312479#context>.
44 <http://www.wikidata.org/entity/Q124215534>.
45 <http://www.wikidata.org/entity/Q86998973>.
46 <https://www.wikidata.org/wiki/Q131422372>.
47 <http://www.wikidata.org/entity/Q23069147>.
48 <http://www.wikidata.org/entity/Q134006487>.
49 <http://www.wikidata.org/entity/Q97499425>.
50 <http://www.wikidata.org/entity/Q4697203>.
51 <http://www.wikidata.org/entity/Q108825313>.

the journal *Willdenowia* has published an index to typifications of names at the end of each issue of the journal (examples of the indices can be found by searching the BioOne[52] platform for “index to typifications” and filtering for Willdenowia as journal title).

The authors of the published type catalogues were disambiguated, the original author name strings resolved and linked to the authors’ Wikidata items whenever these items existed already. Fig. 4 visualises 200 connected authors, scored according to the number of type catalogues published by them.

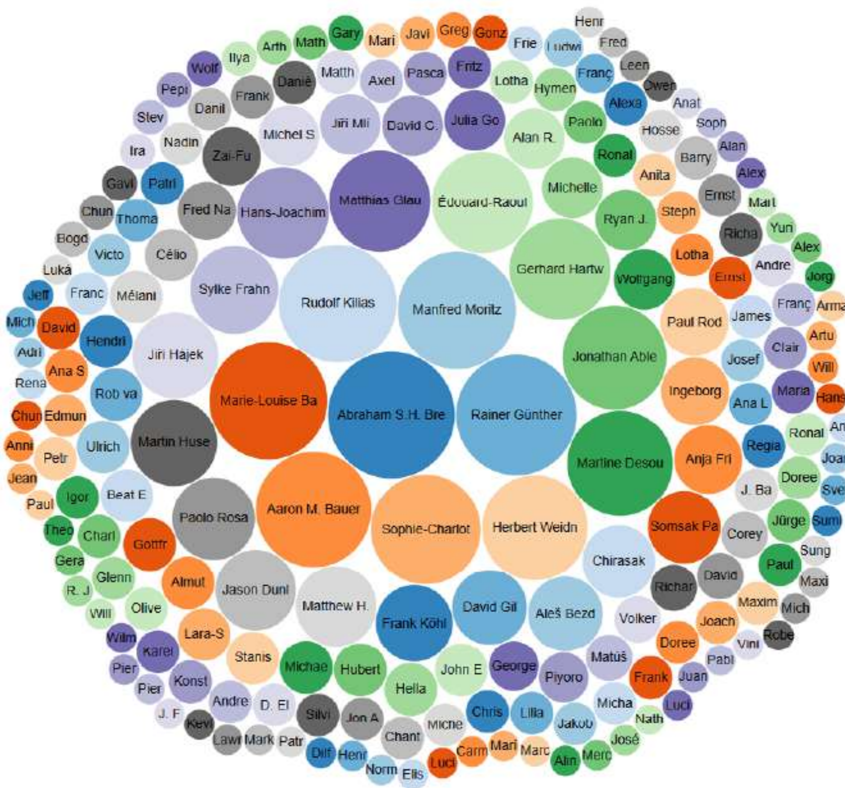


Figure 4 – Authors of published type catalogues (200) scored according to the number of publications (Wikidata query[53]), larger circles correlate to higher numbers of authored articles in the dataset.

[52] <https://bioone.org/search?term=%E2%80%9Cindex+to+typifications%E2%80%9D>.

[53] <https://w.wiki/FtJi>.

4. Discussion and outlook

As a result of the original question and the search for catalogues related to the *Museum für Naturkunde Berlin*, the dataset is limited in its coverage and clearly biased towards this and other German institutions and German-language catalogues as well as relatively easy findable English type catalogues. The project was not designed with a systematic approach to compile all catalogues for a certain field or beyond. For this publication, a cut off was necessary at some point in order to have a dataset for analysis and submission before the *Wikidata and Research 2025* conference in Florence. Therefore, it is currently not a representative dataset, not all taxonomic groups are covered equally and there are many underrepresented regions, groups and languages.

Partly, this is due to the high concentration of type material in large collections in the Global North, rooted in the colonial history museum collections (Johnson et al. 2023). Through improving accessibility to institutional resources, collections, archives, literature and infrastructure, also by supporting local collections and scientists in underrepresented regions, some of these barriers to knowledge-sharing can be removed.

This compilation was created within a limited timeframe, it is an ongoing study and more type catalogues will be added. However, the pilot dataset can be dynamically expanded to include other type catalogues focussing on additional taxa, collections, countries, collectors etc., hopefully also resulting in the addition of further publications in other languages.

For several taxonomic groups, comprehensive websites catalogue data on accepted names, synonyms, type specimens and more. The *World Spider Catalog* (Q3570011[54]), published and operated by the Natural History Museum of Bern, is such a taxonomic database that also collects all arachnological literature (World Spider Catalog 2025). While this taxonomic database is supported by institutional and community commitment, other similar websites often depend on one or a few active specialists sharing their expert knowledge (Pulawski 2025). Such digital catalogues also rely on access to relevant taxonomic literature including published type catalogues and other articles featuring type specimens.

In addition to the printed book (Jarvis 2007), data from the Linnaean plant name typification project are also available as a dataset, accessible via different aggregators and annually updated (Wajer 2022). However, for other taxonomic groups such as the earwigs (Dermaptera), the main sources of information on the species and type specimens are only available in printed volumes (Catalogus Dermapterorum; *Dermapterorum Catalogus Praeliminaris*[55]). Only part of the published type catalogues are openly accessible to everyone. Therefore, a more complete global list of such catalogues could help to inform and prioritise digitisation of missing literature that is not in copyright anymore, for example in the context of BHL.

[54] <http://www.wikidata.org/entity/Q3570011>.

[55] <https://www.aucklandmuseum.com/collection/object/1214642>.

While there are exemplary initiatives for certain taxonomic groups, there does not seem to be a global register for such information as type catalogues or type specimens. A general issue and challenge is the long-term hosting and sustainability of such digital catalogues and databases. In addition to dedicated data curators, there is a need for institutional commitments and the support of a broader community. Even if a global approach is split into larger taxonomic expert networks or communities, harmonisation, integration and sharing of data is critical. Well-curated high-quality datasets related to natural history collections require international collaboration as well as the use of community-agreed standards and persistent identifiers. This includes exchange with different organisations promoting standardisation and open access to biodiversity data such as the Global Biodiversity Information Facility[56] (GBIF), the Consortium of European Taxonomic Facilities[57] (CETAF) and Biodiversity Information Standards[58] (TDWG). For example, a TDWG Task Group[59] is developing a terminology on how to model research expeditions in Wikidata (von Mering et al. 2024), and the BHL-Wiki Working Group[60] is involved in data modelling. Such collaborations – also with the wider Wiki community – help to improve the modelling of type catalogues and other entities in Wikidata and to develop best practice recommendations. GBIF's Global Registry of Scientific Collections[61] (GRSciColl[62]) is a community-curated repository of information about scientific collections, that stores information about their content, location, contacts, associated institutions, and collection codes and identifiers. Relevant literature documenting specimens held in these collections, particularly type catalogues, may represent a valuable source of collection information for GRSciColl.

After linking the published type catalogues to the collections, a next step could be to connect further to the so-called material citations (Darwin Core Term dwc:MaterialCitation[63]) and/or to the specific type specimens in GBIF. Plazi unlocks the potential of biodiversity literature by extracting and liberating data, which are subsequently published through TreatmentBank[64] and the Biodiversity Literature Repository[65] (Agosti et al. 2020; Agosti and Miller 2022).

Given widespread challenges of limited resources, Wikidata could serve as a global, collaborative, and community-curated knowledge base and a sustainable

[56] <http://www.wikidata.org/entity/Q1531570>.
[57] <http://www.wikidata.org/entity/Q5163385>.
[58] <http://www.wikidata.org/entity/Q4914768>.
[59] <https://www.tdwg.org/community/cd/expeditions/>.
[60] <https://meta.wikimedia.org/wiki/Biodiversity_Heritage_Library/Our_community>.
[61] <http://www.wikidata.org/entity/Q98594614>.
[62] <https://scientific-collections.gbif.org/>.
[63] <https://dwc.tdwg.org/list/>.
[64] <https://plazi.org/treatmentbank/>.
[65] <https://zenodo.org/communities/biosyslit/>.

infrastructure for cataloguing biodiversity-related literature (see also Page 2022; 2023). Wikidata could assist in finding publications in different languages, be a hub for identifiers, and as such a connector of different entities. The collection of type catalogues curated in Wikidata could be used as a training dataset for text and data mining. Applying machine learning techniques could bring the information in these publications closer to AI-ready and make linked information actionable. Wikidata could provide a platform to collaboratively curate a dynamically growing corpus of historical and contemporary type catalogues from all disciplines. The global research community is increasingly using Wikidata, but its full potential will only be realised through broader adoption, the dismantling of data silos, and stronger international and interdisciplinary collaboration.

5. Acknowledgements

Wikimedia Italia is thanked for a travel scholarship supporting my in person participation in Florence. This project was supported by a number of colleagues from the *Museum für Naturkunde Berlin*, especially Peter Bartsch, Sylke Frahnert, Jason Dunlop, Kristina von Rintelen, Thomas von Rintelen, Ralf Thomas Schmitt, Frank Tillack. Rebecca Lippert (MfN Berlin) is thanked for her help with compiling the dataset. For discussions on type catalogues as well as information and/or literature I am grateful to Donat Agosti (Plazi), Mathias Dillen (Meise), Anne Greenwood MacKinney (Berlin), Andreas Kroh (Vienna), Sandra Knapp (London), Roderic Page (Glasgow) and Nicholas Turland (Berlin). Special thanks to Michael Ohl (MfN Berlin) for the inspiring exchange and discussions on catalogues, history of science and also for access to literature.

References

Agosti, Donat, and Jeremy Miller. 2022. "The TreatmentBank Factory: Data Access." Naturalis Science Seminar, Leiden. Zenodo. <https://doi.org/10.5281/zenodo.7262609>.

Agosti, Donat, Marcus Guidoti, and Guido Sautter. 2020. "Liberating the Richness of Facts Implicit in Taxonomic Publication: The Plazi Workflow." *Biodiversity Information Science and Standards* 4: e59179. <https://doi.org/10.3897/biss.4.59179>.

Caballer Gutiérrez, Manuel, Jesús Ángel Ortea Rato, José Espinosa Sáez, Leopoldo Moro Abad, and Juan José Bacallado Jr. 2011. "Catálogo del material tipo de los nuevos taxones descritos en Avicennia, Revista de Biodiversidad Tropical (1993–2008)." *Revista de la Academia Canaria de Ciencias* 23 (3): 59–92.

Daston, Lorraine. 2004. "Type Specimens and Scientific Memory." *Critical Inquiry* 31 (1): 153–82. <https://doi.org/10.1086/427306>.

Editorial Committee of the Madrid Code [Turland, Nicholas J., John H. Wiersema, Fred R. Barrie, Kanchi N. Gandhi, Julia Gravendyck, Werner Greuter, David L. Hawksworth, Patrick S. Herendeen, Ronell R. Klopper, Sandra Knapp, Wolf-Henning Kusber, De-Zhu Li, Tom W. May, Anna M. Monro, Jefferson Prado, Michelle J. Price, Gideon F. Smith, and Juan Carlos Zamora Señoret]. 2025. *International Code of Nomenclature for*

Algae, Fungi, and Plants (Madrid Code). Regnum Vegetabile 162. Chicago: University of Chicago Press. <https://doi.org/10.7208/chicago/9780226839479.001.0001>.

Groom, Quentin, Anton Güntsch, Pieter Huybrechts, Nicole Kearney, Siobhan Leachman, Nicky Nicolson, Roderic D. M. Page, David P. Shorthouse, Anne E. Thessen, and Elspeth Haston. 2020. "People Are Essential to Linking Biodiversity Data." *Database* 2020: baaa072. <https://doi.org/10.1093/database/baaa072>.

Groom, Quentin, Christian Bräuchler, Robert W. N. Cubey, Mathias Dillen, Pieter Huybrechts, Nicole Kearney, Niels Klazenga, Siobhan Leachman, Deborah L. Paul, Heather Rogers, Joaquim Santos, David Peter Shorthouse, Alison Vaughan, Sabine von Mering, and Elspeth M. Haston. 2022. "The Disambiguation of People Names in Biological Collections." *Biodiversity Data Journal* 10: e86089. <https://doi.org/10.3897/BDJ.10.e86089>.

ICZN. 1999. *International Code of Zoological Nomenclature*. Fourth edition. London: The International Trust for Zoological Nomenclature.

Jarvis, Charlie. 2007. *Order out of Chaos: Linnaean Plant Names and Their Types*. London: The Linnean Society of London in association with the Natural History Museum.

Johnson, Kirk R., Ian F. P. Owens, and the Global Collection Group. 2023. "A Global Approach for Natural History Museum Collections: Integration of the World's Natural History Collections Can Provide a Resource for Decision-Makers." *Science* 379 (6638): 1192–94. <https://doi.org/10.1126/science.adf6434>.

Kabat, Alan R., and Kenneth J. Boss. 1992. "An Indexed Catalogue of Publications on Molluscan Type Specimens." *Occasional Papers on Mollusks* 5 (69): 157–336. <https://biostor.org/reference/232832>.

Mabberley, David J., and David T. Moore, with the assistance of Jacek Wajer. 2022. *The Robert Brown Handbook: A Guide to the Life and Work of Robert Brown 1773–1858. Scottish Botanist*. Regnum Vegetabile 160. Oberreifenberg: Koeltz Botanical Books.

Miller, Sara E., Lisa N. Barrow, Sean M. Ehlman, Jessica A. Goodheart, Stephen E. Greiman, Holly L. Lutz, Tracy M. Misiewicz, Stephanie M. Smith, Milton Tan, Christopher J. Thawley, Joseph A. Cook, and Jessica E. Light. 2020. "Building Natural History Collections for the Twenty-First Century and Beyond." *BioScience* 70 (8): 674–87. <https://doi.org/10.1093/biosci/biaa069>.

Neumann, Dirk. 2006. "Type Catalogue of the Ichthyological Collection of the Zoologische Staatssammlung München. Part I: Historic Type Material from the 'Old Collection', Destroyed in the Night 24/25 April 1944." *Spixiana* 29: 259–85.

Page, Roderic D. M. 2011. "Extracting Scientific Articles from a Large Digital Archive: BioStor and the Biodiversity Heritage Library." *BMC Bioinformatics* 12: 187. <https://doi.org/10.1186/1471-2105-12-187>.

Page, Roderic D. M. 2022. "Wikidata and the Bibliography of Life." *PeerJ* 10: e13712. <https://doi.org/10.7717/peerj.13712>.

Page, Roderic. 2023. "Ten Years and a Million Links: Building a Global Taxonomic Library Connecting Persistent Identifiers for Names, Publications and People." *Biodiversity Data Journal* 11: e107914. <https://doi.org/10.3897/BDJ.11.e107914>.

Pulawski, Wojciech J. 2025. *Catalog of Sphecidae*. California Academy of Sciences. <https://www.calacademy.org/scientists/projects/catalog-of-sphecidae>.

Uetz, Peter, Sami Cherikh, Glenn Shea, Ivan Ineich, Patrick D. Campbell, Igor V. Doronin, José Rosado, Addison Wynn, Kenneth A. Tighe, Roy Mcdiarmid, Justin L. Lee, Gunther Köhler, Ryan Ellis, Paul Doughty, Christopher J. Raxworthy, Lauren Scheinberg, Alan Resetar, Mark Sabaj, Greg Schneider, Michael Franzen, Frank Glaw, Wolfgang Böhme, Silke Schweiger, Richard Gemel, Patrick Couper,

Andrew Amey, Esther Dondorp, Gali Ofer, Shai Meiri, and Van Wallach. 2019. "A Global Catalog of Primary Reptile Type Specimens." *Zootaxa* 4695 (5): 438–50. <https://doi.org/10.11646/zootaxa.4695.5.2>.

von Mering, Sabine, Robert W. N. Cubey, Dag Endresen, Annika Hendriksen, Siobhan Leachman, Daniel Mietchen, and Joaquim Santos. 2024. "Advancing Community Curation of Research Expeditions: A Collaborative Journey with Wikidata and Biodiversity Information Standards." *Biodiversity Information Science and Standards* 8: e138921. <https://doi.org/10.3897/biss.8.138921>.

von Mering, Sabine, Siobhan Leachman, Joaquim Santos, and Heidi M. Meudt. In press. "Wikidata for Botanists: Benefits of Collaborating and Sharing Linked Open Data." *Annals of Botany*.

von Mering, Sabine. 2025. *Published Type Catalogues / Catalogues of Type Specimens* (1.0) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.15338503>.

Wajer, Jacek. 2022. *The Linnaean Plant Name Typification Project* [Data set]. London: Natural History Museum. <https://doi.org/10.5519/qwv6u7j5>.

Wiktor, Jadwiga, and Władysław Rydzewski. 1991. *Bibliography of Catalogues of Type Specimens in [the] World's Zoological and Palaeozoological Collections*. Serie Uniwersytetu Wroclawskiego, Wydzial Nauk Przyrodniczych, Prace Zoologiczne 22. Wroclaw: Wroclaw University Press.

World Spider Catalog. 2025. *World Spider Catalog*. Version 26. Bern: Natural History Museum Bern. <http://wsc.nmbe.ch>. <https://doi.org/10.24436/2>.

Enhancement and Sustainable Fruition of Cultural Heritage in the Era of Ecological Transition: A Case Study on Open Data and Citizen Science

Alessia Minnella, Francesca Tornari, Stefano Pierpaolo Marcello Trasatti, Iolanda Pensa, Dario Crespi, Marianna Cappellina

Abstract: The doctoral research project focuses on cultural heritage in the Italian context, investigating the correlation between cultural heritage preservation and the ecological transition, emphasizing open knowledge, sustainability and citizen science. Through an interdisciplinary approach the project aims at developing an in-depth case study focused on a museum system and its repository. The study aims at analysing poly-material artefacts through multidisciplinary methodologies in a sustainable context, reaching the goal of the research by means of monitoring biotic and abiotic environmental parameters. Wikimedia infrastructures and communities play a pivotal role in ensuring the accessibility and dissemination of research findings through Open Access and free licensing practices.

Keywords: Cultural heritage, environmental sustainability, citizen science, open access.

1. Introduction

The protection of cultural heritage—and the active involvement of both institutions and citizens in this process—lies at the core of national and international regulations and policy frameworks. Article 9 of the Italian Constitution states that "The Republic promotes the development of culture and scientific and technical research. It protects the landscape and the historical and artistic heritage of the Nation" (Senato della Repubblica, Costituzione Italiana. Testo vigente 2023). According to it, the Cultural Heritage and Landscape Code sets out the essential concepts relating to activities aimed at Italy's Cultural Heritage, i.e. enhancement, conservation and protection (D.lgs. 42/2004). In 1972, UNESCO adopted the Convention Concerning the Protection of the World Cultural and Natural Heritage of Humanity, encouraging international cooperation for

Alessia Minnella, University of Milan La Statale, Italy, alessia.minnella@unimi.it, 0009-0003-2267-6706
Francesca Tornari, University of Milan La Statale, Italy, francesca.tornari@unimi.it, 0009-0007-5563-3277
Stefano Pierpaolo Marcello Trasatti, University of Milan La Statale, Italy, stefano.trasatti@unimi.it, 0000-0002-5456-4376
Iolanda Pensa, University of Applied Sciences and Arts of Southern Switzerland, Switzerland, iolanda.pensa@supsi.ch, 0000-0001-5122-7385
Dario Crespi, Wikimedia Italia, Italy, dario.crespi@wikimedia.it, 0000-0001-5003-4442
Marianna Cappellina, National Museum of Science and Technology Leonardo da Vinci, Italy, cappellina@museoscienza.it

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Alessia Minnella, Francesca Tornari, Stefano Pierpaolo Marcello Trasatti, Iolanda Pensa, Dario Crespi, Marianna Cappellina, *Enhancement and Sustainable Fruition of Cultural Heritage in the Era of Ecological Transition: A Case Study on Open Data and Citizen Science,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.26, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 211-219, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

its conservation and identifying the sites most at risk (UNESCO 1972). In addition, Faro Convention, or Council of Europe Framework Convention on the Value of Cultural Heritage for Society, signed in 2005 by more than 20 States, establishes rights and responsibilities to and for Cultural Heritage, focusing on the promotion of sustainability, accessibility, and democratic participation (Faro Convention Art. 12 2005). To date, the same theme is present in the 2030 Agenda for Sustainable Development signed in 2015 by 193 UN member States. The Agenda defines the '17 Sustainable Development Goals' to be pursued in order to achieve sustainable development, in which, in section 11.4, the need to strengthen efforts to protect and safeguard the world's cultural and natural heritage is emphasised (ONU, Agenda 2030 2015). Climate change poses a significant threat to cultural heritage, which, as a testimony to civilisation, history and culture, must be constantly protected. Rising temperatures, flooding and air pollution are just some of the effects of the phenomenon, the direct consequences of which can be seen in the cultural environment in the archaeological sites, monuments and cities of art (Fabbri and Bonora 2021).

In this frame museums assume a pivotal role in cultural heritage conservation, since, according to ICOM, they represent institutions that researches, collects, preserves, interprets and exhibits cultural heritage (ICOM 2024). Museums' mission to ensure long-term preservation and continued access to cultural heritage is endangered in the ever-changing scenario posed by climate change (Cassar and Pender 2005). Collections enhancement, stated as a central activity for cultural heritage preservations (D.lgs. 42/2004), is becoming an issue in museums, often resulting in accumulation of artefacts stored in repositories, regarded as an underused resource. In fact, most museums' collections are stored in these environments: in a survey proposed by ICOM to museums all over the world it is stated that the majority of interviewed institutions exhibit less than 15% of their collection to the public, and the necessity to increase accessibility to those objects is a common opinion. The demand of bigger spaces collides with the reality of most institutions: the same survey highlights that storage spaces condition is judged to be rather unfavourable by most museums staff, regarding lack of free areas and equipment, as most types of furniture are considered to be in short supply (ICOM 2024).

In this context of conscious use, the significance attributed to repositories is increasing in museums system all over the world, not only for their safety functions, but also as study and research reserves. This change of direction is also reflected by the ever-growing success of "open storage", which allow visitors to access depot's spaces, usually through guided tour, thus increasing cultural heritage awareness and valorisation (Ferruzza and Marras 2020). This and other innovative phenomena aim to increase cultural heritage accessibility, but from preventive conservation point of view, this reality rises new doubts, concerning the material aspect of collection, different environmental conditions and logistic organization. This situation presents a high-risk factor, with potential impact on deterioration dynamics: the joint necessity is to find balance between new demands of accessibility and conservation of artefacts in repositories (Loddo 2018).

Museums embrace several categories of artefacts in their collections, with disparate genesis and consistencies, differentiated based on collections' topic and period. This heterogeneity represents both the richness and complexity of cultural heritage, as it portrays the multi-faceted nature of human identity, while also posing challenges to conservation procedures (Canadelli and Di Lieto 2024). Composite artefacts are present in many museums' collections, as modern synthetic materials are often combined with traditional ones, with heterogeneous productive techniques. Poly-material is one of the only common factors among the heterogeneity of collections, and presents a difficulty for their conservation, as interactions of different material result in complex management choices for conservators, from preventive measures to restorative protocols. Degradation phenomena of these objects are strongly conditioned by complex stratification of different materials and their chemical nature (Zmeu and Bosch-Roig 2022). Usually, the presence of organic materials enhances object's susceptibility to biodeterioration, as these substances represent a potential source of nutrition for heterotrophic organisms (Caneva, et al. 2008). Conservative solution to poly-material artefacts deterioration is not straightforward, as established treatments could be the remedy for the baseline material but cause alteration of the other one (Zmeu and Bosch-Roig 2022). If existing procedures can't be fully and satisfyingly adapted to technical-scientific artefacts conservation, defining new protocols is essential.

The study is framed within the wider context of the ecological transition, addressing the emerging challenges and responsibilities that contemporary conservation practices must navigate in response to dynamic environmental conditions. This research aims to enrich the current academic discussion on the protection of cultural heritage by adopting an analytical perspective focused on the evaluation of biotic and abiotic environmental parameters, alongside the assessment of the preservation state of cultural heritage materials.

The project is part of the PNRR Cultural Heritage initiative Ministerial Decree no. 118 of 02-03-2023 (D.M. 118/2023). The research aims at supporting the Italian cultural institutions and heritage sites in the current phases of dynamic change promoted by the ecological transition. The purpose of the project is going to be achieved through the implementation of a systematic experimentation and the use of tools that, starting from the academic research, will stimulate interaction with civil society (Bertacchini and Pensa 2023), as the third mission activities request. The entire research is based on two fundamental pillars: the accessibility of data, through the application of Open Access, Open Data and Open licences; and the active participation of communities, enabling citizens to become an integral part of scientific research, following the principle of Citizen science. This research investigates the role of open data and citizen science in the protection of cultural heritage. It explores how the integration of open data platforms (Farda-Sarbas et al. 2019) and the active involvement of citizens in scientific research contribute to the preservation and sustainable management of heritage resources.

The implementation of sustainable conservation practices and the adoption of innovative technologies for monitoring and damage prevention in cultural

heritage management are therefore of fundamental importance in the context of the ecological transition of the cultural sector (Marzano, et al. 2024), where museums are increasingly required to make strategic choices—deciding in which areas to invest resources and in which to reduce consumption—according to sustainability principles and environmental priorities. In addition, by combining cultural heritage preservation, open knowledge dissemination, and community involvement, the project highlights the transformative potential of Wikimedia-driven initiatives and tools (Fagerving 2023) in addressing global challenges like climate change, while reinforcing the relevance of cultural heritage in society's ecological transition.

2. Case study

The project aims at developing an in-depth case study which is satisfied and reflected through an interdisciplinary approach.

This study takes the Museo Nazionale Scienza e Tecnologia Leonardo da Vinci (Italy) as its focal point, analysing it through a dual analytical lens: one oriented toward conservation practices, and the other toward fostering active participation via community-driven and Wikimedia-related initiatives. In order to study the correlation between climate change and cultural heritage, the investigation underlying the research project is aimed at: i) analysing the museum practices in the field of sustainability; ii) analysing the effects of climate change contextualised to works exhibited in the indoor environment and iii) defining strategies aimed at heritage protection. The entire project is therefore focused on the accessibility and sharing of scientific data related to cultural heritage, with a focus on scientific reproducibility and involving the active participation by communities.

Museo Nazionale Scienza e Tecnologia Leonardo da Vinci is examined through its distinct cultural heritage contexts, focusing on the repository and the museum system. More in detail, the investigation on the repository is focused on the identification and definition of optimal environmental conditions through the design of a monitoring campaign aimed at examining the biotic and abiotic parameters and the artefacts' surfaces conditions; while the focus on the museum system implies the analysis of the issue related to the sustainability practices applied to the repository environment. These two aspects represent key typologies of heritage topic and offer an opportunity for interdisciplinary analysis of their respective conservation and interpretive frameworks.

2.1 Museum repository

The indoor environment analysed has been identified as a museum repository, that covers a large area and that is characterized by a wide variety of poly-materic objects exhibited, dating back to different historical periods. The environment identified is "*Collezioni di studio*", that is the Museo Nazionale Scienza e Tecnologia Leonardo da Vinci's indoor repository (Museo Nazionale Scienza e Tecno-

logia Leonardo da Vinci, 2021). In addition, the attention towards this cultural institution is due to the fact that in February 2022 it has released data from 20.000 objects of the collections available in LOD form (Linked Open Data) under free CC BY-SA 4.0 Licenses, as part of the national ArCo – Architettura della Conoscenza project developed by the ICCD (ArCo project, 2022). The sheets produced show information at the inventory and at the catalographic level, following ICCD's normative; the 20.000 objects are divided into: artworks and objects (OA); scientific and technological heritage (PST); musical instruments (SM). "Collezioni di studio" is partially open to visitors; it occupies an area of 2.000 m^2, hosting about 8.000 objects consisting of various materials referring to different historical periods. The items are stored into three rooms, named: Motorbike and Red Tent Area, Scientific instruments Area and Ship model Area; all the objects are conserved inside lockers and shelves, or on the ground with supports to prevent direct contact with the floor and ensure stability.

The collaboration has been shaped to follow three principal research lines: environmental micro-climatic and biological monitoring of the exhibition areas, analysing of the state of preservation of poly-materic objects exhibited in relation to the exposure environment (with a specific focus to the metallic ones), and Wikimedia events aimed at updating Wikipedia articles. The cultural heritage data acquired, related to the repository environment monitoring and the state of preservation of the objects under study, will be stored in the Wikibase structure, in order to make research data easily accessible. Wikibase allows to create private istances that can serve as primary sources for data to be later referenced in Wikidata (Rossenova, et al. 2022). The activities planned as "wikiexcursions" and "edit-a-thons" are focused on updating Wikidata and Wikimedia Commons through specific themes focused on cultural heritage and preventive conservation. Within the edit-a-thons, there are also activities aimed at updating specific points or interesting areas of OpenStreetMap (Grinberger, et al. 2022), through the correct geolocation of the big indoor cultural heritage objects under study, such as those on display in the naval and aviation pavilion. The environmental micro-climatic monitoring, thanks to the emerging collaboration with a company that provided sensors, is structured through the placement of 8 pairs of sensors along the perimeter and the centre of the museum repository. The choice related to the location of the sensors within the repository follows a strategy that implies considering the total area of the repository as a "grid", that allows to reconstruct the microclimate of a given limited area. Sensors acquire and record data that can be monitored over time to assess environmental quality. They register both atmospheric parameters (T, RH and Pa) and pollution parameters (PM 1, PM 2.5, PM 4, PM 10, VOCs and CO_2). In addition, the Museum has a microclimate control system that allows monitoring data related to indoor data acquired from 2021 to date in the "Collezioni di Studio" environment, from 3 sensors recording T and RH. Given the possibility to analyse and compare the atmospheric parameters from the previous years to date, an environmental risk model is feasible. A one-year monitoring campaign is now active in order to define the Heritage Microclimate Risk (Historical) and the

Predicted Risk Damage Indices (HMR_{hs} e PRD index), that indicate the microclimate risks and the possibility to reduce material damage (Tringa, et al. 2024).

Alongside biotic environmental component has been preliminary assessed through aerobiological investigation, to assess concentration and abundance of microorganisms able to deteriorate artefacts conserved in museum storage room. Airborne contaminants have been sampled using the method of culture-plate, employing Petri dishes as gravitational samplers, and identified with morphological analysis. This preliminary inspection was necessary to determine the exact protocol to employ during aerobiological monitoring campaign, including sampling location, number and timing. Seasonal monitoring will be conducted with active samplers, employing both culture-dependent and culture-independent techniques for identification, in order to correlate microorganisms' concentration to abiotic environmental factors.

The heterogeneity and poly-materiality of artefacts stored within the "Collezioni di Studio" repository requires the application of multidisciplinary investigation methodologies. For this reason, further biological investigations will be carried out on surfaces of artefacts composed by different materials, which present potential active biodeterioration. Microorganisms will be sampled with sterile cotton swab and both culture-dependent and culture-independent techniques will be employed for identification. Moreover, the investigation of biodeterioration dynamics on artifacts will be conducted *in vitro*, recreating the same materials and conditions of some representative artefacts in laboratory, to assess the interaction between microorganisms and materials, since cultural heritage value seldom admit invasive samplings.

A large number of the items located in "Collezioni di studio" are techno-science-related and include two-wheeled vehicles made of metals and alloys, so the need to set up a diagnostic campaign also focused on metals. For the purpose of evaluating the variation in the state of preservation over time, control samples (carbon steel coupons 2,5x5cm) have been placed in the exposure areas of the museum repository, along the perimeter and the inner exhibition area. The metallic coupons were placed in the Scientific Instrument Area at 11 points of interest and in the Motorbike and Red Tent Area at 8 points of interest. The metallic coupons were placed at both 45° and vertical angles, at two different heights corresponding to those of the shelves with the goods on display. In spring 2025 82 carbon steel coupons have been placed in the selected points of interest in "Collezioni di Studio". Through the acquisition of photos and subsequent periodic visits within two years of research, control samples are analysed with the aim of studying the degradation products, organic deposits, and material loss assessment.

2.2 Museum system

The focus on the museum system reflects the need to thoroughly investigate the internal dynamics of a cultural institution that can gradually lead to the definition of "Green Museum", a cultural entity which fully embodies the princi-

ple of environmental sustainability in the field of Cultural Heritage. A "Green Museum" integrates that principle into its management, exhibition practices and operational activities, with the aim of reducing ecological impact and promoting environmental awareness among the public. This includes energy efficiency strategies, waste reduction, use of sustainable materials and educational programmes focused on sustainability (Brophy, et al., 2008).

The study of the Museo Nazionale Scienza e Tecnologia Leonardo da Vinci is carried out through a multi-dimensional approach, encompassing the examination of sustainability-related guidelines and best practices for cultural buildings, the analysis of HVAC systems (Heating, Ventilation and Air Conditioning system) within the indoor storage facilities, and the investigation of Citizen Science initiatives promoting sustainability in the museum sector. Currently, the museum repository, is also undergoing a rearrangement of part of the collections and the definition of the new indoor heating system. The possibility of analysing the consumptions related to the current HVAC system in the Scientific Instruments Area of "Collezioni di studio" over the course of a one-year campaign, allows to create an actual Life Cycle Assessment model (LCA), a method for assessing the environmental impact, determining its actual potential and the feasibility of energy-saving. Starting from the LCA Inventory of data in the input and output of the system, the model is based on the ISO 14041 standard (ISO 14041) and allows to define a correct strategy, in the face of low energy consumptions (Ge, et al. 2015).

The involvement of Wikimedia communities constitutes a key component of the case study, particularly through the organization of participatory initiatives such as edit-a-thons and writing weeks, aimed at enriching and updating Wikipedia content related to the museum and its themes.

3. Conclusion and future work

This research has the ultimate purpose to develop sustainable strategies and models for the use and wide dissemination of cultural heritage, starting from the current state of the art, intertwining with the common awareness of the fundamental issue of climate change and progressively producing and monitoring data. The project overview emphasises how museums with their already available data and those in the process of being openly released, can be mapped, enriched and interconnected through Wikidata. This approach enables the selection of new relevant data and the updating of an open and accessible information network, fostering the valorisation of cultural heritage. A key component of the project deals with organizing Wikimedia events in order to involve citizens and communities in the current research process. These events aim to raise awareness, share best practices in sustainable heritage conservation, and democratize access to scientific knowledge. Wikipedia, Wikidata, Wikimedia Commons, Wikibase and OpenStreetMap constitute the pillars to assess the entire research project.

The in-depth case study has the purpose to investigate both the environmental sustainability practices applied to museum environments and the main phe-

nomena of heritage degradation, through the analysis of the causes that led to their occurrence. The aim is to monitor and define both the conservation environment of cultural heritage and the degradation phenomena discovered on the surfaces of the materials under analysis. Particularly, due to the poli-matericity of the cultural heritage artefacts exhibit within the "Collezioni di Studio" repository, has been necessary to apply a multidisciplinary strategy. In addition, as a result of the monitoring of all the environmental data acquired and the state of conservation of the works, it is possible to obtain the definition of guidelines and best-practices aimed at the preventive conservation of the historical heritage under study.

This research project offers a contribution to the understanding of methods and practices for the conservation of cultural heritage, aware that three years of experimentation is only the beginning of a process to be carried out on a much longer time scale, in order to assess the impact of climate change on cultural heritage.

References

Albore, Alexandre Saverio, Giulio Malatesta, and Christelle Molinié. 2021. "Open Cultural Data and MediaWiki Software for a Museum: The Use Case of Musée Saint-Raymond (Toulouse, France)." *Environmental Sciences Proceedings* 10 (1): 10. <https://doi.org/10.3390/environsciproc2021010010>.

ArCo project – Architettura della Conoscenza. 2022. "ArCo project – Rete di Ontologie per il Patrimonio Culturale." <https://dati.beniculturali.it/arco-rete-ontologie#dati>.

Bertacchini, Enrico, and Iolanda Pensa. 2023. "Exploring Collaborative Digital Heritage Communities: A Quantitative Assessment of Wiki Loves Monuments in Italy." *Il capitale culturale* 28: 129–50. <https://doi.org/10.13138/2039-2362/3197>.

Brophy, Sarah S., and Elizabeth Wylie. 2008. *The Green Museum: A Primer on Environmental Practice*. Lanham, MD: AltaMira Press.

Canadelli, Elena, and Paola Barbara Di Lieto. 2024. *Da cimeli a beni culturali: fonti per una storia del patrimonio scientifico italiano*. Milano: Editrice Bibliografica.

Caneva, Giovanna, Maria Pia Nugari, and Orazio Salvadori. 2008. *Plant Biology for Cultural Heritage*. Vol. I. Los Angeles: Getty Conservation Institute.

Cassar, May, and Robert Pender. 2005. "The Impact of Climate Change on Cultural Heritage: Evidence and Response." In *ICOM Committee for Conservation: 14th Triennial Meeting, The Hague*. London: James & James. <https://discovery.ucl.ac.uk/id/eprint/5059/1/5059.pdf>.

Council of Europe. 2005. *Council of Europe Framework Convention on the Value of Cultural Heritage for Society*. Faro: Council of Europe.

Decreto Legislativo 22 gennaio 2004, n. 42. *Codice dei beni culturali e del paesaggio, ai sensi dell'articolo 10 della legge 6 luglio 2002, n. 137.*

Decreto Ministeriale n. 118 del 2 marzo 2023. <https://www.mur.gov.it/it/atti-e-normativa/decreto-ministeriale-n-118-del-02-03-2023>.

Fabbri, Kristina, and Alessandro Bonora. 2021. "Two New Indices for Preventive Conservation of the Cultural Heritage: Predicted Risk of Damage and Heritage Microclimate Risk." *Journal of Cultural Heritage* 48: 112–22. <https://doi.org/10.1016/j.culher.2020.09.006>.

Fagerving, Alicia. 2023. "Wikidata for Authority Control: Sharing Museum Knowledge with the World." *DHNB Publications, DHNB2023 Conference Proceedings*. <https://journals.uio.no/dhnbpub/issue/view/875>.

Farda-Sarbas, Mariam, and Claudia Müller-Birn. 2019. "Wikidata from a Research Perspective: A Systematic Mapping Study of Wikidata." arXiv. <https://arxiv.org/abs/1908.11153>.

Ferruzza, Maria Luisa, and Anna Maria Marras. 2020. "I depositi come spazio di ricerca e valorizzazione dei patrimoni museali." In *Beni Culturali: dai depositi alla valorizzazione. Modi, forme, esperienze, norme*. Caltanissetta: Edizioni Lussografica. <https://iris.unito.it/retrieve/e27ce433-4193-2581-e053-d805fe0acbaa/1%20-%20Ferruzza%20e%20Marras%20estr.pdf>.

Ge, Jing, Xia Luo, Jun Hub, and Shuyan Chen. 2015. "Life Cycle Energy Analysis of Museum Buildings: A Case Study of Museums in Hangzhou." http://dx.doi.org/10.1016/j.enbuild.2015.09.015>.

Grinberger, A. Yair, Marco Minghini, Levente Juhász, Godwin Yeboah, and Peter Mooney. 2022. "OSM Science—The Academic Study of the OpenStreetMap Project, Data, Contributors, Community, and Applications." *ISPRS International Journal of Geo-Information* 11 (4): 230. <https://doi.org/10.3390/ijgi11040230>.

ICOM. 2024. *Museum Storage around the World: Working Group on Collections in Storage*. Paris: ICOM. <https://icom.museum/wp-content/uploads/2024/05/Report_ICOM-STORAGE_EN_Final-ok.pdf>.

ISO. 1998. *ISO 14041: Environmental Management — Life Cycle Assessment — Goal and Scope Definition and Inventory Analysis*. Ginevra: International Organization for Standardization.

Loddo, Marta. 2018. "Museum Storage Facilities: What's Next?" *ICAR-International Journal of Young Conservators and Restorers of Works of Art* 2: 144–59.

Marzano, Marianna, and Monia Castellini. 2024. "Cultural Heritage and Sustainability: What Is State of the Art? A Systematic Literature Review." *Sinergie Italian Journal of Management* 42 (1): 65–84.

Museo Nazionale della Scienza e della Tecnologia "Leonardo da Vinci". 2021. "Collezioni di Studio." <https://www.museoscienza.org/it/area-stampa/Collezioni-di-studio>.

Museo Nazionale della Scienza e della Tecnologia "Leonardo da Vinci". 2022. *Catalogo dei dataset del Museo della Scienza e della Tecnologia "Leonardo da Vinci"*. <https://dati.museoscienza.org/lod/>.

ONU. 2015. *Agenda 2030 per lo Sviluppo Sostenibile*.

Rossenova, Lozana, Paul Duchesne, and Ina Blümel. 2022. "The 3rd Wikidata Workshop, Workshop for the Scientific Wikidata Community." *@ ISWC 2022*.

Senato della Repubblica. 2023. "Costituzione Italiana, Testo Vigente", Art. 9. Aggiornamento al 2023.

Tringa, Evangelia, Dimitrios Kavroudakis, and Klea Tolika. 2024. "Microclimate-Monitoring: Examining the Indoor Environment of Greek Museums and Historical Buildings in the Face of Climate Change." *Heritage* 7: 1400-18. <https://doi.org/10.3390/heritage7030067>.

UNESCO. 1972. *Convenzione sulla Protezione del Patrimonio Mondiale culturale e naturale.*

Zmeu, Carmen Nicoleta, and Paula Bosch-Roig. 2022. "Risk Analysis of Biodeterioration in Contemporary Art Collections: The Poly-Material Challenge." *Journal of Cultural Heritage* 58: 33–48. <https://doi.org/10.1016/j.culher.2022.09.014>.

PARTE SECONDA

Interrogare i dati / Data Retrieval

Visualizing Wikidata: The RAWGraphs 2.0 Approach

Michele Mauri, Tommaso Elli

Abstract: Wikidata represents an incredible revolution in open knowledge curation, yet its complexity and structure often limit its visibility and use, particularly when it comes to data visualization. RAWGraphs 2.0 addresses this challenge by providing an intuitive, template-based approach for translating Wikidata's SPARQL query outputs into meaningful visualizations. Drawing on data visualization literature, RAWGraphs employs visualization templates: conceptual and technical schemas that identify visual properties of graphical items and bind them to data. With direct SPARQL integration, RAWGraphs supports transparent, reproducible visual exploration, highlighting data quality, biases, and structural insights within Wikidata.

Keywords: Information design, Wikidata, visualization tool, SPARQL queries, chart templates

1. Introduction

Wikidata is a collaborative knowledge base with significant potential for defining relationships between concepts and their descriptors. Its architecture assigns a unique identifier (known as a Q-number) to each concept. For instance, the "Universe" is labeled as Q1, "Earth" as Q2, and a "Human Being" as Q5. Items in Wikidata are interconnected through properties; for example, "Earth" (Q2) is linked to "planet" (Q634) via the property "instance of" (P31). Each property also has its unique identifier, enabling precise querying across items sharing similar properties. Furthermore, items can include descriptors represented as strings or numbers, such as the population size of cities.

The resulting structure of Wikidata forms a highly complex network. At the time of writing, it comprises over 116 million items and approximately 12,000 properties, continuously expanding in size and complexity ("Wikidata Homepage" 2025).

Despite its potential, exploring and understanding Wikidata remains challenging for many users. The data frequently present issues such as missing values, inconsistent modeling practices, ambiguous hierarchical structures, and language inconsistencies. Moreover, the querying process itself can be complicated. Queries in Wikidata are typically performed using SPARQL (a recursive acronym for SPARQL Protocol and RDF Query Language), a standard devel-

Michele Mauri, Polytechnic University of Milan, Italy, michele.mauri@polimi.it, 0000-0003-1189-9624
Tommaso Elli, Polytechnic University of Milan, Italy, tommaso.elli@polimi.it, 0000-0002-9818-1991

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Michele Mauri, Tommaso Elli, *Visualizing Wikidata: The RAWGraphs 2.0 Approach,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.28, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 223-238, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

oped by the World Wide Web Consortium (W3C). SPARQL is specifically designed to retrieve and manipulate data stored in RDF (Resource Description Framework) formats, such as Wikidata's structure. Its syntax and logic entail a steep learning curve, though recent developments in AI have begun to lower this barrier (Mecharnia and d'Aquin 2025).

SPARQL queries yield various results, but these are most commonly displayed in tabular form. In these tables, each column represents a property or variable, and each row represents a so-called "binding", that is a specific result that matches the query pattern.

Visualization methods enable users to identify patterns, gaps, and relationships that may otherwise remain hidden within complex datasets. As Bounegru et al. (2018) argue, visual representations are not merely aesthetic outputs; they serve as analytical lenses, prompting us to question, validate, and interpret data critically. For Wikidata specifically, visualizations can highlight inconsistencies (such as duplicated population values), expose biases (like the underrepresentation of female biographies), and illustrate structural characteristics (including taxonomies and temporal spans).

This paper explores the potential of the open-source tool "RAWGraphs" for visualizing Wikidata query results. RAWGraphs utilizes a template-based approach, wherein each chart type clearly defines its visual variables (e.g., x-position, size, color) that users can connect directly to data columns. The latest release, version 2.0, introduces direct SPARQL support, allowing users to input queries directly, connect seamlessly to data endpoints like Wikidata, and immediately visualize the results. This feature aims to bridge the gap between Wikidata query results and their meaningful interpretation.

2. Background: why visualize Wikidata?

Wikidata is one of the most recent developments in the extensive ecosystem surrounding Wikipedia (Vrandečić and Krötzsch 2014). Over the past decade, hundreds of millions of items have been added, either manually by users or automatically through bots programmed to extract data from existing Wikipedia pages or other open sources.

This immense collection primarily serves as "central storage for the structured data of its Wikimedia sister projects, including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others" ("Wikidata Homepage" 2025). Its primary use is therefore highly specific: providing up-to-date, language-independent information about various topics. For example, consider the infoboxes used in articles about cities, which include details such as area, population, and current government. Wikidata simplifies the process of maintaining consistent and current information across different language editions of Wikipedia. This targeted use encourages detailed and specific definitions for the properties associated with each item.

However, this method of data production can lead to inconsistencies and biases, especially given Wikidata's vast scale. Such biases were highlighted in a

study examining race and citizenship biases within the knowledge base (Shaik et al. 2021). Visualization techniques can effectively reveal these biases, providing a clearer overview of complex issues. As authors of the mentioned study on biases on Wikidata state: "Data visualizations about gender biases can help bring awareness to the gender bias in Wikidata." (Shaik et al. 2021).

Over time, interest has grown in accessing Wikidata not just in individual, isolated instances, but also through aggregated visual approaches. Several tools have emerged to facilitate specific visualizations. A notable example is Wikidata Atlas (Del Pino and Hogan 2023), which aims to provide cartographic representations of Wikidata items retrieved from queries. The authors identify several key motivations for visualizing aggregated Wikidata: enabling users to discover new knowledge serendipitously, identifying issues related to data quality, and highlighting geographic biases. Another relevant contribution is VizKG (Raissya et al. 2021), a Python framework designed for visualizing Wikidata items. Broadening the discussion to general visualization practices for linked open data sources, in making a literature review on the role of visualization in the exploration of RDF data sources, authors argue that "visual data exploration is particularly useful when little is known about the source data and the analysis goals are vague" (Gómez-Romero et al. 2018). This observation is highly applicable to Wikidata, particularly when considering the extension of its contents.

3. Visualization and Wikidata

Visualization in the context of Wikidata queries has been addressed in the literature through three main approaches: visualization as a query interface, visualization for quality assessment, and visualization as a means of communicating Wikidata contents.

The first approach involves visual interfaces designed specifically to facilitate querying the knowledge base (Camarda et al. 2012; Metilli et al. 2019). In this context, visualization serves directly as a querying tool, enabling users to retrieve information without writing complex SPARQL queries. These visual query interfaces leverage the capabilities of information visualization to simplify data exploration and interaction.

A second strand employs visualization as a method to explore and navigate topics within Wikidata. An example is Scholia (Nielsen et al. 2017), a tool developed for exploring academic-related data, such as researcher profiles and their publications. Scholia demonstrates how visualization effectively organizes complex information, making it easier to navigate and understand. Another example is InteractOA (Elhossary and Förstner 2024), a web platform tailored for biologists, designed to visualize interactions of small RNAs in bacteria. Numerous similar tools are listed on a dedicated Wikidata page ("Wikidata:Tools/Visualize Data" 2025). Reviewing this list reveals that most tools have very specific visualization types, such as tree diagrams for family trees or timelines for historical events.

The third approach involves visualization as a tool for critically assessing data quality (Das et al. 2023; Shaik et al. 2021; Zhang and Terveen 2021). Here,

visualization techniques are employed to identify and expose systematic biases or representation issues within the Wikibase. In this context, visualization is often incorporated into scientific publications to illustrate researchers' findings clearly and effectively communicate them to readers.

Across these three approaches, visualization is generally treated as a way to support access to, exploration of, or critique of Wikidata. Yet they also point to a broader issue: visualizations are often perceived as final outcomes, viewed as the endpoint of data analysis or processing pipelines. However, visualization outputs are often not endpoints but milestones within broader analytical or communicative processes, particularly when analytical goals are vague or exploratory (Mauri and Ciuccarelli 2016). Visualizations function as intermediate checkpoints, prompting further investigation, refining research questions, or highlighting data gaps that warrant additional analysis.

In open and exploratory analyses, visualizations are particularly valuable, as they support data quality assessment, facilitate the identification of significant patterns or biases, and highlight areas requiring deeper investigation. Therefore, visualization is not merely a means of displaying final results; it is an integral component of iterative cycles of inquiry, interpretation, and discovery within Wikidata's complex and continuously evolving dataset.

For these reasons, we propose and discuss an approach aimed at exploiting the potential of a flexible, open, and transparent visualization ecosystem.

Given Wikidata's nature as a massive knowledge base originally designed for highly specific use cases yet suitable for broad, exploratory queries aimed at uncovering new questions, evaluating completeness, robustness, and biases, we believe the approach underlying the tool RAWGraphs is particularly applicable.

4. The RAWGraphs approach

RAWGraphs[1] was initially developed to simplify the creation of data visualizations (Mauri et al. 2017). Since its creation, the tool has emphasized openness, not just in licensing, but also in its design process. The development of RAWGraphs is grounded in the principle that visualizations are components of broader analytical and communicative processes, rather than standalone outcomes. This openness was driven primarily by user needs, especially among visual designers, researchers, and journalists, who required tools that can be integrated with their existing practices and workflows.

To make the creation of visualizations accessible even to non-technical users, RAWGraphs employs a straightforward and consistent data model based on a single flat table. Whether data is loaded from a spreadsheet, pasted as CSV, or retrieved through SPARQL, users interact with a familiar tabular format: rows represent records and columns represent dimensions. This decision significantly reduces cognitive load and barriers to entry, removing the necessity

[1] <https://app.rawgraphs.io>.

of understanding complex data structures such as nested JSON, RDF graphs, or multi-table joins.

Simplifying chart creation further, RAWGraphs separates chart elements into two categories: data-related elements and chart-related elements. Data-related elements include, for instance, dots in a scatterplot, their positions, colors, and sizes. Chart-related elements cover aspects such as font styles used on axes, axes boundaries, dot contours, label font families, and artboard dimensions. This separation follows Jacques Bertin's theory of visual variables (Bertin 1967), which conceptualizes graphical representation as encoding data through defined perceptual channels.

Central to this methodology is the linkage between data dimensions (columns of the table) and the visual variables offered by each chart type. This modular framework enables the creation of "chart templates," which function as predefined visualization recipes, clearly distinguishing data-related from chart-related variables. For example, a scatterplot template may define visual variables such as X-position, Y-position, dot size, and dot color, while chart-related variables may include artboard size, maximum diameter, and dot stroke style. Clearly, categorizing these elements depends on the template developer's intentions and needs.

This modular approach transcends specific technologies, making it broadly applicable across diverse contexts. Following a successful crowdfunding campaign, RAWGraphs released a significant update in 2021, enhancing its technical infrastructure and introducing new features.

Focusing specifically on SPARQL query functionalities, RAWGraphs now allows users to load data directly from SPARQL endpoints. It employs an intuitive interface featuring two input fields: one for specifying the SPARQL endpoint[2] and another for entering the SPARQL query itself. Once loaded, data is visualized as easily as if imported from a traditional tabular file.

Another innovation is the platform's modular architecture, comprising three core repositories: the core system, the user interface (app), and chart templates. The core supports chart template creation and execution, the app provides a user-friendly interface, and the "chart templates" repository contains a curated collection of reusable charts. Templates can be developed independently, simplifying creation and maintenance. Additionally, new chart templates can now be dynamically loaded through the web interface, requiring no specialized technical skills. A notable example is the population pyramid template, developed for specific use cases and shared openly by its creators, visible in figure 8.

5. Case studies: visualizing SPARQL data with RAWGraphs

Even before the 2.0 update, we envisioned strong potential in connecting Wikidata with RAWGraphs. A first exploration using it has been described in a blog post (RAWGraphs Team 2017). In this section, we highlight two common groups of

[2] <https://query.wikidata.org/sparql>.

visual problems (representing hierarchies and values over time) and then show how it is possible to use a custom chart to visualize specific needs.

The tool offers the ability to load data from various sources: a CSV or JSON file, online data sources, or a SPARQL datasource. When the SPARQL option is selected, the interface provides two fields (Fig. 1): one for the "SPARQL Endpoint," which defaults to Wikidata's endpoint, and another for writing the SPARQL query. While the query field includes syntax highlighting and basic error detection, it does not support advanced features for query composition. Therefore, it is often better to first write and test the query using the dedicated Wikidata interface and then paste it into the tool.

5.1 Hierarchies

A common visualization challenge involves understanding categorical structures. Specifically, identifying items within categories and counting items within subcategories. The representation of categories and their nesting has a very long history (Lima 2014, 15). RAWGraphs approaches this by viewing each line of a dataset as a leaf and each column as a hierarchical level.

In RAWGraphs, there are several solutions for representing hierarchies (Fig. 2), derived from both practical applications and data visualization literature. Each template provides distinct affordances: for example, tree-like dendrograms emphasize structural relationships, while treemaps emphasize quantities.

A simple scenario might involve visualizing a list of items alongside a numeric property. For example, identifying the largest cities globally, representing their populations comparatively, and grouping them by country. While seemingly trivial, visualization helps highlight potential data issues. By querying Wikidata for cities (instances of Q515), their population (P1082), and their country (P17), we can construct a hierarchical dataset[3].

When mapping data dimensions to visual variables (Fig. 3), RAWGraphs by default counts the number of records grouped by the same hierarchy. This reveals that some cities have multiple population entries (Fig. 4); for example, Brasília has six records, and Kano has two. This is due to Wikidata's ability to accommodate data from different sources or time points. This approach helps assess data quality and highlights the importance of clearly defined visualization strategies, such as selecting minimum, maximum, or average population values.

A more complex hierarchical visualization could explore uneven categorizations, such as music genres (Fig. 5). Querying Wikidata for instances of bands and their genres, then building hierarchies based on the "subclass of" property of the genre, can reveal disparities in genre classification[4]. This visualization provides insights into the complexity and coverage of Wikidata's genre categorization. From the visualization it is possible to see that the most

[3] Query: <https://w.wiki/Dqsi>.

[4] Query: <https://w.wiki/EtdV>.

Figure 1 – The first step of the RAWGraphs interface, where data can be loaded. In this screenshot, the "SPARQL query" option is selected, and syntax highlighting is visible in the inserted query.

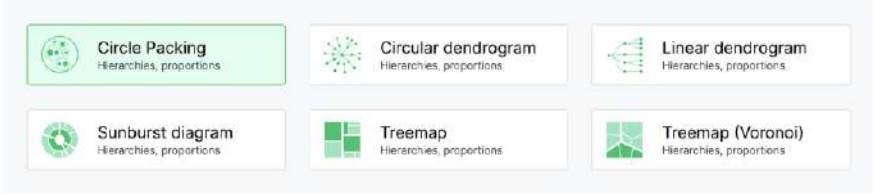


Figure 2 – The available chart templates for visualizing hierarchies in RAWGraphs.

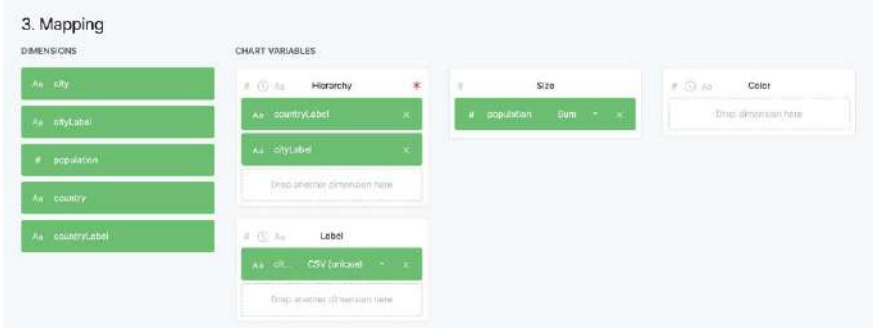


Figure 3 – Mapping of dimensions for the creation of a circle packing in RAWGraphs interface. See results in figure 4.

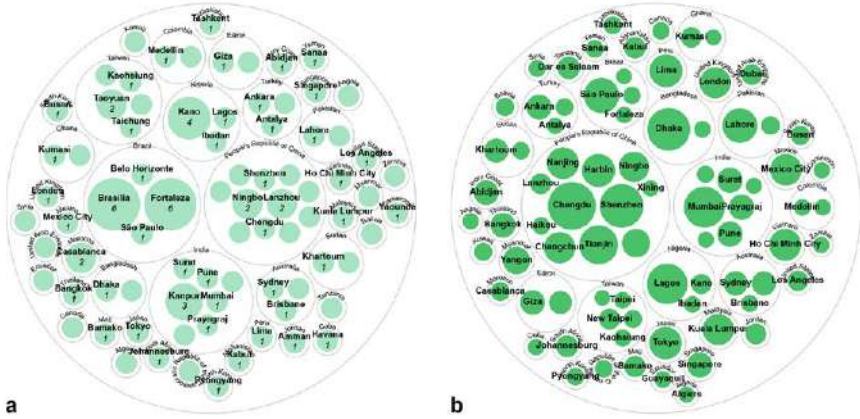


Figure 4 – a) Visualization of the 100 most populated cities according to Wikidata, using RAWGraphs. On the left, no value is assigned to size, so the tool maps it by default to the number of records found. This reveals that some cities, such as Brasília, have up to six population values. b) On the right, the population is mapped to size, allowing us to see the countries where the cities are located and visually compare them. Based on the previous exploration, we can choose to use the minimum, maximum, or average value when multiple numeric entries are present. Access the Wikidata query at: https://w.wiki/Dqsi.

common genre is rock music, which according to Wikidata is a subclass of "popular music". At the same time, "rock music" acts as a parent class for other genres. This lets us hypothesise that rather than a tree structure we are dealing with a rhizomatic and probably recursive structure.

5.2 Temporal series

Time representation is another common challenge in data visualization. A widely used approach is to represent time along the horizontal axis, a standardized convention especially in Western countries, where time typically progresses from left to right. Despite this common strategy, a variety of alternative solutions exist, as reflected in the range of options available in RAWGraphs (Fig. 6).

To illustrate an example of a time series, we used a query that retrieves all asteroids present in Wikidata along with their dates of discovery (query: https://w.wiki/DqxV)[5]. Since there are tens of thousands of entries, we grouped them by "asteroid families," which are groups that share similar proper orbital elements,

[5] Query: <https://w.wiki/DqxV>.

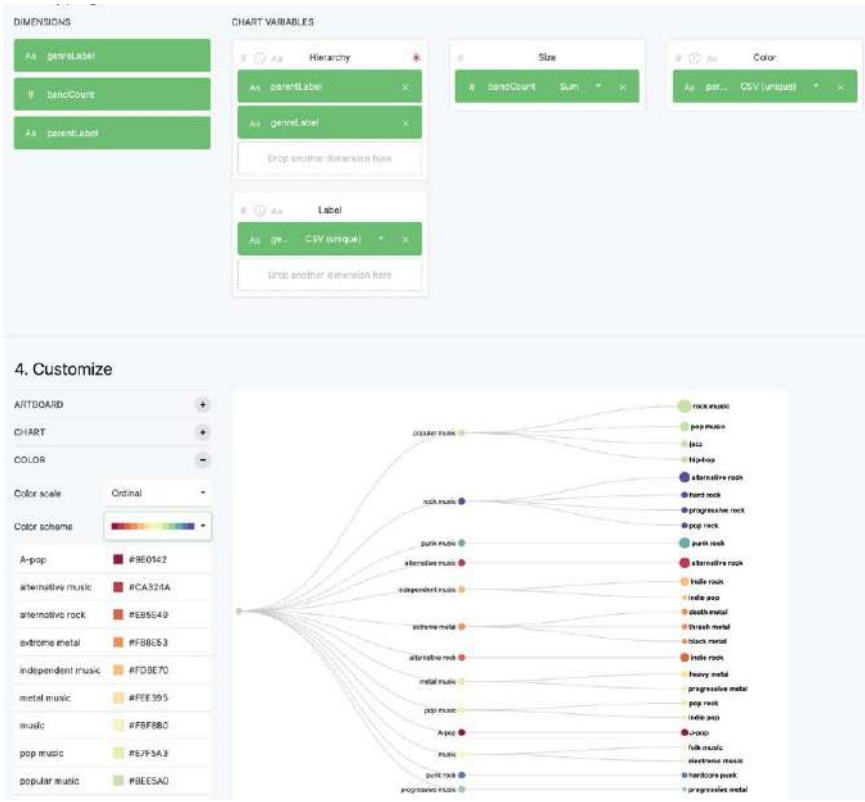


Figure 5 – Mapping and visualization of musical genres by parent genre in RAWGraphs. Access the Wikidata query at: https://w.wiki/EtdV.

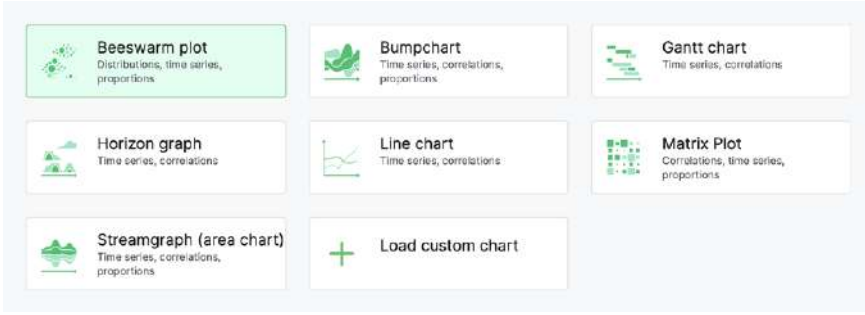


Figure 6 – Available chart templates for time series.

such as eccentricity or orbital inclination. We used a line chart for visualization: each line represents a family, the horizontal axis shows time, and the vertical axis indicates the number of discovered asteroids (Fig. 7).

From the resulting visualization, it appears that most asteroid discoveries occurred during the 2000s, or at least, this is what emerges from the Wikidata records. This raises an important question for those maintaining the data: does this pattern reflect actual astronomical activity, is it a result of how the data is added to Wikidata, or could it be due to a large dataset being imported by a single bot or source?

When working with temporal data, another recurring issue is that most visualization tools do not automatically account for gaps in the data. For example, if the dataset contains a value of 1 in 1950 and a value of 3 in 1960, RAWGraphs will simply draw a line between the two, without indicating the ten-year gap. To address this, it is recommended to prepare SPARQL queries that explicitly fill in these temporal gaps when using defined time units such as years. In cases where a regular time interval is not specified, such as years, months, or hours, it becomes essential to consider how these temporal voids should be represented or interpreted.

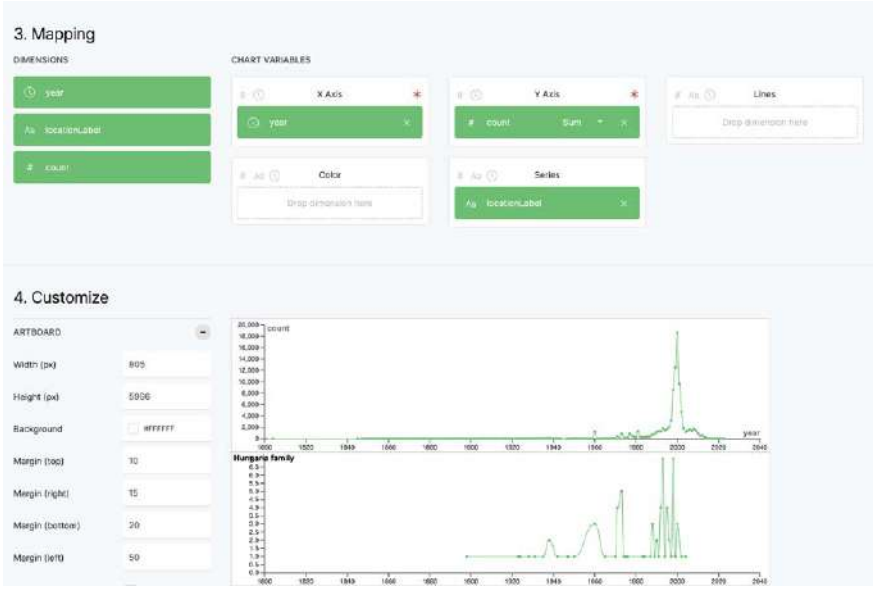


Figure 7 – Mapping and first two asteroids families from the query of all asteroids on Wikidata and their discovery dates. The image shows two series based on location: the first includes asteroids without a specified location, while the second includes those in the "Hungaria Family" group. Access the Wikidata query at: https://w.wiki/DqxV.

A second example of using this type of visualization with RAWGraphs is to illustrate the alternation of individuals in official roles, such as the Prime Min-

ister or President of the Italian Republic[6]. With the appropriate SPARQL query, one can retrieve the names of officeholders along with the start and end dates of their terms. In this case, we are dealing with temporal chunks: elements that have both a beginning and an end point in time. The Gantt chart is the most suitable RAWGraphs template for this scenario, as it draws a rectangle for each term of office, with the horizontal position and length representing its duration.

This second example is particularly compelling because it demonstrates how visualizations can synthesize complex temporal data into formats that are directly suitable for Wikipedia or related projects. However, it also enters a grey area concerning the "No original research" policy of Wikipedia (*Wikipedia* 2025a, "Wikipedia"). This raises the issue of whether the visualization of data from Wikidata constitutes a mere translation, therefore not original research, or if it creates a circular reference within the MediaWiki ecosystem (*Wikipedia* 2025b, "Wikipedia"), making it potentially unacceptable. This issue has been previously discussed in literature examining the relationship between diagrams and Wikipedia (Mauri, Pini, et al. 2017).

6. Custom chart template

The approach described above can be extended to any custom chart template that can be generated from data in tabular form. At present, RAWGraphs includes a variety of built-in chart templates suited to a wide range of visualization needs.

However, in the field of data visualization, many domain-specific solutions have emerged over time, specifically designed to address highly specialized use cases. One such example is the paired bar chart, which is commonly used to represent population distributions by age in the form of "population pyramids." This chart consists of two mirrored bar charts, centered and aligned vertically by age groups, typically arranged from youngest (at the bottom) to oldest (at the top).

RAWGraphs' architecture is explicitly designed to support third-party chart templates, following the same conceptual model used by its native visualizations (RAWGraphs Team 2023). Creating a custom chart requires first defining which components of the chart are data-driven and which are template-based. The developer must also declare the visual variables they intend to expose (e.g., bar length, color, position), and specify what types of data dimensions can be mapped to those variables (e.g., strings, numbers, dates).

Given these inputs, RAWGraphs automatically generates the interface that allows users to drag and drop data dimensions onto visual variables and to configure additional chart properties. This reduces the effort required to build user-facing interactions, allowing contributors to focus on the data-to-graphics transformation logic.

To contribute a custom chart, developers must write the JavaScript code necessary to render the chart using the data mappings provided by the interface.

[6] Query: <https://w.wiki/Dx7C>.

While RAWGraphs is designed to be agnostic with respect to the rendering library, a common practice at the time of writing is to rely on the D3.js library (Bostock et al. 2011), which offers extensive capabilities for building custom data visualizations.

Once created, a custom chart template can be built and shared online. One distinctive feature of RAWGraphs is that users can load third-party chart templates directly in the browser interface, without requiring installation or configuration. This design choice significantly lowers the barrier to reuse and encourages community contributions.

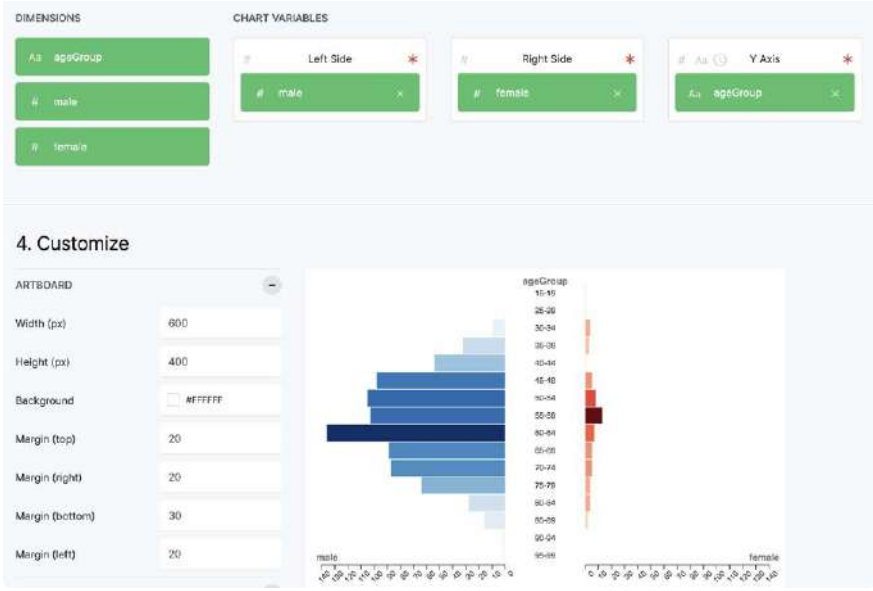


Figure 8 – Usage of the custom chart template "Paired Bar Chart" and its mapping to visualize the distribution of male and female Nobel Prize winners, grouped by the age at which they received the award. Access the Wikidata query at: https://w.wiki/Dsrc.

The Paired Bar Chart was developed by a group of contributors (@blindguardian50 et al. [2023] 2025) and is publicly available for use within the web interface. It can be used, for example, in combination with Wikidata. In Figure 8, we show how it can be used to visualize the number of Nobel Prize winners by gender and age, in order to explore potential patterns or biases[7]. As expected, the vast majority of recipients are male. Interestingly, we can also observe that female recipients tend to receive the prize at a younger age than their male counterparts. This example illustrates how visual exploration can help identify patterns that might otherwise remain hidden.

[7] Query: <https://w.wiki/Dsrc>.

By collecting data directly from Wikidata, it becomes feasible to retrieve properties (such as gender, birth date, award date) that would be difficult to obtain from conventional sources. Naturally, this raises questions about the completeness and accuracy of the data: for example, whether all Nobel Prize winners are represented in Wikidata or whether certain attributes are systematically missing. This type of post-evaluation is a crucial responsibility for anyone creating and interpreting such visualizations. The chart may highlight insights, but it must be read critically, with attention to what is shown and what may be absent.

7. Discussion: benefits and challenges of the RAWGraphs approach

The adoption of the "RAWGraphs approach," grounded in the principles of template-based design and open-ended visual workflows, appears particularly well-suited for visualizing data retrieved through SPARQL queries from Wikidata. While other data visualization tools aim for automated charting or polished dashboards, RAWGraphs instead prioritizes user agency, interpretability, and remixability. Testing this approach with real cases of Wikidata queries, it highlights a number of important benefits, as well as several persistent challenges. These emerge both from the technical architecture of the tool and from the epistemological implications of visualizing collaboratively-curated knowledge graphs.

One of the central strengths of RAWGraphs lies in its ability to offer a seamless visual exploration of tabular data. Given that SPARQL queries typically return results in the form of flat tables, consisting of rows of variable bindings, RAWGraphs' interface and conceptual model well align with this structure. Users can map columns onto visual variables exposed by chart templates and receive immediate feedback through previewed visualizations. This compatibility is particularly significant in the context of Wikidata, where data is semantically rich but often difficult to interpret through raw query results alone. By transforming such results into visual form, RAWGraphs enables users to more easily identify patterns, gaps, and anomalies that might otherwise remain hidden in textual outputs.

Another strength of RAWGraphs is its modular architecture, which allows the creation and integration of custom chart templates. This opens up the possibility of developing domain-specific visualizations tailored to Wikidata's unique structures. For instance, templates designed for genealogical relationships, taxonomies, or spatio-temporal event maps could support more precise representations of knowledge graph content. The ability to load third-party templates directly into the web interface, without installation or advanced configuration, lowers the barrier for experimentation and community development.

The process of producing a visualization with RAWGraphs is relatively smooth and intuitive, especially for users with some familiarity with data tables and basic principles of charting. However, we observed that the current SPARQL integration, while functional, is not fully optimized for complex querying tasks. In most cases, it remains preferable to compose and test SPARQL queries within the Wikidata Query Service interface. The service provides syntax highlighting,

runtime diagnostics, and result previews that are indispensable for debugging. Once the query is finalized, it can be copied into RAWGraphs for visualization. This two-step process is not problematic per se, but it does reflect the fact that a better integration between the two platforms, or a Wikidata-specific version of RAWGraphs would be beneficial.

While the tool's visual interface lowers the entry threshold for non-programmers, the act of visualization itself frequently surfaces the limitations of the underlying data.

The challenges of working with Wikidata-specific data structures remain. In the examples of visualizations explored in this paper, distinct limitations emerged. This includes issues such as duplicate entries, missing values, inconsistent units of measurement, and ambiguous categorization. For example, when plotting population values for cities, it is common to encounter multiple figures associated with different years or sources, resulting in apparent duplication. Similarly, in hierarchical charts such as treemaps or circle packings, users may struggle to produce coherent groupings due to the lack of standardized taxonomies within Wikidata. Genres for musical bands, for instance, may vary widely in specificity, language, and formatting, making it difficult to derive clear structures from the data.

These challenges underscore one of the most significant insights from our experimentation: visualizations derived from Wikidata data often tell us more about the platform itself (its structure, contributions, and biases) than about the phenomena ostensibly being represented. This interpretive shift is not a shortcoming of visualization, but rather a reminder of its critical value. Through visual exploration, users are able to understand not only what the data appears to say, but also what it omits, where it contradicts itself, and how it has been constructed. In this sense, RAWGraphs functions not just as a visualization engine, but also as a diagnostic tool that exposes the partial, evolving, and contested nature of collaborative data infrastructures.

Importantly, the visualization process can also inform and refine the process of query formulation itself. Several examples from our tests have shown that attempting to visualize a query result often leads to a reassessment of the query structure, as users confront unexpected outputs, unbalanced distributions, or implausible values. Visualization becomes a mechanism for iterative query development and data validation: a feedback loop where seeing the data helps users better define what they want to retrieve. This reinforces the idea that visualization should be understood not only as a method of communication but also as a mode of reasoning and exploration.

8. Acknowledgements

We would like to thank the other creators of RAWGraphs 2.0 beyond the authors of this article: Matteo Azzi and Giorgio Uboldi from Calibro Studio, and Mauro Bianchi from InMagik. We also wish to acknowledge Giorgio Caviglia, the original creator of the first version of the tool, for his foundational work on the project.

A special thanks goes to all the individuals, companies, and organizations who supported the RAWGraphs project through their donations. Without their contributions, the development of the platform would not have been possible. In particular, we are grateful to Zazuko, whose support included the development of the code that enables direct integration of SPARQL queries.

Finally, we would like to recognize the many members of the DensityDesign Lab network who contributed to the project through bug fixes, thoughtful suggestions, improvements to documentation, and ongoing engagement with the tool's development and use.

References

@blindguardian50, @steve1711, @TheAlmightySpaceWarrior, @wizardry8, and @kandrews99. (2023) 2025. *Rawgraphs-Paired-Barchart.* JavaScript. June 29, released February 4. https://github.com/rawgraphs/rawgraphs-paired-barchart.

"Wikidata Homepage." 2025. https://w.wiki/7aBH.

"Wikidata:Tools/Visualize Data." 2025. https://w.wiki/DrNN.

Bertin, Jacques. 1967. *Sémiologie Graphique.* Paris: Mouton/Gauthier-Villars.

Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer. 2011. "D3 Data-Driven Documents." *IEEE Transactions on Visualization and Computer Graphics* 17 (12): 2301-9. https://doi.org/10.1109/TVCG.2011.185.

Bounegru, Liliana, Jonathan Gray, Tommaso Venturini, and Michele Mauri. 2018. *A Field Guide to "Fake News" and Other Information Disorders.* Zenodo. https://doi.org/10.5281/ZENODO.1136271.

Camarda, Diego Valerio, Silvia Mazzini, and Alessandro Antonuccio. 2012. "LodLive, Exploring the Web of Data." In *Proceedings of the 8th International Conference on Semantic Systems. I-SEMANTICS '12,* 197-200. New York: ACM. https://doi.org/10.1145/2362499.2362532.

Das, Paramita, Sai Keerthana Karnam, Anirban Panda, Bhanu Prakash Reddy Guda, Soumya Sarkar, and Animesh Mukherjee. 2023. "Diversity Matters: Robustness of Bias Measurements in Wikidata." In *Proceedings of the 15th ACM Web Science Conference 2023. WebSci '23,* 208-18. New York: ACM. https://doi.org/10.1145/3578503.3583620.

Del Pino, Benjamín, and Aidan Hogan. 2023. "Wikidata Atlas: Putting Wikidata on the Map." In *Companion Proceedings of the ACM Web Conference 2023,* 238-41. Austin: ACM. https://doi.org/10.1145/3543873.3587356.

Elhossary, Muhammad, and Konrad U. Förstner. 2024. "InteractOA: Showcasing the Representation of Knowledge from Scientific Literature in Wikidata." *Semantic Web* 15 (6): 2381-93. https://doi.org/10.3233/SW-243685.

Gómez-Romero, Juan, Miguel Molina-Solana, Axel Oehmichen, and Yike Guo. 2018. "Visualizing Large Knowledge Graphs: A Performance Analysis." *Future Generation Computer Systems* 89: 224-38. https://doi.org/10.1016/j.future.2018.06.015.

Lima, Manuel. 2014. *The Book of Trees: Visualizing Branches of Knowledge.* New York: Princeton Architectural Press.

Mauri, Michele, and Paolo Ciuccarelli. 2016. "Designing Diagrams for Social Issues." Paper presented at the Design Research Society Conference 2016, Brighton, June 25. https://doi.org/10.21606/drs.2016.185.

Mauri, Michele, Azzurra Pini, and Paolo Ciuccarelli. 2017. "Designing Diagrams for Wikipedia." *Information Design Journal* 23 (1): 65-79. https://doi.org/10.1075/idj.23.1.08mau.

Mauri, Michele, Tommaso Elli, Giorgio Caviglia, Giorgio Uboldi, and Matteo Azzi. 2017. "RAWGraphs: A Visualisation Platform to Create Open Outputs." In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter,* September 18, 1-5. https://doi.org/10.1145/3125571.3125585.

Mecharnia, Thamer, and Mathieu d'Aquin. 2025. "Performance and Limitations of Fine-Tuned LLMs in SPARQL Query Generation." In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK),* edited by Genet Asefa Gesese, Harald Sack, Heiko Paulheim, Albert Merono-Penuela, and Lihu Chen. International Committee on Computational Linguistics. https://aclanthology.org/2025.genaik-1.8/.

Metilli, Daniele, Valentina Bartalesi, and Carlo Meghini. 2019. "A Wikidata-Based Tool for Building and Visualising Narratives." *International Journal on Digital Libraries* 20 (4): 417-32. https://doi.org/10.1007/s00799-019-00266-3.

Nielsen, Finn Årup, Daniel Mietchen, and Egon Willighagen. 2017. "Scholia, Scientometrics and Wikidata." https://doi.org/10.1007/978-3-319-70407-4_36.

Raissya, Hana, Fariz Darari, and Fajar J. Ekaputra. 2021. "VizKG: A Framework for Visualizing SPARQL Query Results over Knowledge Graphs." In *VOILA! 2021,* 95-102. Aachen: TIBKAT. https://www.tib.eu/de/suchen/id/TIBKAT%3A1816498505.

RAWGraphs Team. 2017. "A Visual Exploration of Musical Bands on Wikidata." *RAWGraphs,* July 5. https://www.rawgraphs.io/post/a-visual-exploration-of-musical-bands-on-wikidata.

RAWGraphs Team. 2023. "Introducing a New Feature in RAWGraphs: On-the-Fly Custom Charts." *RAWGraphs,* July 7. https://www.rawgraphs.io/post/introducing-a-new-feature-in-rgraphs-on-the-fly-custom-chart.

Shaik, Zaina, Filip Ilievski, and Fred Morstatter. 2021. "Analyzing Race and Citizenship Bias in Wikidata." In *2021 IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS),* October, 665–66. https://doi.org/10.1109/MASS52906.2021.00099.

Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: A Free Collaborative Knowledgebase." *Communications of the ACM* 57 (10): 78-85. https://doi.org/10.1145/2629489.

Wikipedia. 2025a. "Wikipedia:No Original Research." April 20. https://w.wiki/Dx6$.

Wikipedia. 2025b. "Wikipedia:Verifiability." April 21. https://w.wiki/Dx7G.

Zhang, Charles Chuankai, and Loren Terveen. 2021. "Quantifying the Gap: A Case Study of Wikidata Gender Disparities." In *Proceedings of the 17th International Symposium on Open Collaboration. OpenSym '21,* 1-12. New York: ACM. https://doi.org/10.1145/3479986.3479992.

Hunting for Lost Heritage on Wikimedia Commons and Wikidata. Un workflow per scovare i monumenti italiani

Marco Chemello

Abstract: Wikidata represents an incredible revolution in open knowledge curation, yet its complexity and structure often limit its visibility and use, particularly when it comes to data visualization. RAWGraphs 2.0 addresses this challenge by providing an intuitive, template-based approach for translating Wikidata's SPARQL query outputs into meaningful visualizations. Drawing on data visualization literature, RAWGraphs employs visualization templates: conceptual and technical schemas that identify visual properties of graphical items and bind them to data. With direct SPARQL integration, RAWGraphs supports transparent, reproducible visual exploration, highlighting data quality, biases, and structural insights within Wikidata.

Keywords: Wikimedia, Wikimedia Commons, Wikidata, Heritage, Churches, crowdsourcing, AI, cultural heritage.

1. Introduzione

Il patrimonio monumentale italiano comprende centinaia di migliaia di edifici, opere d'arte e reperti archeologici, molti dei quali – anche grazie al concorso fotografico annuale *Wiki Loves Monuments* – sono già presenti su Wikidata, talvolta anche con una voce su Wikipedia. Wikidata in effetti da anni costituisce il maggiore catalogo di dati sui beni culturali italiani pubblicamente disponibile.

Molti monumenti e beni culturali, tuttavia, sono documentati solo fotograficamente su Wikimedia Commons ma non hanno una voce su Wikipedia e risultano inoltre privi di un collegamento strutturato a Wikidata (Fig. 1). Alcuni di questi monumenti sono presenti in database pubblici (es. chiese di culto), altri risultano completamente sconosciuti (es. chiese in rovina in aree isolate).

Le liste di Wiki Loves Monuments nel 2023 elencavano 64555 edifici religiosi[1] su Wikidata (su un totale di 85821 di monumenti), ottenuta grazie all'integrazione di varie banche dati pubbliche, il che chiaramente non rappresenta

[1] Wiki Loves Monuments 2023 – Statistiche, Wikimedia Italia (rev. 16 apr 2025). <https://wiki.wikimedia.it/wiki/Wiki_Loves_Monuments/2023>.

Marco Chemello, Wikimedia Italia, Italy, marco.chemello@wikimedia.it, 0000-0001-7446-8800

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Marco Chemello, *Hunting for Lost Heritage on Wikimedia Commons and Wikidata. Un workflow per scovare i monumenti italiani,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.29, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 239-248, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

ancora la totalità degli edifici o ex edifici religiosi presenti in Italia. Inoltre il 97% di questi elementi censiti non era ancora collegato alla relativa categoria fotografica su Commons (o ad altri progetti Wikimedia). Tali lacune informative riducono la visibilità dei beni e ne limitano il riuso nella ricerca, nella didattica e nelle applicazioni turistiche, ma anzitutto rendono difficile o impossibile la loro salvaguardia. Il lavoro di crowdsourcing, specie quando compiuto da reti organizzate di volontari, aiuta la collettività e le istituzioni ad identificare il patrimonio culturale e naturale in pericolo[2].

Il progetto personale *Hunting for Lost Heritage* (2023-25) affronta il problema di colmare le lacune attraverso una strategia di data mining in crowdsourcing basata sul volontariato, offrendo una soluzione pratica e rispondendo a tre domande di ricerca: (1) quante categorie di Commons relative ai monumenti italiani sono prive di collegamento a Wikidata; (2) con quali tempi e risorse i volontari possono riconciliare o creare gli item mancanti; (3) in che misura l'intelligenza artificiale potrebbe ridurre il carico manuale di revisione e arricchimento.

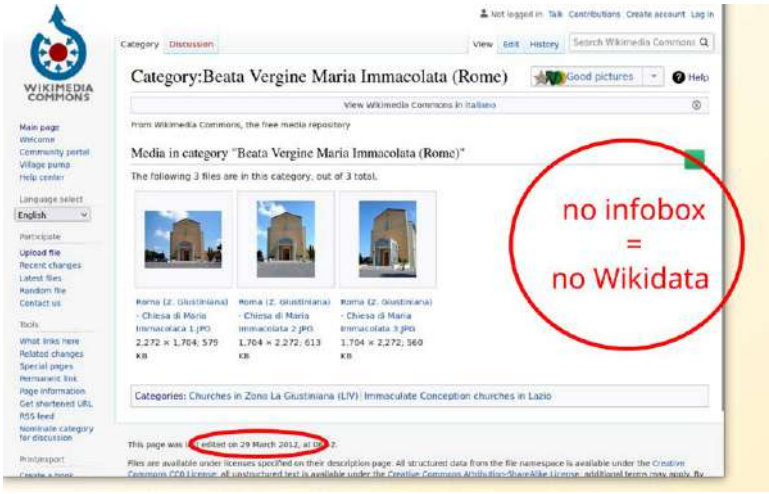


Figura 1 – Esempio di categoria su Wikimedia Commons (creata nel 2012) ancora priva di infobox e collegamento a Wikidata. Chiesa della Beata Vergine Immacolata di Roma.

2. Contesto e stato dell'arte

I volontari della comunità di Wikimedia si dedicano da due decenni a documentare il patrimonio culturale e naturale locale, scattando e caricando milioni di fotografie che lo ritraggono, producendo decine di migliaia di voci di Wikipedia

[2] Ćwik, N. 2024. ÒA New Initiative Aims to Protect Endangered Heritage Using Crowdsourcing.Ó Diff blog, 25 marzo 2024. URL: <https://diff.wikimedia.org/2024/03/25/a-new-initiative-aims-to-protect-endangered-heritage/>.

che lo descrivono, in oltre 300 lingue, e raccogliendo dati dai database pubblici che lo riguardano per incrociarli su Wikidata. La letteratura riconosce, in particolare, il potenziale del concorso fotografico Wiki Loves Monuments come iniziativa di crowdsourcing su scala mondiale;[3] il concorso è stato attestato nel 2012 dal Guinness World Records come il più grande concorso fotografico al mondo[4] ed è basato su un'ampia comunità di volontari locali.

Strumenti come PetScan[5] consentono di estrarre liste complesse da categorie di Commons incrociando proprietà di Wikidata, mentre OpenRefine offre una soluzione open source per riconciliare, creare e modificare in maniera massiccia entità di Wikidata e – più recentemente – di Wikimedia Commons.

Sul fronte normativo, i recenti decreti ministeriali (DM 161/2023[6], parzialmente corretto dal DM 108/2024[7]) hanno sollevato nel mondo della ricerca e non solo notevoli questioni sulla libertà d'uso delle immagini dei beni culturali, ed anche potenziali conflitti, dato che le *Linee guida del Ministero della Cultura sul riuso delle riproduzioni digitali dei beni culturali,* parte del Piano nazionale di digitalizzazione del patrimonio culturale, incoraggiano licenze aperte quando il bene è in pubblico dominio[8].

All'interno del movimento Wikimedia, il progetto Heritage Guard Network (HGN), finanziato dallo Swedish Institute, promuove la salvaguardia dei siti a rischio attraverso crowdsourcing e reti locali di volontari[9].

3. Obiettivi specifici

Il progetto *Hunting for Lost Heritage* intende:

1) Mappare le categorie di Commons prive di collegamento a Wikidata relative a monumenti italiani, edifici religiosi, musei, cimiteri monumentali e opere d'arte,
2) Creare o arricchire gli item di Wikidata con proprietà fondamentali (tipo di bene, ubicazione) e metadati tematici (autore, soggetto),

[3] Azizifard, N. et al. 2022. ÒWiki Loves Monuments: Crowdsourcing the Collective Image of the Worldwide Built Heritage.Ó *ACM Journal on Computing and Cultural Heritage* 16 (1). DOI: 10.1145/3569092. <https://dl.acm.org/doi/10.1145/3569092>.

[4] Guinness World Records. 2012. ÒLargest photography competition.Ó <https://www.guinnessworldrecords.com/world-records/largest-photography-competition>.

[5] PetScan – WMcloud.org, <https://petscan.wmcloud.org/>.

[6] Ministero della Cultura. 2023. Decreto ministeriale 11 aprile 2023, n. 161 – ÒLinee guida per l'acquisizione, la circolazione e il riuso delle riproduzioni dei beni culturali.Ó <https://www.beniculturali.it/mibac/multimedia/MiC/documents/2023/04/18/DM_161_2023.pdf>.

[7] Ministero della Cultura. 2024. Decreto ministeriale 21 marzo 2024, n. 108 – Modifiche al DM 161/2023. <https://www.beniculturali.it/mibac/multimedia/MiC/documents/2024/03/25/DM_108_2024.pdf>.

[8] ÒLinee guida per l'acquisizione, la circolazione e il riuso delle riproduzioni dei beni culturali in ambiente digitale.Ó (vers. 2022-2023) URL: <https://docs.italia.it/italia/icdp/icdp-pnd-circolazione-riuso-docs/it/consultazione/index.html>.

[9] Heritage Guard Network. 2025. ÒHeritage Guard NetworkÓ. Meta-Wiki, rev. febbraio 2025. <https://meta.wikimedia.org/wiki/Heritage_Guard_Network>.

3) Misurare efficienza e precisione del workflow ampiamente basato su OpenRefine,
4) Sperimentare strumenti AI (LLMs) per identificare elementi da creare e/o suggerire label e proprietà.

4. Metodologia

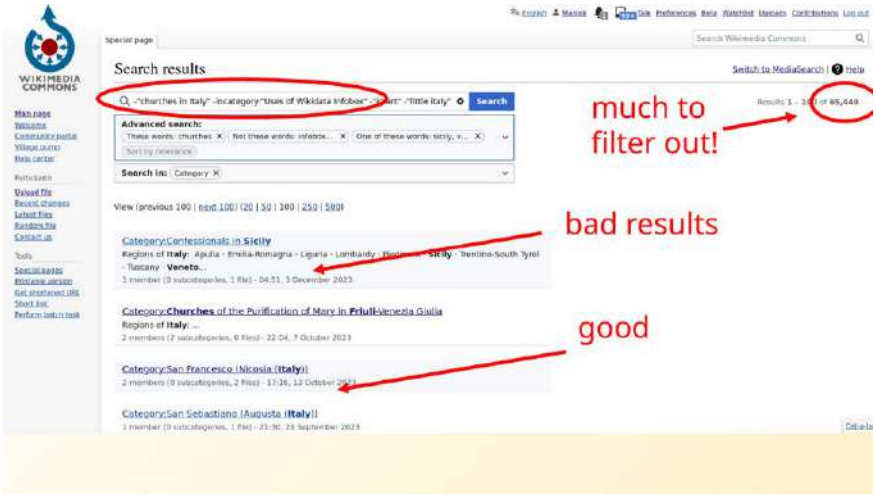


Figura 2 – Esempio di risultati con il sistema di ricerca full-text interno di Wikimedia Commons, cercando categorie con «churches in Italy» senza «Wikidata infobox».

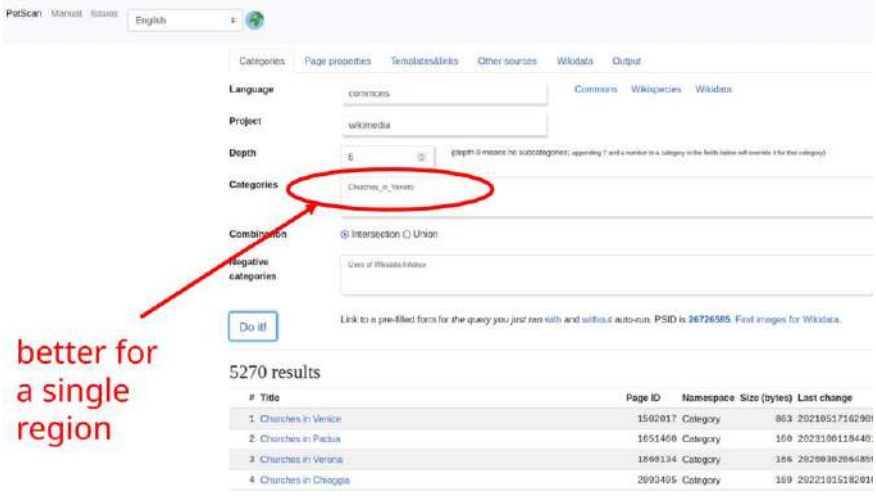


Figura 3 – Esempio di ricerca usando PetScan per ottenere un elenco delle chiese senza Wikidata Infobox. Per evitare errori di time-out, è stato necessario segmentare la ricerca per regione, o per provincia.

4.1 Estrazione dati con PetScan

Sono state impostate query su PetScan selezionando il progetto Wikimedia Commons, filtrando il namespace Category sottraendo la categoria «Uses of Wikidata Infobox» ma anche «paintings», «sculptures» ecc. e restringendo gradualmente la profondità. Categorie madre come «churches in Italy» (che contiene quasi 700 mila immagini organizzate in circa 100mila sottocategorie), «museums in Italy» e «burials in Italy» (Fig. 3) hanno prodotto (una volta segmentati e filtrati) oltre 17.000 titoli di categoria. Sfortunatamente PetScan non si è sempre dimostrato uno strumento efficiente (per problemi tecnici del server o limiti strutturali) e inizialmente, a causa dei frequenti errori di time-out del server, è stato necessario accantonarlo e procedere manualmente utilizzando il sistema di ricerca di Commons, più limitato e impreciso (Fig. 2). L'affidabilità di PetScan è migliorata utilizzando sottoinsiemi più piccoli ed è stato quindi possibile utilizzarlo nella Fase 2. Per non incorrere nell'errore da eccesso di elementi (teoricamente 500 mila, in pratica molti meno), le query con PetScan sono state segmentate prima per regione e poi per provincia (Fig. 3).

4.2 Pulizia preliminare

I CSV esportati sono stati puliti in una prima fase manualmente in text editor o LibreOffice, semplicemente operando un ordinamento alfabetico e rimuovendo i casi più evidenti di duplicati, categorie descrittive generiche o non rilevanti (es. le sottocategorie per gli interni o esterni delle chiese). Questa operazione preliminare aveva lo scopo di ridurre la dimensione del file (da 65 mila a meno di 25 mila righe).

4.3 Importazione in OpenRefine, filtraggio e trasformazioni

I dati sono stati importati in OpenRefine dove, grazie all'estensione Wikidata e tramite il linguaggio GREL, si sono generate colonne derivate per label multilingue e altre dichiarazioni (Fig. 4). Sono state applicate espressioni regolari per normalizzare i nomi e isolare città o dedicazioni. Ciò ha nel contempo permesso di filtrare ulteriormente i dati, eliminando circa il 30% dei record (categorie non pertinenti a singoli beni, o duplicate).

Di ogni potenziale item sono state create label e descrizioni più ricche e complete della media in italiano e inglese, unendo più proprietà/colonne create (Fig. 4), come tipologia («istanza di» – P31), città («unità amministrativa in cui è situato» – P131), Paese (P17), dove possibile località («luogo» – P276), più – a seconda dei casi – titolazioni («prende il nome da» – P138) nel caso di edifici religiosi, rappresentazioni («raffigura» – P180) e autori nel caso di arte figurativa, commemorazioni («commemora» – P547) nel caso di monumenti e tombe. Questa operazione – che era in parte necessaria per garantire l'univocità del matching tra le tante omonimie – ha richiesto un lungo (oltre il 50% del tempo totale) e paziente lavoro di integrazione dei dati, che è stato svolto soprattutto manualmente con l'ausilio di OpenRefine, ma che in futuro potrebbe essere facilitato almeno in parte dall'ausilio di IA (LLMs).

4.4 Riconciliazione e creazione di entità

Utilizzando il servizio di riconciliazione interno di OpenRefine, i record sono stati abbinati a entità di Wikidata esistenti, identificati i matching possibili/probabili da sottoporre a revisione umana, infine marcati per la creazione. Gli item sono stati creati e quelli esistenti integrati delle nuove informazioni, precisandone la provenienza e l'aggiornamento nelle references di ogni dichiarazione.

Sono state adottate diverse strategie per migliorare il matching e ridurre al minimo il rischio di creare entità duplicate (comunque presente). Sono stati condotti molti prudenti test multipli di caricamento dati su Wikidata per individuare le carenze e le non conformità dello schema dei dati.

Oltre ad avere creato o arricchito le entità dei monumenti, come lavoro aggiuntivo sono stati isolati e aggiunti su Wikidata anche dati relativi a numerose località, intestatari, temi artistici e artisti, sia creando nuove entità, sia arricchendo, al termine del lavoro principale, quelle esistenti con dichiarazioni derivate (es. «luogo di sepoltura» – P119, «opere rilevanti» – P800) o inverse (es. «raffigurato da» – P1299).

4.5 Editing massivo via OpenRefine

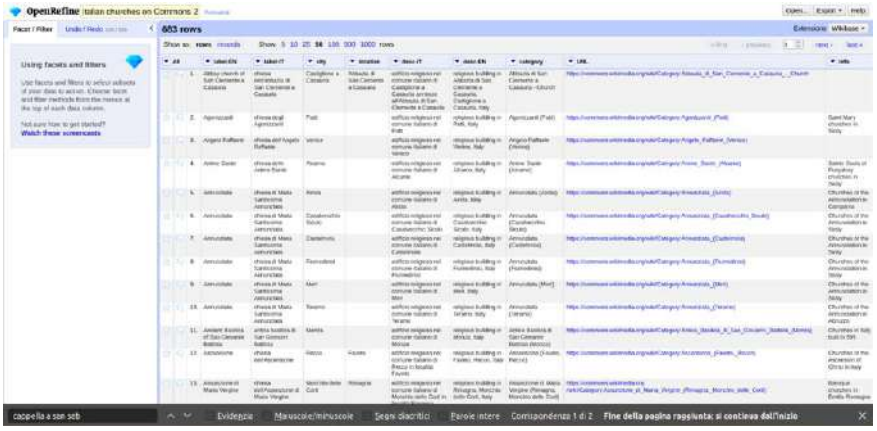


Figura 4 – Foglio di lavoro di OpenRefine con una lista di edifici religiosi ricavati dalle categorie di Commons.

OpenRefine gestisce creazione e aggiornamento in un'unica interfaccia (Fig. 4) e offre svariate funzioni di manipolazione, filtraggio e miglioramento dei dati. Batch di progressiva dimensione, fino a 500 record, sono stati caricati e verificati manualmente, con rollback inferiore all'1%. Ogni batch poteva essere facilmente annullato, in caso di necessità, da qualunque utente.

In alternativa a OpenRefine, per l'editing massivo su Wikidata esiste il tool QuickStatements, utilizzando come base dei fogli dati creati con OpenOffice/LibreOffice. OpenRefine tuttavia presenta vari vantaggi, essendo creato per riconci-

liare i database e possedendo molte funzionalità aggiuntive (ad es. per il filtraggio). D'altro canto, OpenRefine ha una curva di apprendimento più ripida per l'utente.

4.6 Metriche e tempi

Il tempo totale impiegato per la Fase 1 (durata 3 mesi) è stato di circa 100 ore di lavoro/persona, con una media di 1 ora/giorno, per lavorare 6 500 soggetti / item di Wikidata (nuovi o aggiornati). Il tempo medio è stato dunque di circa 1 minuto per entità; nella Fase 2 è sceso a circa 0,7 grazie all'esperienza maturata. È possibile ipotizzare per il futuro una ulteriore significativa riduzione dei tempi attraverso l'utilizzo di IA (LLMs) nella maggior parte delle fasi, anche se non potrà mai mancare un'attenta revisione umana in ogni fase del processo.

4.7 Sperimentazione AI (LLMs)

Nella Fase 3 (ancora in corso) sono stati testati vari modelli di chatBot LLM a partire da quelli pubblici gratuiti (tra i quali Chat GPT-3, GPT-4o, GPT 4.5, DeepSeek), utilizzando prompt per suggerire label/descrizioni e individuare nuovi elementi da creare Wikidata.

I primi esperimenti hanno evidenziato importanti criticità con la maggior parte dei modelli, che non sembrano ottimizzati per lavorare con Wikidata con la precisione e l'affidabilità necessarie. Questo potrebbe essere inoltre influenzato dalla qualità dei prompt, inizialmente semplici, che potrebbero essere migliorati e resi più complessi. Solo Chat GPT-4.5 (a pagamento) ha finora dimostrato di essere in grado di produrre risultati sufficientemente affidabili, individuando alcuni ridotti set di elementi da creare su Wikidata sulla base dell'analisi delle categorie di Commons (Fig. 5).

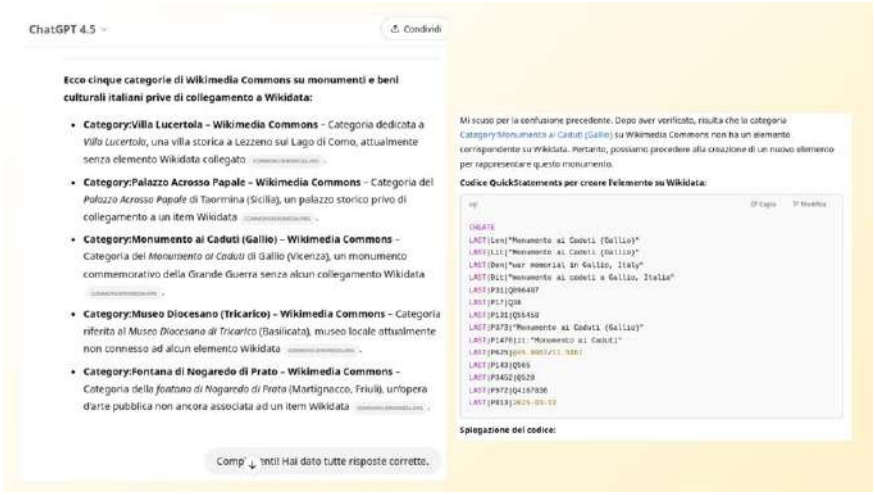


Figura 5 – Risultati con ChatGPT 4.5 (colonna 1) e generazione codice QuickStatements (colonna 2).

In generale gli LLM testati sono stati quasi sempre in grado di generare del codice per QuickStatements per modificare i dati su Wikidata, con qualche correzione iniziale (Fig. 5).

5. Risultati

Nella Fase 1 (edifici religiosi in Italia) sono state identificate oltre 6.500 categorie 'orfane' esistenti su Wikimedia Commons; sono stati creati 3.000 nuovi item su Wikidata e arricchiti 4.000 esistenti, aggiungendo il link alla categoria.

Nella Fase 2 (musei, cimiteri, opere d'arte italiane) le categorie individuate hanno superato le 10.000, con 9.000 item nuovi e 1.500 aggiornati su Wikidata. Perlopiù si trattava ancora di beni culturali in Italia; talvolta le entità si riferivano ad opere d'arte italiane presenti in musei stranieri.

Il progetto ha, nella Fase 1, aumentato la copertura dei monumenti italiani in Wikidata di circa il 5% rispetto al baseline di Wiki Loves Monuments 2023 (edifici religiosi): un risultato di molto superiore alle aspettative iniziali.

La creazione dei collegamenti tra le categorie di Commons e Wikidata ha aperto la strada al lavoro della comunità: vari utenti nelle settimane e mesi successivi hanno potuto correggere ma soprattutto aggiungere ulteriori informazioni in numerose entità di Wikidata, come le coordinate (spesso via bot) e le immagini rappresentative (scelte invece manualmente), superando il 'confine' percepito tra le due comunità. Tale lavoro è tuttora in corso.

6. Discussione

Il workflow completamente basato su OpenRefine è accessibile a volontari con esperienza dei progetti Wikimedia di livello intermedio: si stima che una sessione di formazione di 2-4 ore potrebbe essere sufficiente per raggiungere produttività maggiori di 200 edit in un'ora. L'approccio è replicabile per altri domini (altre nazioni, altri temi), specie dove le categorie di Commons seguono pattern coerenti.

7. Limiti e sviluppi futuri

- Scalabilità: il limite di 50.000 risultati (nel caso migliore) di PetScan richiede segmentazione o PagePile,
- Bias geografici: copertura ancora scarsa in molti luoghi isolati e del Sud Italia e Isole; future campagne potrebbero mirare ad esempio a Sicilia e Sardegna,
- Qualità metadati museali: mancano collegamenti authority control (es. ULAN/VIAF) per molte opere d'arte,
- Explainability in AI: necessità di interfacce che espongano lo score di confidenza dei suggerimenti,
- Miglioramento dell'uso degli strumenti LLM, con prompt più complessi che forniscano risultati più affidabili,
- Affidabilità delle informazioni di crowdsourcing su Wikimedia Commons: in alcuni casi (molto ridotti, 1-2%) i nomi delle categorie erano attribuzioni

errate, oppure alcune categorie erano duplicate. Nel complesso il progetto ha permesso di individuare e correggere molti di questi errori, aprendo al contempo la possibilità di individuare più facilmente i rimanenti tramite Wikidata.

8. Conclusioni

Il progetto *Hunting for Lost Heritage* dimostra che è possibile arricchire la conoscenza libera del patrimonio culturale facendo data mining sui progetti Wikimedia in molti modi, non sempre evidenti. L'uso combinato di PetScan, OpenRefine e volontari formati può contribuire a colmare rapidamente le lacune dei database pubblici e di Wikidata sui monumenti italiani. La metodologia è aperta, documentata e replicabile da volontari, enti e istituzioni GLAM di tutto il mondo che intendano rilasciare i propri cataloghi come Linked Open Data. L'apporto di IA (LLMs) potrà probabilmente nel prossimo futuro ridurre il lavoro manuale e aumentare l'efficienza dei volontari.

9. Ringraziamenti

Si ringraziano i volontari del movimento Wikimedia internazionale e di Wikimedia Italia per avere prodotto con il loro lavoro, negli ultimi due decenni, l'enorme patrimonio di immagini di beni culturali su Wikimedia Commons di cui disponiamo, nonché i relatori dei *Wikidata Days* 2024 di Bologna per i suggerimenti gentilmente prestati. Si ringraziano gli organizzatori del primo convegno *Wikidata e la ricerca* di Firenze del 2025.

Riferimenti bibliografici

AIB Associazione Italiana Biblioteche. 2023. "Osservazioni sul DM 11 aprile 2023, n. 161." *AIB Notizie,* 12 maggio 2023. <https://www.aib.it/notizie/osservazioni-sul-dm-11-aprile-2023-n-161/>.

Azizifard, N., et al. 2022. "Wiki Loves Monuments: Crowdsourcing the Collective Image of the Worldwide Built Heritage." *ACM Journal on Computing and Cultural Heritage* 16 (1). <https://doi.org/10.1145/3569092>.

Chemello, M. 2025. *Hunting for Lost Heritage on Wikimedia* [presentation slides]. <https://commons.wikimedia.org/wiki/File:Hunting_for_lost_heritage_on_Wikimedia.pdf>.

Ćwik, N. 2024. "A New Initiative Aims to Protect Endangered Heritage Using Crowdsourcing." *Diff blog,* 25 marzo 2024. <https://diff.wikimedia.org/2024/03/25/a-new-initiative-aims-to-protect-endangered-heritage/>.

Grossi, P. G. 2024. "Nuove regole sulle riproduzioni dei beni culturali: saremo più libere e liberi di usarle?" *Arkivia,* Wikimedia Italia, 18 aprile 2024. <https://www.wikimedia.it/news/nuove-regole-sulle-riproduzioni-dei-beni-culturali-saremo-piu-libere-e-liberi-di-usarle/>.

Guinness World Records. 2012. "Largest Photography Competition." <https://www.guinnessworldrecords.com/world-records/largest-photography-competition>.

Heritage Guard Network. 2025. "Heritage Guard Network." *Meta-Wiki,* rev. febbraio 2025. <https://meta.wikimedia.org/wiki/Heritage_Guard_Network>.

Meta Wikimedia. s.d. "PetScan." *Meta-Wiki.* <https://meta.wikimedia.org/wiki/PetScan/en>.

Ministero della Cultura. 2023. *Decreto ministeriale 11 aprile 2023, n. 161 – Linee guida per l'acquisizione, la circolazione e il riuso delle riproduzioni dei beni culturali.* <https://www.beniculturali.it/mibac/multimedia/MiC/documents/2023/04/18/DM_161_2023.pdf>.

Ministero della Cultura. 2024. *Decreto ministeriale 21 marzo 2024, n. 108 – Modifiche al DM 161/2023.* <https://www.beniculturali.it/mibac/multimedia/MiC/documents/2024/03/25/DM_108_2024.pdf>.

OpenRefine. 2024. "Reconciling." *OpenRefine Manual* v. 3.8, 20 dicembre 2024. <https://openrefine.org/docs/manual/reconciling>.

Wiki Loves Monuments Italia. 2025. "Progetto Wiki Loves Monuments Italia: storia e statistiche." *Wikipedia/Wikimedia Italia.* <https://it.wikipedia.org/wiki/Progetto:Wiki_Loves_Monuments_Italia/Storia_e_statistiche>.

Wikimedia Deutschland. 2025. "Wikidata: Embedding Project." *Wikidata*, aprile 2025. <https://www.wikidata.org/wiki/Wikidata:Embedding_Project>.

Whose History We Keep: Benchmarking our Records of the Past's Protagonists

Lennart Finke, Sarah Sophie Pohl, Luca Lukačević,
Janek Große, Stefan Haas

Abstract: Historical science compounds millennia of records of human activity, which have become available in digital, searchable form. Focusing on the largest such dataset—Wikidata—we attempt to find quantitative answers to the questions: Whose periods and which places' people do we know most about? What interests us about them? Who are we forgetting? What trends underlie the number of people registered and written about over time? We examine the individuals who have ended up in these datasets in relation to their readership, thereby shedding light not only on history itself but also on the practice of writing it.

Keywords: Digital History, Methodology of History, Open Data, Wikidata.

1. Introduction

History as a quantitative science is supplied with increasingly comprehensive data, allowing us to ask questions that seemed out of reach previously. Large data sets assembled by historians can unify the biographies of millions of individuals into a single place, making them available in standardised, quantitative formats. Schich et al. (2014, 558–62) made an innovative first step and provided a scientific derivation of Freebase, a now defunct semantic data base. Yu et al. (2016, 1–16) then created *Pantheon* from the open encyclopaedias Wikidata and Wikipedia. They cite the *Book of Human Accomplishments* by Murray (2003) as another inspiration. An impressive qualitative and quantitative leap forward is *A Brief History of Human Time* by Laouenan et al. (2022). They automate a cross-verification process with natural language processing, sporadically verifying it by hand, to create a validated database of an impressive 2.2 million people.

As our entry point, we revisit a generalisation of a question by Arbesman (2013), who calculated the "fraction of famous people" in the world by referring to the size of the living people category on the English Wikipedia. He gives a

Lennart Finke, ETH Zurich, Switzerland, lfinke@student.ethz.ch, 0009-0003-6908-314X
Sarah Sophie Pohl, University of Göttingen, Germany,
Luca Lukačević
Janek Große
Stefan Haas, University of Göttingen, Germany, stefan.haas@phil.uni-goettingen.de

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Lennart Finke, Sarah Sophie Pohl, Luca Lukačević, Janek Große, Stefan Haas, *Whose History We Keep: Benchmarking our Records of the Past's Protagonists,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.30, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 249-261, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

fraction between $4.1 \cdot 10^{-4}$, dividing by the anglophone world population and $8.6 \cdot 10^{-5}$ dividing by the total world population. After entertaining related questions, we also give an epistemological critique of these quantitative tools, uncovering some areas where big data might not lead to big insights and instead mislead us. After all, it takes two to make a record – the recorded and the recording individual. We will always see our own reflections in who we deem worthy of recording and what we record about them.

2. Methods

We collect and analyse the largest sources of historical data currently in use, where we focus first on Wikidata derivatives. Our approach is twofold: we look at previously collected datasets, but also query Wikidata ourselves.

We want to find the time distribution of the biographical information collected on Wikidata. The database is accessible via a SPARQL endpoint, a web

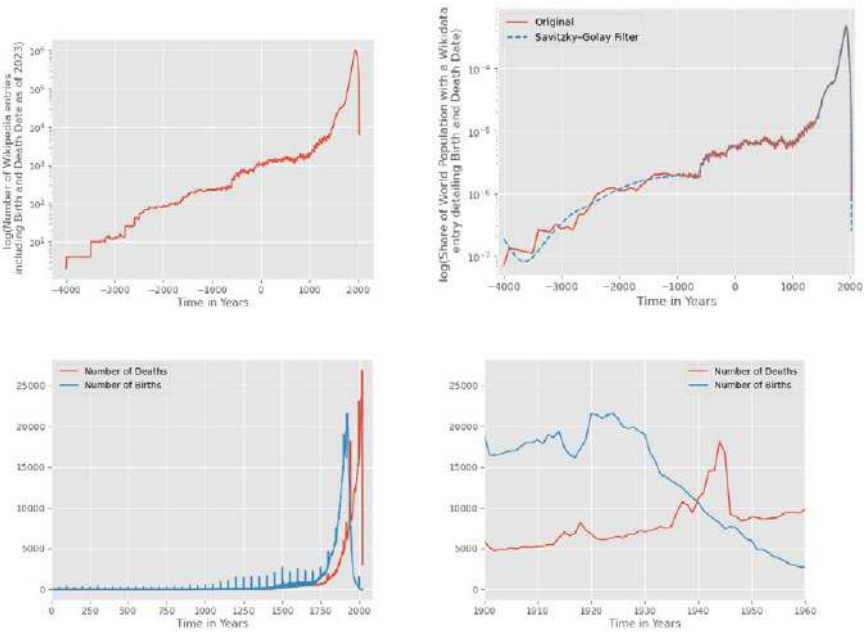


Figure 1 – Top row: Raw Births and Deaths per Year in all language editions of Wikipedia on a linear scale from 0 AD until today, with a close-up for the first half of the century on the right. Bottom left: The integrated difference giving the "population of Wikipedia" on a semi-log scale. Bottom right: Dividing by the world population gives the fraction of the people earth with a Wikipedia entry who were alive in a given year. To stress the 200-yearlong periodic oscillations between 500 BC and 1000 AD, a smoothing filter is shown in blue.

service that allows for server-side execution of a SPARQL query. We pose queries of the following form:

```
SELECT (COUNT(*) AS ?count)
WHERE {
 ?item is HUMAN;
   has property DATE_OF_BIRTH;
   has property DATE_OF_DEATH;

 FILTER EXISTS {WIKIPEDIA_ARTICLE about ?item}
 FILTER (year < Year(DATE_OF_BIRTH) < year + 1)
}
```

and thereby obtain the count of people listed on Wikidata having a recorded time of birth, time of death, who were born between year and year + 1 and who have a Wikipedia article in any language. For readability, property IDs have been substituted with their human-intelligible equivalent and query optimisation was omitted. We do the analogous query to find out how many people died between year and year + 1. Existence of birth and death dates ensures a reasonable lower bound on notoriety of the people included in our analysis, but also—more importantly—renders it possible to compute their lifespan. Our queries begin with the year 4000 BC. We then subtract the deaths from the births for each year and sum over the result, giving us the "population of Wikipedia" as it varies over time. Divided through interpolated estimates of the world's population, this gives a fraction of the world population with a Wikipedia article, which can be interpreted as a proxy for the fraction of notorious people if we follow Arbesman. The world population from before 1940 AD is referenced from the History Database of the Global Environment (HYDE) of Goldewijk et al. (2011) and from the United Nations Department of Economic and Social Affairs (2022) for years 1941 until 2022. Figure 1 illustrates the results. We repeat this process for the *A Brief History of Human Time* database, as shown in Figure 2.

This method of posing direct queries can be used to find the distribution of births on months and days of the week. We append the query above with an additional condition

```
WHERE {
 FILTER (month < Month(DATE_OF_BIRTH) < month + 1)
}
```

and additionally with the weekday. In both calculations, the first day of every year was omitted because some dates are only given to the year and consequently listed as the first of January. Averages instead of absolute numbers were used to take into account the differing number of days per month.

For geospatial data, we use the *Brief History of Human Time* data set and continue their considerations (Laouenan et al. 2022, 11–13). We give a world map detailing the birth places by area in Figure 4.

3. Findings

3.1 Millenia, centuries, decades and years

The directly queried absolute births and deaths, visualised in the top row in Figure 1, exhibit an increase from 4000 BC until 1930 AD and 2020 AD respectively, that is on the whole exponential. Since we are collecting the numbers for individuals who already have a death date entered and thus many living people not listed in the query, this is to be expected. It is certainly a worthwhile matter for speculation how this curve might look in the future. Birth and Death rates overtake each other many times, until a period from around 1500 to 1939 AD where the birth rate is consistently higher, and the data set thus grows. This is in part due to the fact that discrete spikes can be observed at full centuries that suddenly become less steep around 1500, which is discussed later in section [Precision]. Infamous wars, famines and epidemics including even the Great Plague in Europe around 1350 are not directly visible, contrary to our expectation. The only exceptions are the two World Wars; the first induces a noticeable drop in births of Wikipedians and a slight increase in deaths, most probably connected to the Great Influenza epidemic. The second World War has an even larger toll; a truly frightening effect not only of causalities, but likely also bombings of civilians, famines and (other) ideologically motivated genocide.

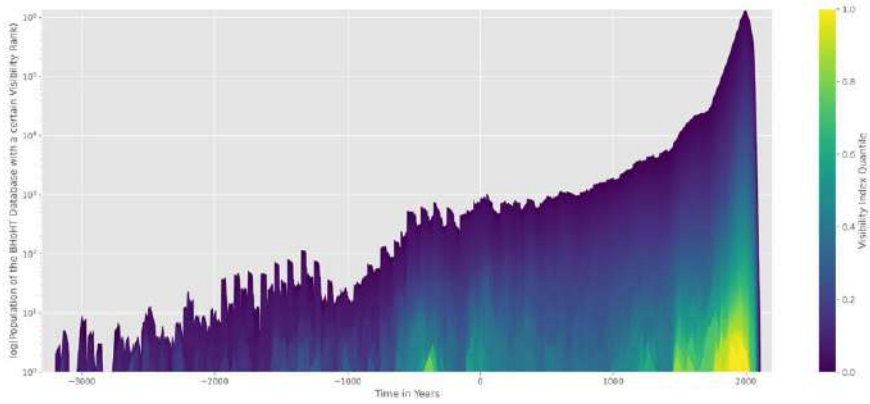


Figure 2 – Fraction of the world population listed in the Brief History of Human Time data base on a semi-log scale, coloured by notability index quantile also on a log-scale. The notability index is a measure introduced and calculated by Laouenan et al. (2022) that, for every individual, compiles five values — the number of articles in different language editions, the cumulative length of these articles, the amount of attributes listed in their Wikidata entry, the number of external links from Wikidata and how often the associated biographies were clicked from 2015 to 2018—into a single number indicative of the person's notability. 15 highlighted lines were drawn at logarithmically spaced quantiles to aid reading.

Now taking the difference of the two rates and running a cumulative sum, we arrive at a sort of "population" of Wikipedia. We now divide by the world population to get the fraction of individuals with a Wikipedia article including birth and death date alive at a given time. An overall exponential increase from a single individual in 4000 BC to a peak at 1940 AD of just over 106 people takes place in absolute numbers, as a fraction from 10^{-7} to $6 \cdot 10^{-4}$ of the total population of earth. However, the growth is stepped through in multiple phases and certainly not a strict exponential or logistic curve. Most strikingly, we can observe a repeated cycle of a rapid expansion of the population followed by a stagnation as if pieced together by separate logistic curves. The most prominent explosion is at around 600 BC, a smaller one at 100 BC, then oscillations with a period of around 200 years at $7.5 \cdot 10^{-7}$ until 900 AD with multiple ups and downs. The next spurt begins at 1000 AD, followed by another logistic phase from 1500 to around 1740 AD, where the last seemingly exponential growth spurt takes the population to its peak in 1930. Here, the fraction is at $4.7 \cdot 10^{-4}$, which is slightly above Arbesman's upper value for 2013 ($4.1 \cdot 10^{-4}$) in part because Wikipedia has grown significantly since his analysis. Simply repeating his lower bound calculation of dividing the number of people in the "living people" category of the English Wikipedia (= 1062985 at time of writing) by the number of people alive ($8.03 \cdot 109$) yields $1.329 \cdot 10^{-4}$, compared to his lower value of $8.6 \cdot 10^{-5}$.

The fraction of humanity in the *Brief History of Human Time* data set obtained with the same calculation (Figure 2) is qualitatively similar and peaks at $2.3 \cdot 10^{-4}$ in 1990 AD.

3.2 Months and weeks

Restricting ourselves to the number of people who enter Wikipedia in the years 1900 to 1920 daily, where the yearly birthrates are roughly constant, we find a significant dependence on the month of birth. It is very curious indeed that February and March are overrepresented while June and July are

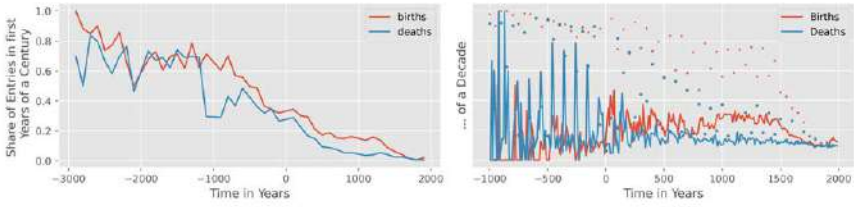


Figure 3 – Fraction of entries in Wikidata with an associated Wikipedia entry in any language that have a birth date or death date listed in the first year of its century (left) or decade (right). On the right, the first decade of each half-century are shown separately as a scatter plot, because some dates are listed to be known up to a half-century. If all entries had exact years associated with them, the fraction would trail around 0.01 on the left and around 0.1 on the right.

underrepresented, because most entries in Wikipedia are people from the Northern hemisphere. On it, the more common birth months are the exact opposite, as for instance Martinez-Bakker et al. (2014) demonstrate. In contrast, we cannot report any significant findings in regard to the weekday of birth.

3.3 Temporal precision

Naturally, we know less exact dates about individuals who have lived further in the past. In the top left of Figure 1 for instance, there are periodic peaks of yearly births and deaths at full centuries, indicating that they were only known to the editor up to the century—as is for example the case for pharaoh Ramses I. The same is true for decades. Quantifying this effect, we give a plot of the fraction of entries with birth and death dates in the first years of centuries or decades in Figure 3. We can observe the respective fractions decrease from around 1 to their expected values for fully precise biographical data of 0.1 for decades and 0.01 for centuries, the latter in a surprisingly linear fashion — over the course of 5000 years. As anticipated, death dates are consistently listed to higher precision than birth dates across all time. A discontinuity can be observed around 1500, with the precision of birth dates suddenly improving quickly after a plateau since 700 AD. This could possibly be explained by the invention and rapid adoption of the printing press in Europe, where many of the recorded individuals in this time frame are from. The effect can also be observed quite distinctly in the smoothing of the resulting population curve, the row of Figure 1.

4. Location by year of birth

The map showing the places of birth classified by era, suggests the cradle of written history in Mesopotamia, Egypt, Palestine, Israel and Greece. The distribution then moves to China and Rome before spreading quickly in the Middle Ages — which is perhaps unexpected — to Central Europe, West Europe and Japan. The Enlightenment era carries the people of Wikipedia across East Europe and Americas. India, Central Africa and South Africa also start a substantial contribution to the data shortly after. Finally, we can see the humanity colour in even remote locations like islands and the polar regions. Some eras of cultural bloom immediately stick out as suspiciously undocumented, such as the entirety of Imperial China after the Three Kingdoms era. This is surprising because it sustained the largest population of any country throughout most of history, including of course the time leading up to the decade before our own. Inhabitants of the People's Republic of China are largely barred from Wikipedia, but others have not filled the gap. Even China's modern history as an increasingly industrially successful country has left the Wikipedia editors unimpressed. The same, to a slightly lesser degree, can be said of India from the ancient dynasties to the Maratha reign and from the Mughal empire to today. Although the world population at times comprised mostly of subjects of China and India, they remain invisible. Other seemingly underrepresented

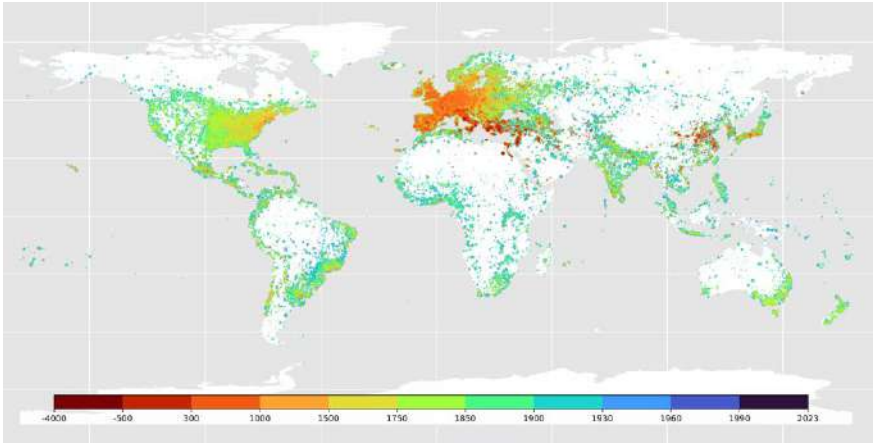


Figure 4 – This map is marked with approximately 2.2 million dots, each for one person in the cross-verified version of the Brief History of Human Time data set at the location of their birth. The colour specifies a historical period in which the person was born, according to the year as classified on the bottom bar. Recent periods are printed first, enabling us to see the relatively few people who were born in ancient times over the exponentially more people who entered the data base in newer times. The dots each have an opacity of 0.2 and a radius proportional to the number of the corresponding person's number of Wikipedia editions.

dominions include the Kingdom of Morocco, the Ottoman Empire, the Incas and the Azteks.

On the other hand, we can see an impressive share of the data set popping up in certain regions despite a comparatively negligible population. The most extreme examples are Medieval Europe and Great Britain, but also certainly also the young the United States of America, considering its 5 million inhabitants around 1800 made up merely 0.5% of the world population (Gauthier 2002, Appendix A).

5. Gender

Laouenan et al. (2022) give a comprehensive analysis regarding the fraction of women in their database, so we need not linger too long on this particular point. Certainly worth a second look however is the distribution of non-binary individuals, who are, to clarify, not people with a gender that is not recorded or unknown, but deliberately registered as so. Simply by the numbers, it is not surprising that the entries are few and often located in today's urban centres of industrial societies, especially in Anglophone countries where the availability of neutral pronouns might increase visibility and biographical accuracy. We also want to take a mo-

ment to recognise that not all dots indicating such a person stems from the modern era, such as for example the Public Universal Friend, an American preacher born in 1776 (Larson 2014, 576–600). We see at the same time a possibility for data-driven research to find or to drown out suppressed identities in this ocean of 2.2 million people. Non-binary genders that have been recognised as such since earlier times seem not to have found much of a representation. We find, for example, only 4 people listed as Hijra in Wikidata, with none being born before 1900.

6. Language

We investigate the editorial interest in different language communities of the large encyclopaedias, by counting the number of articles written in one language domain about different countries. Figure 5 shows the correlation between the country of origin of the person described by entries in Wikipedia with the language edition the article is written in. We first notice the overwhelming number of articles about individuals from the United States, as also seen in Figure 4. Some other countries also appear as being represented particularly often, such as Germany, France, Italy and Japan. Next, a strong home bias becomes apparent: Italophones write about Italians, Grecophones about the Greek and especially Germanophones about each other. The geographically close countries also tend to show a mutual affinity, such as Romance languages with Ibero-American and Mediterranean countries or Slavic languages with Slavic countries. This extends beyond language barriers also, as Japan with its popularity among Asian neighbours demonstrates. Some Wikipedia editions, such as the Vietnamese, stick out as particularly diverse and cover many other countries' individuals. It is sobering that more mature editions of Wikipedia (i.e. English, German, French) are not more diverse than less edited versions (i.e. Vietnamese, Hungarian, Finnish). Digital globalisation has not in itself been enough to dissolve the language barriers between us.

7. Discussion

We would now like to attempt the jump from our particular data to the historical record at large and therefore need to discuss the systematic issues when using Wikidata as the basis for a historical analysis beforehand. An issue that remains lingering in the background throughout is the data quality of our source Wikidata. Apart from missing information, which can be pruned programmatically, it is also clear that false information does manage to stay in open encyclopaedias for longer than in counterparts written by experts. However, removing conflicts between two sources, as *A Brief History of Human Time* does, already eliminates a large part of these errors. Wikidata keeps track of multiple data sources as available, enabling more sophisticated mitigation in principle, although we did not use this feature for our analysis.

While we consider all language editions and can thereby reduce regional bias significantly, we acknowledge that a disproportionate fraction of Wikipedia

editors is from regions where English, German, Russian or Japanese is spoken, creating an immediate bias to toward the written record in these languages. The English language with its many second-language speakers might further attract editors to primary sources in English, which in turn disproportionally describe Anglophone people. There may exist large parts of recorded history that has not found its way into the database because of this bias, for instance the Chinese cultural heritage we earlier noted to be largely missing.

Overall, we are optimistic that at least the Western historical record has been translated quite comprehensively into this very large data set. The data on reader interest on the other hand, which we used to infer interest into historical figures might be less representative. It seems highly likely that the people accessing an encyclopaedia might seek out more information about historical figures than they would do at other times, simply because information about contemporaries or the recently deceased is found in other sources like magazines or social media. People who are in the habit of using Wikipedia are younger, more highly educated, disproportionally male and wealthy, as user demographics show (Rainie and Tancer 2007). We can perhaps find a parallel to the historical figures themselves and conjecture that historians, editors and finally readers are drawn towards people which they identify with through shared traits. We have no direct data to support this though; a follow-up investigation on this point would certainly be interesting. This also links back to our findings about in-group language preference for editors as seen in Figure 5.

As we continue to uncover the layers of identity in Wikipedia's population, we inevitably run into the question of why the historical record has handed us their biographies rather than the other 117 billion humans who have been alive, 62 billion of which after 0 AD. This question then splits in two as we shift our focus towards the present; who can we keep records about and who do we foster a large enough interest in to do so? We have long reached the technological capabilities to publish an encyclopaedic entry or even a whole biography about nearly everybody, but of course choose not to. We may infer a certain set of criteria that a society needs to produce such documents and further that only a small percentage will fulfil this criterion. But is this percentage increasing over time, as our investigation might suggest? Or is our memory of the past degrading and hiding people who would have been deemed more than worthy of being recording by their contemporaries? These are the two possibilities that would explain the circumstance that for our approximation of "in the historical record," the fraction rises from 10^{-7} in 4000 BC to more than $2.3 \cdot 10^{-4}$ in 1990 AD. The difference between those two hypotheses is nicely illustrated by the life of Menocchio, a miller from North Italy who was burned at the stake by the Inquisition after being accused of heresy for his unorthodox religious views. As discovered and wonderfully recounted by Ginzburg (2002) on the basis of unusually comprehensive court records from the trial in 1599, his biography is fascinating and inspiring for us century citizens and can teach us more than most contemporary biographies. And yet, it is clear that only a series of highly improbable coincidences is the vector of our knowledge about him; that for every biography

like this, a thousand others have not found their way to us. It could therefore be useful to actively look for gaps in our knowledge, for example through quantitative means, and come to empirically backed conclusions where our blind spots lie. We are compelled to take a position on the extreme opposite to Great Men theory; who among us does not have something to contribute to the human story that is worthy of study? Why should that have been different in the past? On the other hand, it is difficult to look for what we do not know, reassuring us that diligent historians and philologists will always be needed to shed light on our past. We are not at all in a position to demand artificial attention to certain groups and instead find it interesting to simply observe where attention is paid.

Let us then venture to make some more speculative statements and find overarching patterns in our evaluation. Firstly, the amount of people who we describe as being in the historical record increases much faster than the population. There is no particularly simple model, but the growing exponential is a first good approximation as with the world population itself. However, even the exponential declines too slow in a sense—not only are we forgetting our past but also increasingly storing more information, leading to more available records. Readers are primarily interested in reading and writing the history of our contemporaries and the recently deceased. However, it cannot be at all concluded that this interest also fades off exponentially as we go back into the past. (To stress again, Western) readers collectively take a special interest in certain periods that partly counteracts the scarcity of information. We may know few people born around 0 AD, but the ones that are known are more popular. One could see an opportunity here to dispel a narrative about the Medieval era, at least for the globe taken as a whole. Instead of the number of articles declining, it is simply the number of reads that give the appearance of "dark ages". This low-supply-high-demand effect also appears for different genders. We know fewer women and non-binary people, but they consistently receive more interest from readers if they do appear. We may take this to mean that marginalised people need to accomplish something more noteworthy to be passed on as part of the written record, or that readers looking for representation of their own identities hold on to the few figures they find. It could also simply mean that we are drawn to stories of people who accomplished something big despite their circumstances. This also plays into our surprising result that months where few births are expected on the contrary show a larger proportion of Wikipedians: People placed in slightly more isolated, unlikely circumstances will more often impress us. Perhaps they were able to receive more attention from their family and community, in this case. We have not spoken much about socioeconomic factors much, as we found no reliable way to track them, but can rely on Lauenan et al.'s findings on the different occupations in their database to see clearly that most members of it are influential and wealthy. In earlier times, we see many statesmen and religious leaders, then giving way to aristocracy and later scholars as well as artists. In recent times, the most common occupations are athletes and entertainers. Very telling in our findings, we could not observe any effect of various humanitarian disasters. We interpret this as, on the one hand,

the influential classes of society not being affected by these events so much, but also a slight confirmation of the hypothesis that the fraction of people deemed worthy to be passed on does not change by much.

8. Conclusion

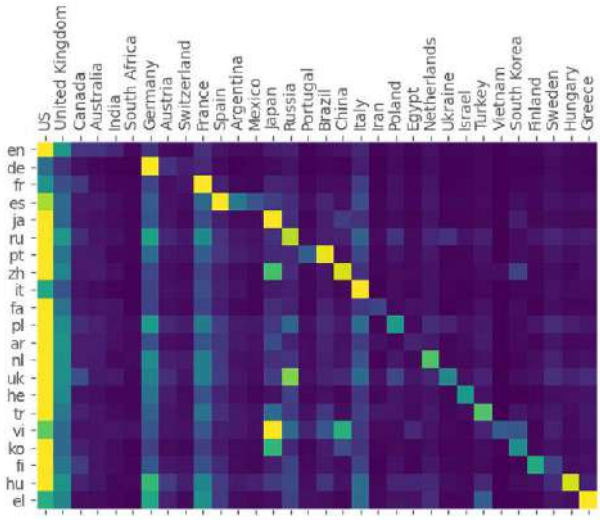


Figure 5 – Each row in the image matrix corresponds to one language edition of Wikipedia as indicated by ISO 639-1 codes on the left, each column to a country or dominion. One cell shows the number of entries in A Brief History of Human Time about a person born in the associated country, that have an article in the associated language. Rows are L1-normalised, so the sum over each row is the same. This means that if all language editions were interested in countries equally, the rows would look more or less identical even for differently sized editions. The values are then linearly put into a perceptually uniform colour map. The order of the rows is by date of founding for the language edition, to which the countries were then aligned by their official and commonly spoken languages.

Summarising the entire historical record in a unified, accurate and unbiased dataset is impossible. Nonetheless, we have looked at and assembled ourselves large collections of biographical encyclopaedic knowledge that try to be such a summary—and found them succeed to a meaningful degree. Comparing them to our perspective on human history, we have also described their shortcomings and who they might be overlooking, which gave rise to a bigger picture of what we know or, more importantly, care about in the past's people. The numbers reveal our collective focal points—Rome, London, Boston, Kyōto—and ask us to understand how much of the effect is fascination and how much fixation. The

profile of our sample population is skewed in many ways; here we see a challenge and an opportunity for intersectional history. A potential for linguistic communities to better understand each other could be demonstrated, opening the way for collective gaps in understanding to be filled. We have tracked the accuracy of historical data improving throughout the decades. We further had a closer look at the dynamics of the world's growing collection of biographical stories over time, revealing one the one hand general trends of withering memory but also of ever more comprehensive information collected.

We cannot help but wonder how our investigation will look in a few decades: Will we see Wikipedia-like databases for the eras to come that represent the world of today more accurately? We conjecture that it is indeed so. As opposed to certain other aspects of modern life, like wealth, we can see the opposite effect of an increasingly even spread of attention over a broader range of people as we go further in time. We excitedly look forward to further historical research that discovers the remaining individuals who were not able to enter our investigation because they have been forgotten and hope that our work is useful in illustrating the need to do so. Certainly, the effort to engineer yet better datasets out of these findings is not yet at an end, meaning that in the future a more detailed and fine-grained analysis will become possible. We hope that this will give not only the Caesars, but also the Menocchios and Public Universal Friends of history a voice that reaches us.

References

Arbesman, Samuel. 2013. "The Fraction of Famous People in the World." *Wired*, January 2013. <https://www.wired.com/2013/01/the-fraction-of-famous-people-in-the-world>.

Gauthier, Jason G. 2002. *Measuring America: The Decennial Censuses from 1790 to 2000*. Washington D.C.: US Department of Commerce, Economics and Statistics Administration.

Ginzburg, Carlo. 2002. *Der Käse und die Würmer: die Welt eines Müllers um 1600*. Berlin: Wagenbach. (*Il formaggio e i vermi*. Milano: Adelphi, 1976).

Klein Goldewijk, Kees, Arthur Beusen, Gerard van Drecht, and Martine de Vos. 2011. "The HYDE 3.1 Spatially Explicit Database of Human-Induced Global Land-Use Change over the Past 12,000 Years." *Global Ecology and Biogeography* 20 (1).

Laouenan, Morgane, Palaash Bhargava, Jean-Benoît Eymeoud, Olivier Gergaud, Guillaume Plique, and Etienne Wasmer. 2022. "A Cross-Verified Database of Notable People, 3500BC-2018AD." *Scientific Data* 9 (1).

Larson, Scott. 2014. "Indescribable Being: Theological Performances of Genderlessness in the Society of the Publick Universal Friend, 1776–1819." *Early American Studies*: 576–600.

Martinez-Bakker, Micaela, Kevin M. Bakker, Aaron A. King, and Pejman Rohani. 2014. "Human Birth Seasonality: Latitudinal Gradient and Interplay with Childhood Disease Dynamics." *Proceedings of the Royal Society B: Biological Sciences* 281 (1783).

Murray, Charles. 2003. *Human Accomplishment: The Pursuit of Excellence in the Arts and Sciences, 800 BC to 1950*. New York: HarperCollins.

Rainie, Lee, and Bill Tancer. 2007. *36% of Online American Adults Consult Wikipedia*. Washington D.C.: Pew Internet & American Life Project. <https://www.pewresearch.org/internet/2007/04/24/wikipedia-users>.

Schich, Maximilian, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási, and Dirk Helbing. 2014. "A Network Framework of Cultural History." *Science* 345 (6196): 558-62.

United Nations, Department of Economic and Social Affairs, Population Division. 2022. *World Population Prospects 2022, Online Edition*. <https://population.un.org>.

Zhao, Yu Amy, Shahar Ronen, Kevin Hu, Tiffany Lu, and César A. Hidalgo. 2016. "Pantheon 1.0, a Manually Verified Dataset of Globally Famous Biographies." *Scientific Data* 3 (1): 1–16.

Automatic Verification of References of Wikidata Statements

Elena Simperl, Odinaldo Rodrigues, Albert Meroño-Peñuela, Kholoud Saad Alghamdi, Gabriel Maia Rocha Amaral, Nathan Schneider Gavenski, Jongmo Kim, Miriam Redi, Yiwen Xing, Bohui Zhang, Yihang Zhao

Abstract: Wikidata is one of the world's most important data assets. It is used by search engines, virtual assistants, fact checkers, and in over 800 Wikimedia projects. Wikidata contains 1.65 billion statements about over 116 million data items, edited by nearly 25 thousand editors. Manually checking whether Wikidata statements are supported by references is a slow process that does not scale with size. Given the overall number of statements to check, the collaborative nature of Wikidata, and the fact that referenced documents can change over time, preserving the quality of the references is an onerous process requiring continuous intervention. This paper present ProVe – a tool to assist in the automatic verification and assessment of the quality of the references of Wikidata items.

Keywords: Wikidata, quality assurance, reference verification, artificial intelligence, computational tools.

1. Introduction

A Knowledge Graph (KG) is a large network of interconnected entities, encoding their properties and relationships to one another (Krötzsch and Weikum 2016; Paulheim 2016). KGs serve as sources of machine-readable and semantically structured data used by several web applications, including Wikipedia

Elena Simperl, King's College London, United Kingdom, elena.simperl@kcl.ac.uk, 0000-0003-1722-947X
Odinaldo Rodrigues, King's College London, United Kingdom, odinaldo.rodrigues@kcl.ac.uk, 0000-0001-7823-1034
Albert Meroño-Peñuela, King's College London, United Kingdom, albert.merono@kcl.ac.uk, 0000-0003-4646-5842
Kholoud Saad Alghamdi, University of Jeddah, Saudi Arabia, kholoud.alghamdi@kcl.ac.uk, 0000-0003-0261-9977
Gabriel Maia Rocha Amaral, King's College London, United Kingdom, gabriel.amaral@kcl.ac.uk, 0000-0002-4482-5376
Nathan Schneider Gavenski, King's College London, United Kingdom, nathan.schneider_gavenski@kcl.ac.uk, 0000-0003-0578-3086
Jongmo Kim, King's College London, United Kingdom, jongmo.kim@kcl.ac.uk, 0000-0002-4984-1674
Miriam Redi, Wikimedia Foundation, mredi@wikimedia.org,
Yiwen Xing, Oxford e-Research Centre, United Kingdom, yiwen.xing@eng.ox.ac.uk, 0000-0003-1521-6616
Bohui Zhang, King's College London, United Kingdom, bohui.zhang@kcl.ac.uk, 0000-0001-5430-1624
Yihang Zhao, King's College London, United Kingdom, yihang.zhao@kcl.ac.uk, 0009-0009-2436-8145

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Elena Simperl, Odinaldo Rodrigues, Albert Meroño-Peñuela, Kholoud Saad Alghamdi, Gabriel Maia Rocha Amaral, Nathan Schneider Gavenski, Jongmo Kim, Miriam Redi, Yiwen Xing, Bohui Zhang, Yihang Zhao, *Automatic Verification of References of Wikidata Statements,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.31, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 263-271, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

infoboxes, search engines, and voice-activated assistants, amongst others (Ji et al. 2022; Malyshev et al. 2018). In most KGs, information is stored as a set of statements, semantic triples of the form ⟨*subject, predicate, object*⟩, denoting a property of the subject in the triple (Färber et al. 2018). Ensuring KGs are trustworthy depends on well-documented and verifiable provenance to the information they encode (Zaveri et al. 2016). For example, if the statement "beer is bitter" is encoded in a KG by the triple ⟨*beer, has characteristic, bitterness*⟩, this triple should reference a source which supports the statement.

Mechanisms that help to evaluate and ensure the quality of information provenance are thus crucial to the verifiability of KGs (Zaveri et al. 2016; Piscopo et al. 2017; Wang et al. 2021). However, such processes are currently mostly performed manually (McAndrew and Strathmann 2021) and do not scale with size. Yet, on vital KGs such as Wikidata and DBpedia, manual verification is prohibitive due to their sheer size (Piscopo et al. 2017)—Wikidata has currently over 1.65 billion statements—and more support to assist with verification is needed.

ProVe (Provenance Verification) leverages research conducted for and evaluated by the Wikidata community, responding to their data assurance needs (Amaral et al. 2021; Amaral, Rodrigues and Simperl 2022; Amaral, Rodrigues, and Simperl 2023). It consists of an add-on user interface embedded in Wikidata's editing pages (a Wikidata gadget) and web API, which use Natural Language Processing (NLP) models, public datasets on data verbalisation and fact verification, and rule-based methods.

Given a Wikidata statement and an external URL reference (e.g., through the P854 property or an external identifier property that has formattable URLs), ProVe verbalises the statement, automatically retrieves the referenced document and looks for the passages within it that are relevant to the statement's claim. ProVe then evaluates the overall stance of the referenced document with respect to the statement, displaying the stance along with the most relevant passage found as justification. This process is repeated for all statements whose subject is the given item, with the results summarised in a convenient, unobtrusive table.

ProVe is work in progress, but we have an MVP whose functionality is described in this paper, along with important technical details about the reference evaluation process. The paper then concludes with a discussion of limitations and plans for future work.

2. Related work

FEVER (Fact Extraction and VERification) (Thorne et al. 2018) is a large dataset containing over 185K claims, that can be used for fact verification against textual sources. Despite its general applicability, claim verification in KGs faces additional challenges because statements need to be turned into sentences first. This "verbalisation" process needs to consider important assumptions about the way information is represented in the KG (more on this later). ProVe employs FEVER for some natural language tasks (see the Reference Verification Process section).

Several works have previously focused on the verification of the quality of information in KGs (McAndrew and Strathmann 2021; Piscopo et al. 2017; Wang et al. 2021; Zaveri et al. 2016; Amaral, Rodrigues, and Simperl 2023; Amaral, Rodrigues, and Simperl 2022; Amaral et al. 2021). However, to the best of our knowledge, ProVe is the only available tool integrated within Wikidata and based on published research that *automates* the reference verification process. ProVe has been running since Summer 2024, even though we only started collecting detailed usage information since March 2025. Our statistics show we have been receiving around 400 requests a day, coming from 22 countries and all continents.

3. General overview

Figure 1 gives an overview of ProVe's architecture. It consists of the main server that processes user requests submitted from the gadget or API; ML models within it used for some natural language tasks; a web service and API which can be used in external applications; a Wikidata *gadget* that interacts with editors and summarises the assessment of references of individual items; and the main database storing the results of evaluations for future analyses.

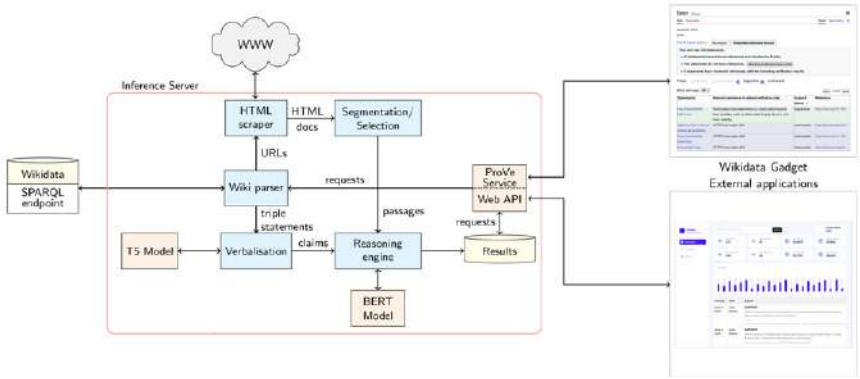


Figure 1 – Overview of ProVe's system architecture.

4. Wikidata gadget

Most users can easily interact with ProVe via its *gadget*, which can be installed according to the instructions found on Wikidata's project page for the tool. Once installed, the gadget works within Wikidata's item editing page, systematically checking externally referenced statements. For each statement-reference pair, ProVe will then display the computed support stance of the referenced document along the most relevant passage found within the document. This is all conveniently displayed in a sortable table, with links to the editing sections within the page for each of the statements. All individual statement-reference quality as-

sessments are then combined to give a number in the [-1,1] interval—ProVe's quality score for the item. This is displayed within the Wikidata page (see infobox, bottom right of Figure 1), but can also be retrieved programmatically for analysis outside Wikidata (see below).

5. Web API

The gadget is intended for simple, item-oriented information about the quality of external references of statements. For programmatic applications that need information about multiple items at specific points in time, ProVe provides aWeb API[1] too. Through the API, external applications can check whether the references of items have been previously analysed, request them to be (re-)evaluated, obtain summarised and comprehensive evaluation results, as well as historical evaluation data.

6. Reference verification process

The general workflow of ProVe's reference verification process is shown in Figure 2. For each statement and associated external reference, ProVe first verbalises the statement (A), then retrieves the text of the referenced document, removing markup elements and segmenting it into passages (B). Passages are then ranked according to their relevance to the verbalised statement (C), and the support stance of the most relevant passages with respect to the statement analysed. The overall support stance of the referenced document for the statement is computed and provided along with the evidence found (D). This is typically the passage within the document found to be the most relevant to the verbalisation of the statement. Finally, the results for each statement-reference pair are aggregated to give the user the item's ProVe score (E).

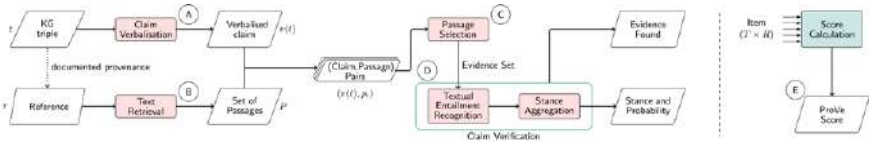


Figure 2 – Overview of the reference quality evaluation process.

Further details of each step are given below.

7. Verbalisation

Although the verbalisation of a statement such as ⟨*beer, has characteristic, bitterness*⟩ is arguably simple, in general there are several hurdles to overcome. Firstly, components of a statement <*subject, predicate, object*> are provided with

[1] <https://kclwqt.sites.er.kcl.ac.uk/apidocs/>.

a set of alternative labels describing the component, and a judicious choice for a suitable label needs to be made. Secondly, there are implicit assumptions about the roles played by the *subject* and *object* of the relationship, as well as what the relationship is used for. For example, in the triple <*william, child, george*> the intended meaning is that the subject of the statement is the parent, and its object is the offspring. Although just conventions, these assumptions need to be considered to produce faithful verbalisations. To generate verbalisations that are fluent and resemble natural text, ProVe employs a T5-base model (Raffel et al. 2020) fine-tuned on the WebNLG 2017 dataset (Gardent et al. 2017). As part of original research done to underpin ProVe, a dataset with verbalised Wikidata statements was generated and made publicly available (Amaral, Rodrigues and Simperl 2022).

8. Text retrieval

Layout and other structural information embedded into referenced documents make the extraction and meaningful re-combination of text non-trivial. In addition, the text itself can contain semantical constructs spread across sentences making ad-hoc segmentation not suitable for subsequent entailment evaluation of the passages. ProVe employs several custom rules to transform and remove HTML markup, after which the text is segmented using spaCy's sentence segmenter with the *en_core_web_lb* model (Honnibal et al. 2020). Pairs of consecutive segments are generated to cater for constructs such as pronominal anaphora, and single and combined segments are sent for subsequent passage selection.

9. Passage selection

Once the claim has been verbalised and the referenced text segmented into passages, passages are ranked according to their relevance to the claim (independently of their support stance, which is analysed later). ProVe uses a pre-trained BERT transformer (Soleimani, Monz, and Worring 2020) fine-tuned on the FEVER dataset to give each passage a relevance score in the interval [-1,1]. Since some of the passages overlap (due the combinations described in the previous step), ProVe only keeps the five highest ranked passages that do not overlap along with their relevance scores. The scores are used in the claim verification step below.

10. Claim verification

Evaluating the support stance of the referenced document for the statement as whole is performed in two stages. First, the stance of each of the most relevant passages is evaluated by a pre-trained BERT model (also fine-tuned on FEVER) for recognising textual entailment yielding a probability distribution for the classes *supportive, refuting,* and *not enough information* (i.e., inconclusive). At a

second stage, the relevance scores of all passages selected along with their support stance probability distributions are aggregated to give an overall strength and support stance of the document for the statement.

11. ProVe score (E)

Evaluating the support of individual statements by their external references is critical, but for editors working on items, an overall indication of the support for the item is also very important. Since a Wikidata item T can appear as the subject of many statements, each of which can have many references, the support stances of the item's statements need to be aggregated. This is done as follows. Let S(T)=[s1,…,sn] be the support stances for all statement-reference pairs of the item T calculated as described in the claim verification step above, where the value si (i=1,…,n) is either -1 for refuting references, 0 (for inconclusive), or 1 (for supportive)[2]. The ProVe score for item T, in symbols PS(T), is calculated as the sum of values in S(T) divided by n.

It should be easy to see that PS(T) is a value in [–1,1] with the following intended meaning. Positive values indicate that the number of supporting references surpasses the number of refuting references and negative values indicate the opposite. In either case, inconclusive references bring the score closer to 0. As a result, proximity to 1 is associated with "good" quality of the references; proximity to 0 is associated with inconclusive or missing references; and proximity to –1 indicates high levels of disparity between claims and references. Of course, there is scope to provide customised scores that consider different types of statements and references.

The ProVe score of an item along with a summary of the total of references in each stance category is shown to editors when the Wikidata page for the item is loaded (see "infobox" at the bottom right of Figure 1). A summary table with sortable columns and filtering capability also provides convenient links to statements for potential correction of references (Fig. 3).

Users can request the computation or re-computation of scores by pressing appropriate buttons in the interface. The scores allow users to factor in reference information in the prioritisation of items to edit and to perform custom analyses with the extra information provided via the web API, e.g., the progress achieved in reference quality improvement of specific items.

12. Limitations

ProVe can take any non-ontological KG triple as long as its components are accompanied by labels in English. This requirement is because the NLP modules so far have only been trained for English. Extensions to other languages

[2] For completeness, ProVe classifies irretrievable references as inconclusive since it cannot evaluate their stances w.r.t. their claims.

are under consideration. ProVe focuses on *external* references only, since these are arguably harder to verify. In addition, visual elements, such as pictures and charts, can serve as evidence for KG statements. However, the automated extraction of text from these types of evidence in a format that language models can understand is not trivial and ProVe cannot deal with them yet. ProVe employs a combination of techniques (including pre-trained sentence segmenters) to extract passages from various forms of structured text, e.g., within tables, but this is not always effective. Finally, we are working towards parsing PDF documents too, but this is not yet operational.

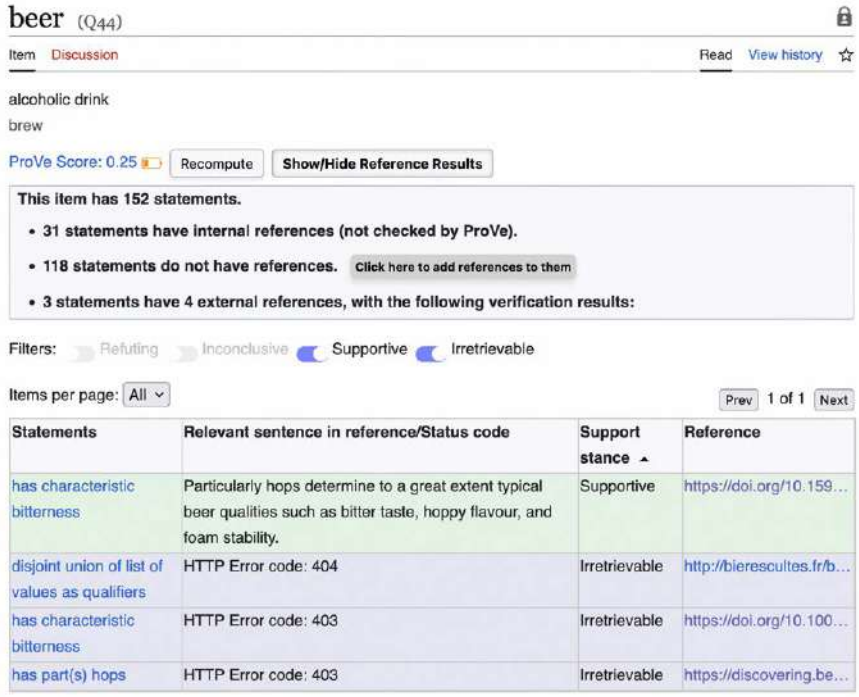


Figure 3 – ProVe's main user interface in Wikidata.

13. Community involvement

We welcome suggestions of the community for improvements to ProVe and the development of new functionality. Users can register interest by adding their names to the tool's participants list[3].

[3] <https://www.wikidata.org/wiki/Wikidata:WikiProject_Reference_Verification#Participants>.

14. Discussion and future work

ProVe is limited to the verification of statements presented as triples, not being directly applicable to actual sentences in natural language. Hence, it cannot be used in applications such as Wikipedia. However, it is technically possible to skip the verbalisation phase (step (A) in Figure 2), replace the verbalised statement by a sentence of interest, and continue the evaluation process from there. For this to be effective, we would need a mechanism to properly extract from the input text the sentence we would like to verify and ensure the NLP model(s) employed in the relevance and entailment tasks are suitable for the domain of the text. This is left as a potential future extension.

We have identified particularly challenging types of references that would benefit from special treatment in the text retrieval phase (step (B) in Figure 2). Alternative mechanisms for text segmentation for use in subsequent textual entailment recognition are under consideration. Finally, we would like to make available to end users more of the information obtained during the evaluation process. This is more easily done by modifying the end points of the Web API, although further tweaks to the Wikidata gadget are also possible.

References

Amaral, Gabriel, Alessandro Piscopo, Lucie-Aimée Kaffee, Odinaldo Rodrigues, and Elena Simperl. 2021. "Assessing the Quality of Sources in Wikidata Across Languages: A Hybrid Approach." *Journal of Data and Information Quality* 13 (4): 1–35.

Amaral, Gabriel, Odinaldo Rodrigues, and Elena Simperl. 2022. "WDV: A Broad Data Verbalisation Dataset Built from Wikidata." In *The Semantic Web – ISWC 2022*, edited by Ulrike Sattler, Aidan Hogan, Maria Keet, Valentina Presutti, João Paulo A. Almeida, Hideaki Takeda, Pierre Monnin, Giuseppe Pirrò, and Claudia d'Amato, 556–74. Cham: Springer International Publishing.

Amaral, Gabriel, Odinaldo Rodrigues, and Elena Simperl. 2023. "ProVe: A Pipeline for Automated Provenance Verification of Knowledge Graphs against Textual Sources." *Semantic Web*. https://doi.org/10.3233/SW-233467.

Färber, Michael, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. 2018. "Linked Data Quality of DBpedia, Freebase, Opencyc, Wikidata, and Yago." *Semantic Web* 9 (1): 77–129.

Gardent, Claire, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. "The WebNLG Challenge: Generating Text from RDF Data." In *Proceedings of the 10th International Conference on Natural Language Generation*, 124–33.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. "SpaCy: Industrial-Strength Natural Language Processing in Python." https://doi.org/10.5281/zenodo.1212303.

Ji, Shaoxiong, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications." *IEEE Transactions on Neural Networks and Learning Systems* 33 (2): 494–514. https://doi.org/10.1109/TNNLS.2021.3070843.

Krötzsch, Markus, and Gerhard Weikum. 2016. "JWS Special Issue on Knowledge Graphs." *Journal of Web Semantics*. International Center for Computational Logic. https://iccl.inf.tu-dresden.de/web/JWS_special_issue_on_Knowledge_Graphs/en.

Malyshev, Stanislav, Markus Krötzsch, Larry González, Julius Gonsior, and Adrian Bielefeldt. 2018. "Getting the Most out of Wikidata: Semantic Technology Usage in Wikipedia's Knowledge Graph." In *International Semantic Web Conference*, 376–94.

McAndrew, Ewan, and Clea Strathmann. 2021. "Quality Assurance and Reliability." *Wikidata Quality Assurance and Reliability*. Edinburgh: The University of Edinburgh. https://www.ed.ac.uk/information-services/help-consultancy/is-skills/wikimedia/wikidata/quality-assurance-and-reliability.

Paulheim, Heiko. 2016. "Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods." *Semantic Web* 8: 489–508.

Piscopo, Alessandro, Lucie-Aimée Kaffee, Chris Phethean, and Elena Simperl. 2017. "Provenance Information in a Collaborative Knowledge Graph: An Evaluation of Wikidata External References." In *International Semantic Web Conference*, 542–58.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, et al. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research* 21 (140): 1–67.

Soleimani, Amir, Christof Monz, and Marcel Worring. 2020. "BERT for Evidence Retrieval and Claim Verification." In *European Conference on Information Retrieval*, 359–66.

Thorne, James, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. "FEVER: A Large-Scale Dataset for Fact Extraction and VERification." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 809–19. New Orleans: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1074.

Wang, Xiangyu, Lyuzhou Chen, Taiyu Ban, Muhammad Usman, Yifeng Guan, Shikang Liu, Tianhao Wu, and Huanhuan Chen. 2021. "Knowledge Graph Quality Control: A Survey." *Fundamental Research* 1 (5): 607-26. https://doi.org/10.1016/j.fmre.2021.09.003.

Zaveri, Amrapali, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Soeren Auer. 2016. "Quality Assessment for Linked Data: A Survey." *Semantic Web* 7 (1): 63–93.

PARTE TERZA

Visualizzare e comunicare i dati / Data Visualization and Communication

Developing a Creative Model for Wikidata Analysis in the GLAM Sector

Enrique Tabone

Abstract: In 2019, the author started developing research exploring the visualization of datasets from Wikidata for artistic practice at the University of Salford. Initially, analysing gender representation in collections led to the development of an inquiring model for other art or museum collections. Later, a similar approach was adopted on prehistoric female figurines, held at museums in Malta and Gozo. This brought the work towards a coherent conclusion. The structured datasets are visualized, sonified, and accompanied by her creation of art objects. Reflections on data representations in art collections and museums have provided opportunities to explore concrete ways to look into a collection rather than at a collection. This data science point of view aims to enable the discovery of relationships between items within the dataset while stitching them together through shared properties, in creative ways.

Keywords: Wikidata, data art, data sonification, data visualisation, GLAM.

1. Introduction

This paper presents the initial applications of a creative model for the use of Wikidata to make art. The methodology developed around this research project treats data not only as an exploratory resource but also as a medium for artistic and interpretative exploration. The primary work described builds on collaborative knowledge graphs of structured cultural knowledge that leverage the basic principle of Wikidata, which holds information in ways that make it accessible to both humans and machines.

Within GLAM institutions, this means that artworks, artists, movements, and thematic concepts can be interlinked across collections and national boundaries in ways that go beyond any single dataset. To facilitate a thorough familiarization with Wikidata before attempting to make art with it, it was essential to work on an initial dataset to learn the basics of structured data that can be queried through triplets. The selected dataset was explored to identify systemic gender imbalances within a specific art collection. Using SPARQL queries and visualizations, the first round of data analysis identified underrepresented female and nonbinary artists in the University of Salford's Art Collection, which belongs to the institution where this initial phase of the research was conducted. These types of artists are often absent in institutional narratives predating the contem-

Enrique Tabone, University of Salford, United Kingdom, enriquedesigns@gmail.com, 0009-0004-9584-7497

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Enrique Tabone, *Developing a Creative Model for Wikidata Analysis in the GLAM Sector,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.33, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 275-286, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

porary period, but this is not immediately evident in most art collections. The goal was to interpret the data not only as factual observation but also to enable creative retelling of the collection's narratives.

The importance of interpreting data visualizations is emphasized in this way of working. Visualizations of connections through networks highlight links between artists, institutions, and themes, while geographical mapping uncovers acquisition and origin trends across regions. For example, plotting data about place of birth for artists within the University of Salford's collections revealed regional biases in collecting practices, which underpin the current acquisition policy. Other outputs from the collected data can be presented in timeline visualizations. These prove particularly useful in tracing institutional priorities over time, such as surges in artwork acquisitions from particular regions or styles. This way of working was developed further to curate and connect datasets of university art collections across the UK. The intention of this was to offer accessible pathways for institutions to engage and connect in similar ways. Outputs from this phase included an interactive dashboard, resulting in mapping 99 UK university art collections. These visual interpretations not only aid scholarly analysis but can also serve as accessible tools for public exhibition contexts.

Following this initial exploratory phase into the mechanics of querying structured data through Wikidata, the research was directed towards the inclusion of an artistic approach as an output. The dataset used for this phase of the research project involved Heritage Malta's prehistoric figurine collection, held at two museums in the Maltese islands.

Significant outcomes from this phase of the research included the integration of data visualizations in video art, which was accompanied by a sculpture based on the visualization of this dataset. Another notable outcome was a work of sound art resulting from data sonification in a way that translates symbolic data into an audio composition. This allows for speculative interpretations of cultural meaning embedded in the artifacts. This creative approach demonstrates the possibilities of revitalizing archaeological data and providing multi-sensory access to ancient narratives. Data visualization and sonification are also innovative methods for digital curatorial practice to engage diverse audiences. Data curated this way makes for enhanced interpretation, aligning with the FAIR principles of findable, accessible, interoperable, and reuse.

Different aspects of this research project have been presented in exhibitions in the UK (2022) and Italy (2023), as documented in peer-reviewed journal articles co-written with Toni Sant (2023 and 2024). These published outputs explore the practical applications of Wikidata in the digital curation of art collections. The importance of structured data in enhancing the accessibility and interpretability of cultural artifacts is emphasized throughout. Wikidata's open and collaborative platform allows for the collaborative enhancement of standard GLAM metadata, facilitating connections between disparate collections and potentially fostering more inclusive representations. By leveraging Wikidata's capabilities, artists and curators can create dynamic narratives that reflect the multifaceted nature of cultural heritage, promoting greater public engagement and understanding.

2. Initial work with Wikidata

The first step of working towards the application of Wikidata for artistic creation is to engage with the data platform to collect, analyse, and store data. The selected dataset for this exercise, as detailed in this paper, involves art and/or museum collections. Effective data analysis begins with thorough data curation. This method includes: (a) identifying and verifying items related to selected artists, artworks, and institutions, (b) augmenting entries with metadata (e.g., medium, creation date, acquisition source), and (c) linking items through shared properties (e.g., region, motif).

Structured data, in this case available through the Wikidata platform, is useful in various ways. However, it is not always easy to extract useful knowledge from the data beyond the obvious, unless it is processed through data visualization tools. Working as a professional artist for over a decade, I came to data curation from my interest in art curation. I developed my first insights into data curation as a postgraduate student at the University of Salford's Digital Curation Lab. My first practical experience with digital curation led me to SPARQL queries in Wikidata, and subsequent data visualizations arose from that work. The RDF language quickly felt very familiar to me, probably because I favor visual learning and communication.

My initial experience with data analysis through data visualization came in 2020, when I worked on exploring gender gaps in the University of Salford's Art Collection. One of my aims with that project was to create better awareness of the works by women artists in that institution's art collection. I addressed this by collecting data and exploring ways to make that information accessible to a broader audience. This was instigated by the University of Salford's Art Collection desire to commission and collect more works by female and non-binary or queer artists.

The digital curation work on this art collection involved significant input on the reorganization of a rudimentary spreadsheet containing a list of artists that formed part of this collection. The spreadsheet was expanded into a structured data schema, which could be ingested into Wikidata.

When combined with a SPARQL query, which was also presented through data visualization tools, this basic exercise helped the collection curators more easily see the geographic provenance of the artist whose work is in the university's art collection come from, enabling them to see the spread of locations represented in the collection, simply through information about the artists' place of birth. In most cases, this affirmed the institution's regional acquisition policy. Other useful properties for which metadata was acquired include the date of birth of the respective artists, as well as the year when specific works of art were acquired by the university. When queried and visualized, this provided a way to easily identify gaps in the timeline, such as when works by women artists were not being acquired for the collection.

Mapping the data geographically and looking at timelines through the dataset instantly revealed insights into the collection that the curators previously

did not have access to. The connections that are enabled through linking data about the same artists in this collection with other collections elsewhere is particularly intriguing and quite useful. This highlights a basic goal of digital curation as a discipline, which is to engage in ways for adding value to data. This initial phase of the broader research project outlined here enabled better awareness of women artists in the University of Salford Art Collection and in the process, increased the visibility of women artists in the collection. The dataset enabled the identification of gaps through data visualization and provided support for the acquisition of relevant works by women artists. Patterns and gaps tell stories about a collection.

This particular dataset also provided a resource from which to help create or improve Wikipedia articles about women artists, as well as engage further with the collection more broadly, regardless of gender. This exploratory work with the university's art collection inspired other postgraduate students at the Digital Curation Lab to employ this methodology on their own projects on other art and/or museum collections. This third-party application of this methodology proved that this method of inquiry is easily adaptable in other contexts within the GLAM sector. This adaptability led me to investigate the possibilities of expanding this type of data analysis into creative outputs. In simpler terms, this work raised the question: how can art be created from the interpretation and manipulation of queries on datasets in Wikidata?

The initial application of SPARQL queries focused on identifying underrepresented artists, particularly women and non-binary creators, in the University of Salford's art collection. A refined query structure was subsequently adapted to other institutional collections, generating datasets that could yield other types of interpretation, including creative perspectives. The subsequent SPARQL queries informed the need to employ data-cleaning strategies, revealing data gaps and prompting curatorial interventions.

Data representations in the form of visualizations or sonifications facilitate an engagement with datasets that goes beyond surface-level observation, allowing for an exploration into the internal structure and dynamics of the data. This approach reflects a fundamental principle of data science, wherein the primary objective is to identify and analyze relationships among individual elements, linking them through shared attributes and underlying patterns. Rather than presenting data as static collections of facts, such representations encourage interpretive interaction, promoting a deeper understanding of how the data functions as a coherent whole.

As data humanist Giorgia Lupi (2017) explains, the role of design in data representation is both functional and interpretive. Design serves the data by shaping how it is perceived, and it is often the first element encountered in a visualization. This framing effect positions design not merely as a tool of presentation, but as an integral component of meaning-making. In this way, data visualization begins to intersect with data art, where aesthetic and conceptual considerations enhance the communicative and expressive capacity of the representation.

Structured data, and particularly that offered through the Wikidata platform, offers significant utility across various domains. These datasets support

multilingual digital knowledge platforms such as Wikipedia and also provide essential content for artificial intelligence systems, including voice assistants like Siri, Alexa, and ChatGPT. Such systems rely on structured data to retrieve and deliver relevant content in response to user input, underscoring the foundational role that structured knowledge plays in contemporary AI applications.

Nevertheless, the process of extracting insights from structured data raises questions on how to handle information beyond readily apparent facts, deeper patterns, and connections that can remain obscured without the appropriate tools for analysis. To address this line of thinking, data-visualization tools are increasingly employed as a means of making complex datasets more legible and meaningful. These tools not only enhance comprehension but also open new interpretive possibilities, enabling researchers to uncover otherwise hidden structures within the data.

3. Data visualisation as art

The next phase of this research project is based on a Wikidata dataset about artifacts within the two collections of female prehistoric figurines and fragments held at the National Museum of Archaeology in Valletta, Malta, and the Ġgantija Archaeological Park in Gozo. Significant numbers of prehistoric figurines, depicting mainly humans and animals, were discovered and collected in the central Mediterranean island nation of Malta as early as the late nineteenth century and throughout the twentieth century (Pace 1996; Vella Gregory 2005).

Working with incomplete or variably categorized data, the research work began with the construction of a series of Wikidata items covering artifacts from the Ġgantija Archaeological Park and the National Museum of Archaeology in Malta. Each figurine was linked to properties such as depicts (P180), material used (P186), and location (P276). Through a process of enrichment, the dataset was made more interoperable, allowing it to be compared to prehistoric figurines in other collections (e.g. British Museum, Louvre etc.). This enabled not only internal reflection but also external scholarly dialogue. The next methodological stage translated these data into visual networks and eventually sound compositions, extending curatorial analysis into aesthetic experimentation.

This approach involved modeling figurines in clay, drawing inspiration from the representations within the dataset. This tactile process offered insights into the proportions and aesthetics of the original artifacts. This embodiment of data underscores the potential of artistic practice to breathe life into digital information, making abstract data tangible and relatable.

The model is composed of five key steps:

1) The first stage involves the systematic identification of relevant items on Wikidata, including artworks, artists, institutions, and themes. SPARQL queries are used to retrieve structured data for analysis. These queries serve both as retrieval mechanisms and as curatorial filters, allowing institutions to uncover gaps, overlaps, and patterns within their collections.

2) Once data is retrieved, it is reviewed for completeness, accuracy, and relevance. This often includes adding missing metadata (such as the material of the work), correcting erroneous values, and linking items via shared properties.
3) It is important that the visual representation of data reveal patterns not evident in raw tables. Network graphs, timelines, and maps are used to illustrate relationships between artists, artworks, and acquisition trends. For example, gender-based visualizations can show the underrepresentation of women in art collections over time or across institutions.
4) A distinguishing feature of this methodology is its integration of creative practice. Drawing from the notion of "data humanism" (Lupi 2017), data is not only visualized but also potentially sonified, transformed into aesthetic artifacts that prompt emotional and intellectual engagement. This process makes metadata tangible and invites alternative interpretations, expanding curatorial narratives beyond the museum wall.
5) Outputs such as interactive dashboards, exhibitions, and installations serve as platforms for public interaction. Feedback from audiences is not merely evaluative but participatory, viewers may become contributors to Wikidata, expanding the scope and richness of the dataset. We describe this as the democratization of curatorial data (Sant and Tabone 2023).

Artistic expression is a broad research method that relates very directly to practice as research (Bishop 2023). It can be argued that artistic expression is only a research method within a framework of practice as research. Within this project, this assertion easily holds true. Nevertheless, there are some formal art techniques that cannot be overlooked. The first of these is the way Rudolf Arnheim (1974) discusses rules of visual grouping through what he calls the "principle of similarity" (Arnheim 1974, 67). In this way, shapes and patterns (including the physical material used to create an object) can be grouped by the way they resemble each other. By extension, Arnheim's principle of similarity can be related to a construct such as "similarity of location," even if this is probably better aligned with what he credits Gestalt psychologist Max Wertheimer, calling the "rule of proximity or nearness" (Arnheim 1974, 67). Figure 1 is designed to better illustrate this concept.

What appears to be a random angular line, if not an abstract shape, is based on a visual exercise creating a line joining the dots on a map of the main sites associated with discoveries of prehistoric female figurines in the Maltese islands. To be more precise, the map is generated through a Wikidata SPARQL query mapping the locations where Heritage Malta's prehistoric female figurines were found. The line is drawn following what Arnheim calls a "consistent shape" exercise. This follows from his assertion that "the straight line is more identified than the irregular" (Arnheim 1974, 71). Building on Arnheim's theory of visual perception, the locations are stitched together through a line in the way humans tend to recognize, interpret, and give meaning to visual forms.

The first work of art created through this research process is a sculpture made from transparent plexiglass rods, designed to be part of an art installation called

Naked Data. The choice of transparent plexiglass in the sculpture created within the framework of this research project can also be viewed in the context of the concept of transparency inherent in the ways open knowledge works through Wikidata and other Wikimedia platforms.

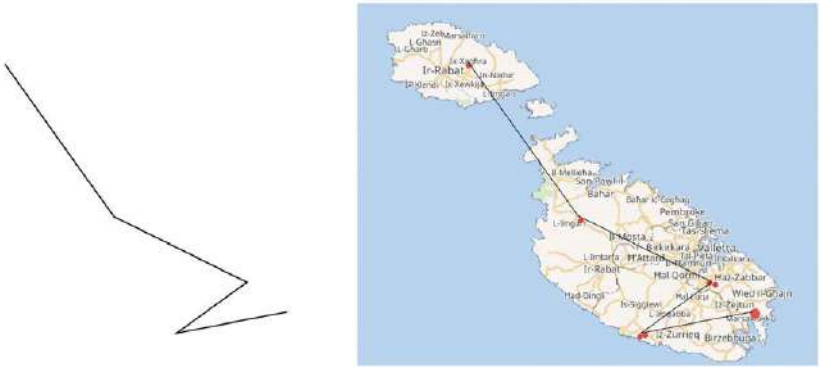


Figure 1 – A line drawing is derived from joining location dots on a map showing the main sites associated with archaeological excavations across the Maltese islands that have yielded objects held within Heritage Malta's collection of prehistoric female figurines.

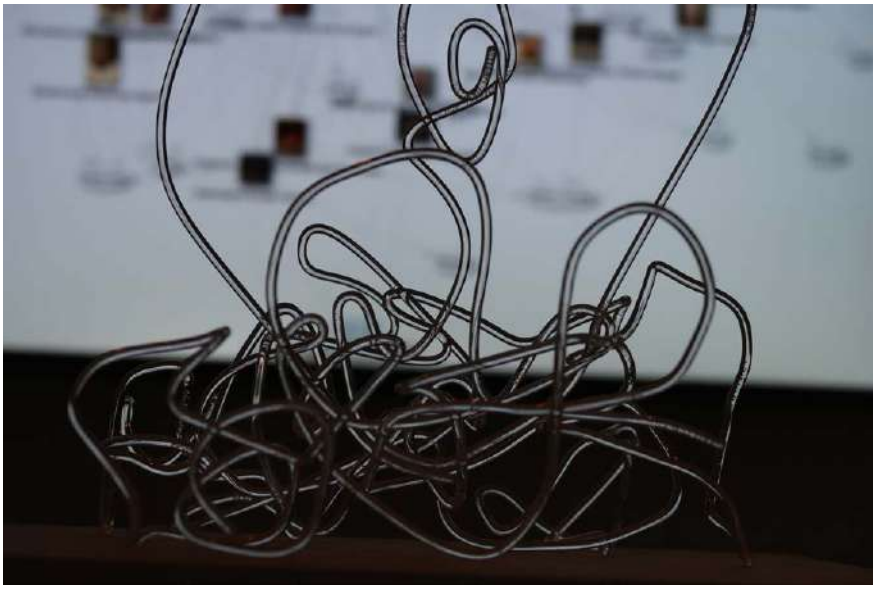


Figure 2 – Detail from Naked Data (2022), composed to show the contrast on stiffness and rigidity between data visualization and data art, intended to present the data in a less rigid mode.

Another significant aspect within this method is known as "parallax", whereby seeking information is a playful exercise that is less goal oriented than regular data retrieval and/or analysis (Drucker 2013). This is certainly a good way to describe the practice as research approach to artistic creation in data art. In a less goal-oriented data retrieval mode, such as an artistic/creative research-driven approach, the engagement is through curiosity and awareness.

Artistic expression through a practice as research method, as presented here, can provide new insights for curators in terms of what they can do with their collections and the way they are displayed to their audiences. Rather than focusing only on screen displays within the museum, curators can create new experiences that can also be experienced off-site, providing opportunities for vertical, horizontal, or tangential explorations that are otherwise not possible within the static museum display experience. A specific example of this, as it relates to the research project presented here, involves a way to potentially link up the objects displayed at the two museums more directly in ways that most visitors presently overlook, unless they happen to have visited both museums (which are on different islands) in close temporal proximity. There is also the possibility of presenting speculative data, even if this is an artistic rather than an archaeological approach.

A second output resulted from the artistic work using the same method on the Heritage Malta dataset to initiate an artistic exploration involving the visualization of scientific data, with special attention to the aesthetic qualities afforded by this technological engagement. This resulted in the exhibition of an artwork that consisted mainly of a video-art installation, depicting aspects of data art as embodied in this part of the broader research project. This was presented at the Stanley Picker Gallery in Kingston, London, in September 2022, along with a physical object developed from the artistic interpretation of the data visualization contained in the video. The purpose of the physical object was to materialize the data work beyond digital technology, to return it to forms that are more congruent with the historical art objects captured in the dataset. The exhibition was curated by Bill Balaskas.

4. Data sonification as a way to make art with Wikidata

Beyond visual representation as shown in the 2022 London exhibition, the research explored sonification, translating data attributes into auditory experiences. By mapping variables like gestures and anatomical features to sound parameters, the composition created and audibly conveyed the dataset's nuances. This method not only provided a fresh perspective on the artifacts but also highlighted the emotional resonance of data when expressed through sound.

This process was especially critical in the case of the prehistoric figurines collection in Malta, where categorical inconsistencies were found in existing datasets. Objects from the National Museum of Archaeology and Ġgantija Archaeological Park were re-categorized, linked semantically, and prepared for visualization, as described earlier, and eventual sonification.

The application of the creative model to the chosen objects from Heritage Malta's collection provided insights into the symbolic significance and cultural distribution of prehistoric female figurines. Mapping visualizations traced the geographic spread of these artifacts, while other creative visualizations highlighted the materials of figurines and what they depict. Data sonification further explored symbolic elements within the dataset by translating recurring motifs into dynamic sound patterns. As intended here, the data representation aspect is both scientific and artistic in its approach. Scientific in the way it handles facts and figures associated with them, and artistic in its ability to present creative experiences that enable diverse audiences to understand things they may overlook when presented to them otherwise.

Using properties such as "depicts" (P180), freely accessible open creative applications translated recurring themes and poses from prehistoric figurines into pitch, rhythm, and timbre. Each of the 227 data points became a unique sonic trigger. Processed in digital audio workstations, these sounds were composed into an immersive audio installation. The result was an evolving artwork that foregrounds data aesthetics and interpretation. The application of the model to the selected items from Heritage Malta's collection provided insights into the symbolic significance and cultural distribution of prehistoric female figurines. Mapping visualizations traced the geographic spread of these artifacts, while other creative visualizations highlighted the materials of figurines and what they depict. Data sonification further explored symbolic elements within the dataset by translating recurring motifs into dynamic sound patterns.

The practice of sonification, in particular, enabled this project to offer a novel form of artistic curatorial narration (Worrall 2018). By mapping dataset variables (e.g., gestures, anatomical features) to audio parameters (e.g., pitch, rhythm), the resulting soundscapes embodied the data's relational structure. Elsewhere, we have observed that such methods do not merely communicate data but "activate it," making the invisible audible and emotionally resonant (Sant and Tabone 2024).

Figure 3 – The Naked Data sound file is available at <https://on.soundcloud.com/Ex6J7>.

This art installation interprets a dataset of prehistoric female figurines from Heritage Malta through visual and sonic mediums. In the process, it provides an example of the ways data curation and artistic practice can be fused. By transforming structured data into immersive experiences, the art project challenges traditional curatorial methods and invites audiences to engage with cultural artifacts in novel ways. Like visualizations, data sonification opens the door for more inclusive

museum practices. Non-visual learners, or those with disabilities, may find sonification a more effective means of engagement. Moreover, it invites the broader public to experience data in novel ways, fostering a more diverse and interactive museum experience. Being aware that acquiring skills in music composition and/or sound production can enable a conventional visual artist to create new works that are potentially more meaningful or impactful from a technical perspective.

This was demonstrated through a sound-art installation diffused through an individual pair of headphones, accompanied by purposely created physical objects presented during an exhibition at the Palazzo del Rettorato of the University of Turin, curated by Federica Patti. This work should be considered within the context of the broader body of work created between 2019 and 2023 around the theme of reimagining prehistoric female figurines from a feminist perspective (Sant 2023). The sound art piece developed through the data-sonification exploration within the framework of this project has also been integrated in a solo art exhibition, presented at Spazju Kreattiv in Malta, in March and April 2023. Within the context of the Malta exhibition, the sound-art work presented in Turin enabled the presentation of artistic insights into Heritage Malta's collection of prehistoric female figurines in an art setting outside the structure of the formal research work conducted within the Digital Curation Lab. This work of art was designed to extend the concept of materiality beyond specific contents of Wikidata, which are taken to be a medium of production rather than dissemination.

5. Conclusion

Working with Wikidata to create art emphasizes the role of the artist as both data manager and storyteller. By curating through data, the project demonstrates how structured information can be reinterpreted to craft compelling narratives beyond the obvious. A distinguishing feature of this model is the use of data sonification, translating dataset parameters into auditory elements. This technique allows for the exploration of data beyond the visual, opening new sensory dimensions of interpretation. This method of making art advocates for the integration of linked open data technologies with creative methodologies in the cultural heritage sector. The model developed here demonstrates how SPARQL, Wikidata, and artistic expression can collectively transform the way cultural institutions curate, understand, and share their collections.

Further research stemming from these initial explorations involves fostering public engagement, encouraging audiences to explore and contribute to the underlying datasets, thus democratizing the curation process. This is being conducted through the WikiProject Każini (Q126912787), within the community structures afforded by Wikimedia Community User Group Malta, which is a local affiliate of the Wikimedia Foundation. Within this new line of exploration, data sonification is also being investigated within the data curation from Wikidata and Wikimedia Commons.

Now that it has been established theoretically and applied to two ways of making art (visual and sonic), the model could also be applied by others to interpret

underrepresented narratives, such as those expected in indigenous collections from postcolonial viewpoints. Further research might explore real-time data interaction in museum environments, AI-assisted visualization, or participatory data curation models involving local communities. As digital infrastructures continue to evolve, so too must curatorial practices. The convergence of data science and artistic research, as illustrated here, offers a promising avenue for more inclusive, dynamic, and insightful engagements with the cultural record.

The evolving field of digital curation in the GLAM sector is increasingly shaped by interdisciplinary models that blend data science, artistic practice, and public engagement. The creative model presented here, integrating linked open data, SPARQL querying, and artistic methods such as data visualization, mapping, and sonification, demonstrates how cultural heritage data can be reimagined as experiential narratives. This model not only offers technical tools for data interpretation but also inspires new curatorial philosophies grounded in openness, collaboration, and creativity.

As highlighted in this paper, data can serve as an artistic medium in its own right. The translation of structured metadata into sound compositions and sculptural representations invites audiences to engage with heritage in multisensory, affective ways. This approach foregrounds the role of the artist-curator as an interpreter of data, emphasizing how curation is both a creative and political act. Moreover, the project exemplifies how GLAM institutions can repurpose open datasets like Wikidata to deepen interpretation, provoke inquiry, and democratize access to cultural knowledge. Future directions for this research model include its application to underrepresented collections and communities, fostering more inclusive digital heritage practices. There is significant potential for expansion into postcolonial and decolonial curatorial strategies, where data and metadata can be reframed to prioritize indigenous knowledge systems, minoritized narratives, and alternative epistemologies. Additionally, further exploration of immersive media technologies such as interactive audio installations and virtual exhibitions can deepen public engagement and reimagine the boundaries of museum experience. This approach is now being applied to a new research project with the University of Salford's GLAM Lab in collaboration with the Greater Manchester Museums Group.

Ultimately, this work illustrates how data curation, when approached creatively and critically, can serve not only as a means of knowledge production but also as a powerful mode of cultural storytelling. As digital infrastructures continue to evolve, the GLAM sector stands to benefit from integrating artistic research and open data methodologies in transformative and future-facing ways.

References

Arnheim, Rudolf. 1974. *Art and Visual Perception: A Psychology of the Creative Eye*. Berkeley: University of California Press. https://doi.org/10.2307/426441.

Bishop, Claire. 2023. "Information Overload." *Artforum*, April 2023. https://artforum.com/print/202304/claire-bishop-on-the-superabundance-of-research-based-art-90274.

Drucker, Johanna. 2013. "Performative Materiality and Theoretical Approaches to Interface." *DHQ: Digital Humanities Quarterly* 7 (1).

Lupi, Giorgia. 2017. "Data Humanism: The Revolutionary Future of Data Visualization." *Print Magazine*, 30 January 2017. https://www.printmag.com/article/data-humanism-future-of-data-visualization/.

Pace, Anthony, edited by. 1996. *Maltese Prehistoric Art, 5000-2500 BC*. Valletta: Fondazzjoni Patrimonju Malti.

Sant, Toni, and Enrique Tabone. 2023. "Creative Uses of Wikidata for Art Collections." *International Journal of Performance Arts and Digital Media* 19 (3): 445–62. https://doi.org/10.1080/14794713.2023.2253335.

Sant, Toni, and Enrique Tabone. 2024. "Art through Wikidata: A Digital Curation Perspective." *Mimesis Journal* 13 (2): 379–89. https://doi.org/10.13135/2389-6086/10021.

Sant, Toni. 2023. *Enrique Tabone, Catalogue Raisonné*. Malta: Kite Group.

Vella Gregory, Isabelle. 2005. *The Human Form in Neolithic Malta*. Malta: Midsea Books.

Worrall, David. 2018. "Sonification: A Prehistory." In *Proceedings of the 24th International Conference on Auditory Display – ICAD 2018*, 177–82. Houghton: The International Community for Auditory Display. https://doi.org/10.21785/icad2018.019.

A Visualisation System Based on Wikidata for Supporting and Monitoring the Wiki Loves Monuments Italian Contest

Tommaso Elli, Andrea Benedetti, Angeles Briones,
Dario Crespi, Michele Mauri

Abstract: This paper presents the Wiki Loves Monuments Observatory, a data-driven platform designed to support and monitor the Italian edition of the Wiki Loves Monuments contest through Wikidata and Wikimedia Commons. The system includes a backend for automated data collection and a user interface that enables exploration, coordination, and reporting at multiple administrative levels. Developed through a participatory design process, the tool supports contest organisation, volunteer engagement, and cultural heritage documentation. Reflections highlight the challenges of involving a heterogeneous volunteer community and the need to expand co-design practices. The system offers a replicable model for integrating open data and community-driven design within Wikimedia ecosystems.

Keywords: Wikidata, Wiki Loves Monuments, Cultural Heritage, Co-Design, Public Engagement.

1. Introduction

Wikimedia Italia is a social promotion association that has been active since 2005 in the cultural field, supporting the dissemination of free knowledge. Through the work of its staff and many volunteers (more than 2500), it supports the projects of the Wikimedia Foundation[1] and OpenStreetMap[2]. To

[1] Wikimedia Foundation Inc. is a non-profit organization established in 2003. Its mission is to encourage the development and dissemination of free content in all languages and to make the entirety of its projects' content freely available to the public. The most well-known of these projects is the encyclopedia Wikipedia, which ranks among the ten most visited websites in the world. Adapted from: Wikimedia Foundation. (February 21, 2024). Wikipedia, The Free Encyclopedia. <https://it.wikipedia.org/wiki/Wikimedia_Foundation>.

[2] OpenStreetMap (OSM) is a collaborative project aimed at creating free-content maps of the world. The project gathers geographic information on a global scale with the goal of

Tommaso Elli, Polytechnic University of Milan, Italy, tommaso.elli@polimi.it, 0000-0002-9818-1991
Andrea Benedetti, University of Milan La Statale, Italy, andrea.benedetti@unimi.it, 0000-0001-7121-459X
Angeles Briones, angelesbriones@gmail.com,
Dario Crespi, Wikimedia Italia, Italy, dario.crespi@wikimedia.it, 0000-0001-5003-4442
Michele Mauri, Polytechnic University of Milan, Italy, michele.mauri@polimi.it, 0000-0003-1189-9624

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Tommaso Elli, Andrea Benedetti, Angeles Briones, Dario Crespi, Michele Mauri, *A Visualisation System Based on Wikidata for Supporting and Monitoring the Wiki Loves Monuments Italian Contest,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.34, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 287-301, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

improve the quantity and quality of information available on the online encyclopedia, since 2012, Wikimedia Italia has organised the photo contest Wiki Loves Monuments Italy (WLM), which aims to enhance documentation of cultural heritage through data and photography (Bertacchini and Pensa 2023; Azizifard et al. 2023). All materials, including high-quality images produced by professional photographers, use licenses that allow free reuse for any purpose. The Italian contest is part of a broader international initiative: in 2023, Wiki Loves Monuments was held in 46 countries, confirming its status as the largest photography competition in the world, as certified by the Guinness World Records. As of 2023, thirteen years after its first edition, the contest has collected over 3,000,000 images from over 82,000 participants worldwide[3], uploaded on Wikimedia Commons[4]. The strength of the project lies in the composition of its large group of organisers: Wiki Loves Monuments is organised almost entirely by a wide network of volunteers who, either individually or in coordinated groups, work each year on a voluntary basis to improve monument-related data, promote the contest, and engage new photographers (Fig. 1). The organisation of Wiki Loves Monuments in Italy also has a unique feature: it is the only national contest structured on multiple levels—one at the national scale and others covering smaller regions. Each year, WLM Italy includes a variable number of regional or local competitions, entirely managed by local volunteers. In addition to organising the contest, volunteers are responsible for its communication and for increasing participation, for example, by organising "wiki-expeditions"—trips aimed at collecting free images of Italy's cultural heritage, especially in remote areas less visited by tourist routes. These outings also serve as opportunities to activate and highlight cultural heritage in underrepresented territories.

In its early editions, WLM Italy relied on manually compiled lists of monuments, updated on a yearly basis. Since 2018, the contest has used Wikidata (Vrandečić and Krötzsch 2014), the database behind Wikimedia projects and the world's largest open data repository. Volunteers can create and update Wikidata entries on a wide range of topics—including cultural heritage—either manually or using semi-automated tools to import data from other databases with compatible licenses. This system has made it possible to manage lists containing a large number of monuments, growing from the 11,500 monuments included in

producing and distributing open maps and cartographic data. These maps can be freely used for any purpose, including commercial ones, provided that the source is credited and the same license is applied to any derivative works based on OSM data. Adapted from: OpenStreetMap. (April 21, 2025). Wikipedia, The Free Encyclopedia. <https://it.wikipedia.org/wiki/OpenStreetMap>.

[3] Further statistics on the WLM contest at both the global and national levels are available at the following address: <https://wikiloves.toolforge.org/monuments>.

[4] Wikimedia Commons is the repository of multimedia contents of Wikimedia projects. Read more at: <https://commons.wikimedia.org/>.

the 2018 edition to almost 100,000 in 2024[5]. This work is not only a prerequisite for uploading high-quality photographs but also enhances the metadata of the images and the data within Wikidata. Each monument is described using statements related to its name and location (geographic coordinates, country, municipality, district if applicable, street address), but can also include additional information such as construction period, ownership, surface area, conservation status, and more. Moreover, individual entries are often linked to corresponding items in other databases, enabling integration across different Wikimedia projects. OpenStreetMap volunteers also support territorial mapping, and all mapped monuments can be linked to their respective Wikidata entries—thus creating a unified system of open data, metadata, and freely accessible information for public use.

Figure 1 – The working group of Wiki Takes Catania, 1-3 September 2023. Photo by Iolanda Pensa, CC BY-SA 4.0, via Wikimedia Commons.

In 2021, the association engaged with a team of information designers because it identified the need for tools to support volunteers' work, facilitate activity coordination, monitor results, and engage new audiences in order to improve the quality and quantity of the content produced.

[5] Further information about the Italian edition of WLM is available at the following address: <https://it.wikipedia.org/wiki/Progetto:Wiki_Loves_Monuments_Italia/Storia_e_statistiche>.

2. Background

In the Wikimedia ecosystem, there are tools for mapping, visualising, and analysing documented cultural heritage[6]. *WikiShootMe* is an interactive map that displays Wikidata items, Wikipedia articles, and Commons images with coordinates to help volunteers identify monuments to photograph[7]. *Wikidata To Do*[8] organises Wikidata entities by countries to easily identify those which are missing data of different kinds, such as places or objects without a picture. *Cassandra*[9] is an analytical platform that offers a user-friendly dashboard for GLAM institutions (i.e., Galleries, Libraries, Archives and Museums) supporting "gaining statistical insights regarding their shared collections across the Wikimedia projects in the world." *Wiki Loves Monuments Map*[10] provides an interactive geovisualization of registered monuments and their images, with filters to distinguish photographed/unphotographed sites, type of monuments and availability of corresponding Wikipedia articles; it is currently available in France, Ireland, Poland, Sweden, the United Kingdom, and the USA (missing Italy). *Wiki Loves Competitions Tool*[11] aggregates metrics across Wiki Loves photo contests (WL Earth, Folklore, Science, Monuments etc.), providing photo upload quantities across countries and contests. In Italy, the community has developed specific tools. *Statistiche Wikilovesmonuments*[12] is a campaign monitor that provides aggregated statistics on regional contest participation. *Wiki Loves Monuments Italia*[13] allows users to search monument records via Wikidata identifiers positioned on a map; it is worth mentioning that after the conduction of the presented project, the tool evolved, and it now provides specific functions to fix monuments metadata and associate images to Wikidata and Wikipedia entries.

The presented tools promote the discovery and documentation of heritage monuments and incentivise the use of structured data through Wikidata to enrich and interlink content. Nonetheless, at the moment of project execution, and if looking specifically at the Italian contest, there remained notable gaps. No tool allowed for (1) an aggregated visual analysis of data related to the Italian contest and all Italian cultural properties available on Wikidata; (2) an analysis of monuments in respect to their typology; (3) a longitudinal and geographical analysis of previous contest editions.

[6] An extensive list of tools is available at: <https://magnustools.toolforge.org/>.

[7] WikiShootMe is available at: <https://wikishootme.toolforge.org/>.

[8] Wikidata To Do is available at: <https://wikidata-todo.toolforge.org/>.

[9] Cassandra is available at: <https://stats.wikimedia.swiss/>.

[10] Wiki Loves Monuments Map is available at: <https://maps.wikilovesmonuments.org/>.

[11] Wiki Loves Competition Tools is available at: <https://wikiloves.toolforge.org/>.

[12] Statistiche Wikilovesmonuments is available at: <https://statistiche.wikilovesmonuments.it/>.

[13] Wiki Loves Monuments Italia (also known as "App di Ferdi") is available at: <https://cerca.wikilovesmonuments.it/>.

3. Methodology

The design process of the presented system employed a blending of user-centred and co-design approaches. After the first contact with Wikimedia, the design team learned about the WLM contest organisation, availability of monuments list, and quality of produced pictures, doing desk research on available websites. Then a designer ran participant observation (Kawulich 2005) and structured interviews (Seidman 2006) in the context of wiki-expeditions. The occasion was fruitful to learn by doing with organisers and volunteers about the contest and the related activities, such as finding and shooting monuments, but also about the importance of connecting with local institutions and volunteers. Some specific issues emerged during the process: unlike other countries, Italy lacks effective legislation on freedom of panorama; furthermore, the dissemination of photographs of monuments may fall under different legal regimes, depending on factors such as ownership status, cultural heritage protection laws etc. As a result, the organisers of the contest request that monument custodians provide explicit authorisation, allowing images to be uploaded to Wikimedia Commons under a permissive Creative Commons license that enables commercial reuse[14]. This aspect was considered during the design process.

3.1 Co-Design Workshop

The described activities enable the collection of rich insights that inform the structure of a co-design workshop held in spring 2022. The workshop aims to collectively define a design brief to guide the development of a digital infrastructure for cultural heritage visualisation. Two multidisciplinary groups, each composed of five participants, including designers and representatives from Wikimedia, participate in the activity. The workshop unfolds across four structured phases, each supported by tailored materials. It addresses two core stakeholder requirements: (1) the need to create analytical visualisations (dashboards) to monitor the campaign (phases 1 and 2), and (2) the need to design visualisations for outreach and engagement, to broaden participation and attract new volunteers (phases 3 and 4).

Phase 1: Consolidating Design Requirements. Participants receive 16 cards describing potential uses enabled by visualisation tools. Examples include: "mitigating tourism-related bias (tourist areas tend to be overrepresented in photos)", "supporting the planning of wiki-expeditions", "fostering friendly competition between local contests", "designing visualisations that can be updated for future editions", or "highlighting under-photographed but authorised monuments". Additional blank cards are available to propose further ideas. Participants assign priorities, discard less relevant cards, and collectively rank and select the most critical uses.

[14] Read more about the "freedom of panorama" at <https://en.wikipedia.org/w/index.php?title=Freedom_of_panorama&oldid=1287100277#Italy>.

Phase 2: Mapping Properties, Uses, and Users. Participants receive 20 cards representing visualisation properties, which describe information types and system functionalities. Examples include: "displaying areas or monuments without photographs", "showing types of monuments", "highlighting authorised or unauthorised monuments", and "displaying monument-specific data". Participants match each property with one or more of the selected uses from Phase 1 and specify the target user group, choosing among four pre-identified categories or adding new ones: Wikimedia Italia staff and board members; local contest organisers; photographers (contest participants); and jury members.

Phase 3: Concept Mapping for Engagement Strategies. Participants create a conceptual map to structure engagement actions. They define the goals of public outreach, identify relevant audience segments, determine the types of content to share, and select suitable communication channels.

Phase 4: Validation and Articulation through Sentence Templates. Participants synthesise and articulate concrete requirements by completing structured sentence templates based on the previous phase. The template reads: "To reach the goal of [objective], we aim at involving [audience], showing them [content] via [communication channel]."

3.2 Workshop Outcomes

Phases 1 and 2 clarify relevant use cases and their technical implementation, while Phases 3 and 4 define the communication objectives and content structure for volunteer engagement. The design team consolidated workshop results, turning them into a narrative (in the format of a slide presentation) to be agreed upon with stakeholders. The activities support the early identification and definition of key functionalities, laying the foundation for a robust, modular, and updatable infrastructure that can evolve in response to future needs.

The process leads to strategic decisions concerning the system's scope. A functional separation emerges between monitoring the WLM campaign and guiding photographers. The new dashboard focuses exclusively on monitoring, while the task of guiding contributors is delegated to a separate application. The dashboard must generate reports for social accountability, public presentations, and media kits. Participants emphasise the importance of communication with the public and journalists, particularly to encourage local administrations to authorise more monuments. Data updates should occur at regular intervals, specifically at the beginning, midpoint, and end of the campaign, accounting for the fact that uploads often peak toward the end of the contest. The monitoring system must provide insights at multiple administrative levels: regional, provincial, and municipal.

The applied methodology identifies the following design requirements: (1) support the organization of the national and local Wiki Loves Monuments competitions; (2) foster volunteer engagement both to improve the cultural heritage database and to encourage the production of new photographs; (3) enable communication of the contest through social media, newsletters, presentations, and

public events; (4) monitor the representation of Italian cultural heritage across Wikidata, Wikimedia Commons, and Wikipedia.

4. Design and Implementation

After consolidating the outcomes of the co-design workshop, the design team initiates the ideation phase. Stakeholders are involved throughout the design and implementation process to provide feedback and validate or negotiate design decisions.

4.1 Data

The first step was the design of the data collection, which was based on a series of SPARQL queries to Wikidata to compile a comprehensive list of cultural heritage monuments, including associated metadata. One of the primary challenges encountered lies in identifying what Wikidata entity qualifies as a monument. Definitions of a cultural property[15] remain intentionally open-ended, offering general criteria rather than rigid classification; they require expert judgment, consensus, and interpretative reasoning. The project adopts a list of entities identified as instances of cultural heritage property classes in a prior Wikimedia project[16], which are used as monument typologies. The initial list is integrated with panoramic views of Italian municipalities and any items that participate in the Wiki Loves Monuments (WLM) contest, regardless of typology, recognisable on Wikidata because they include a WLM Italy ID and an authorisation. After removing duplicates, each monument is then geolocated to its corresponding municipality using a spatial join between its coordinates and administrative boundaries from ISTAT (CC-BY 3.0), which include official updates on municipal borders over time. Monuments with missing or incorrect coordinates (e.g., those located in the sea) are grouped under the label "Unknown Region." For each monument, the system retrieves available image metadata from Wikimedia Commons by referencing the WLM Italy ID (Fig. 2).

The resulting dataset is designed to enable year-by-year calculations, starting from the first edition of the contest in 2012. For each administrative area, the system quantifies: (1) the total number of monuments registered in Wikidata, (2) how many are authorised for participation, and (3) how many have been photographed at least once in the context of WLM.

[15] For example the definition provided by UNESCO at <https://whc.unesco.org/en/compendium/154com>.

[16] The project is titled "Visualizzazioni del patrimonio culturale dei comuni italiani sui progetti Wikimedia" and was conducted with Synapta. Results are available at: <https://bit.ly/43dWRU6>.

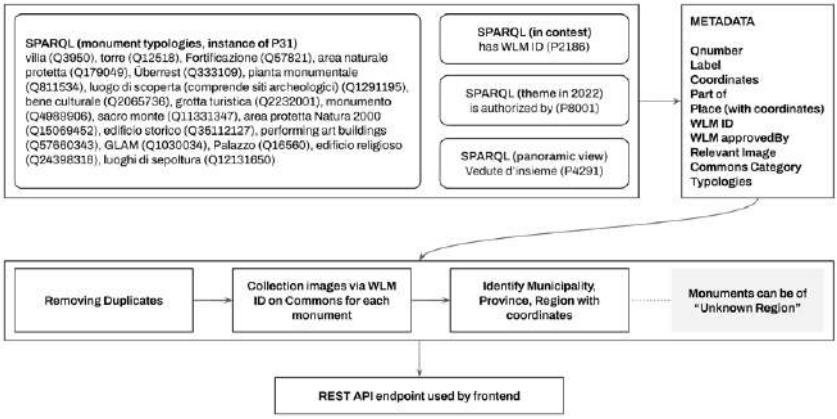


Figure 2 – The diagram illustrates the data collection pipeline. SPARQL queries retrieve instances of selected Wikidata classes to build a comprehensive list of Italian monuments. The system then links available images from Wikimedia Commons to each monument to track coverage and monitor the WLM Italy campaign activity.

4.2 Visualisation and UI

The design team designed a fan-shaped data visualisation model to represent monument status over time. This visual model shows the annual distribution of three monuments' statuses: recorded in Wikidata, authorised for the contest, and documented by photographs. Each administrative area displays its own "fan chart", with the central point pointing to its geographic location. A dedicated "Unknown Region" chart aggregates monuments lacking valid location data. The visual model supports geographic exploration on different administrative levels (i.e., regions, provinces, municipalities) and is embedded in a user interface that allows filtering by various parameters and access to the full monument list (see next section).

The model emerged through an iterative and heuristic process. Initial hand-drawn sketches explored viable uses of visual variables to encode the available data (Bertin 1967). Early digital prototypes use random values to test scalability and legibility of the model in a variety of value ranges, given that actual data were not available yet. The prototypes then incorporate real data without spatial positioning to evaluate visual behaviour under more realistic conditions, and finally align the model with geographic coordinates. Each step reveals unexpected visualisation challenges, such as element overlap and scale calibration, that are resolved through iterative trial-and-error attempts. The team prioritised early integration of real data to confront the increased challenge that data introduces in the visual and interaction design process (Walny et al. 2019).

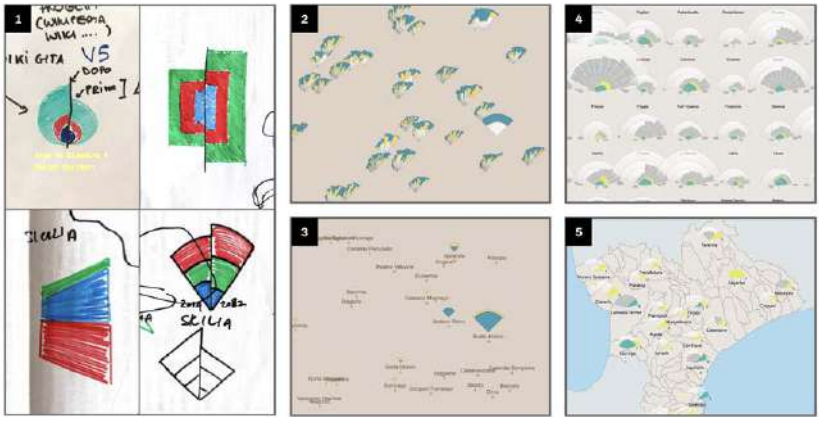


Figure 3 – Illustrates the fan-model ideation process: (1) freehand sketches, (2) digital prototypes with random values, (3-4) prototypes with real data, and (5) final geolocated visual models.

5. Results

The project is developed by the design team for what concerns visualisation and graphical user interface, in collaboration with a software development studio for the implementation of the backend for data collection and API endpoint. The designed system consists of a backend and a user interface (UI).

The backend is a Django-based server application designed to model, collect, and serve data related to Italian monuments. It defines a structured SQL database, executes SPARQL-based scraping tasks to extract monument metadata from Wikidata and Wikimedia Commons, and exposes a REST API for frontend access[17]. The application stack includes PostgreSQL for persistence and Redis for caching and inter-process communication. It is also deployed using Docker Compose. Cron Tools is used to schedule periodic snapshots (i.e., updates). The database manages entities such as monuments, image metadata, and scraping instructions, and supports importing shapefiles and updating geographical datasets using custom scripts.

The UI[18] empowers volunteers, heritage professionals, and local organisers to plan outreach campaigns, coordinate activities (e.g., wiki-expeditions), and enrich digital cultural repositories leveraging a map visualisation and a filterable list of cultural properties. It offers both aggregated and granular views, enabling users to understand coverage patterns, identify areas in need of documentation, reveal temporal trends in community-driven efforts, and produce visual reports. It is composed of two views.

[17] The documentation is available at the following link: <https://wlm-it-visual.wmcloud.org/api/schema/swagger-ui/#/>.

[18] The application is accessible at: <https://data.wikilovesmonuments.it/>.

Figure 4 – The map view allows exploring the Italian territories at different administrative levels: from the entire country (at the top) to regions (Emilia Romagna, bottom left) and to municipalities of provinces (province of Ferrara, bottom right). The "Export visualisation" button allows for the production of static visuals of the current map status. Legend is on the right.

The geographic view (Fig. 4) aggregates the number of monuments by region, province, or municipality and indicates their status: registered, authorised for the contest, or already photographed. Monuments without geographic coordinates, which prevent assignment to a municipality, province, or region, are grouped in the top-right corner under the label "Unknown Region." Data is visualised using a custom fan-shaped model (previously described) that represents the state of documentation at a specific point in time. The fan summarises the historical progression of registration and photographic coverage per geographic area. A histogram in the lower-left corner shows the aggregate data across all visible areas on the map. The view allows temporal exploration by selecting a historical time frame and supports filtering by monument status or typology. Current typologies are defined at scraping time and include, for instance, fortifications, religious buildings, GLAM institutions, palaces, nature reserves, monumental

trees etc. Categories are stored in the database and can be updated through a dedicated function. The view also includes export features for communication purposes, enabling the generation of images with legends optimised for small or large formats. This facilitates visual content production for social media posts, reports, or slide presentations.

The list view (tabular view) provides detailed access to individual monument records (Fig. 5). It displays each monument's status along with a wide range of metadata, including the associated Wikidata entity, creation date, number of images submitted during the contest, contest entry date, and the dates of the first and most recent contest images. A collapsible sidebar on the left allows filtering by geographic area, monument type, and status. It also highlights monuments with images on Wikimedia Commons that are not yet linked to a Wikidata image, supporting data curation by volunteers. The sidebar includes a glossary that clarifies column headings, which may be too concise or domain-specific for immediate understanding, and a download section that enables users to export the full dataset of currently mapped monuments in spreadsheet format.

The frontend is implemented using Next.js (a React-based framework), D3.js, and various NPM packages. All backend and frontend code is released under an open license[19].

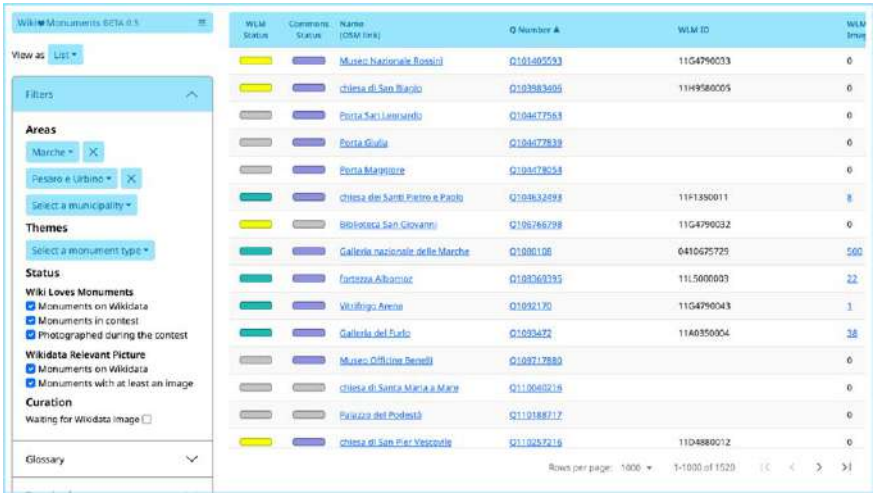


Figure 5 – The list view enables direct access to data for each monument. Filters are consistent with those available in the geographic view, except for the temporal filter. The interface includes specific functions to support data curation activities.

[19] The source code of all described applications is available at: <https://gitlab.wikimedia.org/repos/wikimedia-it/wlm>.

6. Discussions

The relevance of this work lies not only in the technical demonstration of using Wikidata as a foundational infrastructure, but also in the participatory design approach (Dörk et al. 2020; Morelli et al. 2021; Sanders and Stappers 2008) that underpins its implementation. The context offered virtually limitless possibilities for designing a data-driven application because the available content is rich, extended, and free to access on Wikidata, Commons and the other Wikimedia projects. Additionally, there are many different typologies of stakeholders: although the association has a board of directors, it becomes clear that the community of participating volunteers represents the primary user group to prioritise, even though this group is difficult to reach and define due to its high degree of heterogeneity. Identifying and prioritising the features to be implemented is a complex and non-trivial task that was tackled with activities designed to define, in concert with stakeholders, what could be made available and how. The design methodology exploited user-centred methods (i.e., desk research, participant observation and structured interviews) to learn about the context and to inform the realisation of a co-design workshop. This step was crucial, as it allowed for the early identification and prioritisation of all necessary features, facilitating the design of a robust, modular, and upgradable infrastructure capable of adapting to future needs. Besides that, the activity fostered a strong relationship between commissioners and designers, grounded in knowledge sharing, mutual respect, and collective decision-making. As a result, the system is not merely a dashboard visualisation, but a flexible framework where community contributions and open data principles converge.

6.1 Impact

The implementation of the presented tool, released in 2022, serves as the foundation for the development of a companion application dedicated to volunteers[20]. Built on the same infrastructure, the new application introduces features to facilitate user contributions rather than monitoring.

The outcomes of the project and the photo-contest contribute to the enrichment of Wikimedia projects, including Wikipedia articles. The entire software suite is released under an open-source license, with source code hosted in a public repository. This ensures transparency, reusability, and community participation. Several international communities have already expressed interest in replicating or adapting the tools.

Photographs and data collected during the contest are released using tools and licenses that allow unrestricted use, including for commercial purposes, provided proper attribution is given. These materials can be used by institutions or commercial actors to promote Italian cultural heritage—even in remote ar-

[20] The application is available in form of Progressive Web App at: <https://app.wikilovesmonuments.it>.

eas—supporting cultural tourism and local economic development. Moreover, the dataset contributes to research, analysis, and documentation efforts related to Italian cultural heritage.

6.2 Reflections and further works

The implementation of the project, particularly its methodology, offers an opportunity to reflect on the outcomes and implications of stakeholder involvement. The co-design activities prove effective in clarifying the needs and defining a focused design space within which to operate productively. However, only a small subset of volunteers and board members participated, especially when compared to the broader Wikimedia Italia volunteer community.

Although the design process delivered well-conceived and functioning products, part of the volunteer community may perceive it as top-down, particularly those with design, technical, or cultural expertise. These contributors tend to be the most active and consistent, and indeed, the Wikimedia ecosystem already includes numerous tools. They may address a specific function but sometimes present overlapping features, resulting in a wide but fragmented landscape. The learning curve to use these tools effectively for a complex and articulated task is often steep. Following the standards of the digital industry (e.g., UI/UX design), one may think that the goal is to unify and simplify interfaces to improve accessibility and expand participation. However, it remains unclear whether such centralisation, and the decision-making process it entails, would be welcomed by volunteers who take pride in their work and seek to be involved as active contributors. Further research is needed to understand how software production unfolds within the Wikimedia ecosystem, identify strategies for sensibly broadening the co-design process in a participatory way, and collaborate with volunteers.

7. Conclusions

This paper presents a design-driven system for visualising and monitoring the Italian edition of the contest Wiki Loves Monuments. The project demonstrates how structured data and open infrastructures can support cultural heritage initiatives, not only by enabling quantitative tracking but also by fostering engagement, coordination, and visibility. Through co-design activities, the design team identified key user needs and translated them into a data-driven and visualisation system based on Wikidata and Commons, whose infrastructure is composed of a data backend and a frontend UI capable of supporting both strategic monitoring and data curation via a geographical and a list view. The system highlights the value of Wikidata as a central data source and confirms the potential of visualisation tools to make complex data accessible and actionable for different stakeholders. It also contributes to the broader Wikimedia ecosystem by filling key gaps in the Italian context, such as the lack of tools for typological analysis or longitudinal comparison across campaign editions.

By synthesising these technical and strategic insights, the contribution offers a replicable model for future efforts to leverage Wikidata in projects related to cultural heritage and volunteer engagement. On one hand, it provides examples of data integration and curation strategies; on the other, it exposes the need for co-design methods to accommodate and anticipate multiple stakeholders' needs and perspectives. Importantly, the project raises critical questions about participation and governance in tool development within open communities. While the approach produced effective outcomes, future iterations must attempt to better involve the wider volunteer base, which is diverse and skilled. Still, it is difficult to fully engage through small-group co-design. The risk of perceived top-down decisions remains, particularly in a landscape where many volunteers actively collaborate in the creation of tools and freely accessible knowledge. In this regard, the project offers both a working prototype and a methodological reflection. It calls for further exploration into collaborative design practices that align with the distributed, community-led nature of Wikimedia, and into strategies that reconcile usability, inclusivity, and openness in shared digital infrastructures.

8. Acknowledgements

The authors would like to acknowledge the contributions of all volunteers involved in Wikimedia projects, with particular appreciation for those engaged in Wiki Loves Monuments Italy, the participants of the 2022 wiki-expeditions to Calabria, Marco Chemello and Piergiovanna Grossi. Participants in the co-design workshop include: Iolanda Pensa, Catrin Vimercati, Marta Arosio, and Dario Crespi (for Wikimedia Italia); and Francesca Gheli, Alessandro Quets, Andrea Benedetti, Angeles Briones, Michele Mauri, Elena Aversa, and Tommaso Elli (for DensityDesign, Design Department, Politecnico di Milano). The back-end application described in this paper was developed by the Inmagik development studio. The frontend was designed and implemented by Elli, with support from Benedetti, Mauri, and Briones. The sections Background, Methodology, Design and Implementation, and Discussion were written by Elli; the remaining sections were co-authored by all contributors.

References

Azizifard, Narges, Lodewijk Gelauff, Jean-Olivier Gransard-Desmond, Miriam Redi, and Rossano Schifanella. 2023. "Wiki Loves Monuments: Crowdsourcing the Collective Image of the Worldwide Built Heritage." *Journal on Computing and Cultural Heritage* 16 (1): 1–27. https://doi.org/10.1145/3569092.

Bertacchini, Enrico, and Iolanda Pensa. 2023. "Exploring Collaborative Digital Heritage Communities: A Quantitative Assessment of Wiki Loves Monuments in Italy." *Il Capitale Culturale. Studies on the Value of Cultural Heritage* 28: 129–50. https://doi.org/10.13138/2039-2362/3197.

Dörk, Marian, Boris Müller, Jan-Erik Stange, Johannes Herseni, and Katja Dittrich. 2020. "Co-Designing Visualizations for Information Seeking and Knowledge

Management." *Open Information Science* 4 (1): 217-35. https://doi.org/10.1515/opis-2020-0102.

Kawulich, Barbara B. 2005. "Participant Observation as a Data Collection Method." *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 6 (2). https://doi.org/10.17169/FQS-6.2.466.

Morelli, Angela, Tom Gabriel Johansen, Rosalind Pidcock, et al. 2021. "Co-Designing Engaging and Accessible Data Visualisations: A Case Study of the IPCC Reports." *Climatic Change* 168 (3–4): 26. https://doi.org/10.1007/s10584-021-03171-4.

Sanders, Elizabeth B.-N., and Pieter Jan Stappers. 2008. "Co-Creation and the New Landscapes of Design." *CoDesign* 4 (1): 5–18. https://doi.org/10.1080/15710880701875068.

Seidman, Irving. 2006. *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences*. 3rd ed. New York: Teachers College Press.

Vrandečić, Denny, and Markus Krötzsch. 2014. "Wikidata: A Free Collaborative Knowledgebase." *Communications of the ACM* 57 (10): 7885. https://doi.org/10.1145/2629489.

Walny, Jagoda, Christian Frisson, Mieka West, et al. 2020. "Data Changes Everything: Challenges and Opportunities in Data Visualization Design Handoff." IEEE Transactions on Visualization and Computer Graphics 26 (1): 12-22. http://dx.doi.org/10.1109/TVCG.2019.2934538.

A Wikibooks-Wikidata Integration for Crowdsourced Curatorship at the Museu Paulista in Brazil

João Alexandre Peschanski, Solange Ferraz de Lima

Abstract: The article presents an initiative for the digital dissemination of the Museu Paulista collection, based on the reuse of data available on WIkidata for the creation of a Wikibook in Portuguese version of the Open Book platform, where works are written collaboratively and published under a free licence. The wikibook *The photographs of Guilherme Gaensly in the collection of the Museu Paulista* features a set of 140 photographs and postcards by Gaensly, especially urban landscapes. In the technical infrastructure of the wikibook, each page was created by just one person, but the information on it comes from edits made either directly on Wikidata or in applications that indirectly generate edits on Wikidata. The page also contains a section on participatory curatorship, a section on connections with other images and two sections in the footer with contextual references to Gaensly's work and the Paulista Museum's GLAM-Wiki.

Keywords: Wikibooks, GLAM-Wiki, Digital dissemination.

1. Introduction

This article describes and analyses the experience of digital dissemination of the collection of the Paulista Museum of the University of São Paulo on the Portuguese version of Wikilivros, an open book platform where works are written collaboratively and published under a free licence (Silva 2017). We discuss the socio-technical infrastructure of the Wikibook entitled *As fotografias de Guilherme Gaensly* no acervo do Museu Paulista (2020a) (*The photographs of Guilherme Gaensly in the collection of the Paulista Museum*), produced in the context of the digital dissemination strategy of Paulista Museum/ Universidade de São Paulo. We discuss how the wikibook experience can point paths to shared curatorship of collections preserved in museums through interaction with Wikidata and, thus, generate knowledge collaboratively.

The analysis takes into account the Paulista Museum's strategic planning for digital dissemination in the 21st century, characterized by an "emphatic commitment" to making data and images from its collection freely and openly available on collaborative platforms, such as those of Wikimedia (Carvalho et al., 2021). This commitment took place in the context of research into dissemination practices aligned with inclusive and representative curatorial practices,

João Alexandre Peschanski, Wikimedia Brasil, Brazil, joalpe@wmnobrasil.org, 0000-0002-2352-1787
Solange Ferraz de Lima, University of São Paulo, Brazil, sflima@usp.br, 0000-0002-9411-1988

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

João Alexandre Peschanski, Solange Ferraz de Lima, *A Wikibooks-Wikidata Integration for Crowdsourced Curatorship at the Museu Paulista in Brazil,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.35, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 303-310, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

with a focus not only on visualizing the collection, but with the expectation that digitization could be incorporated into a broader circuit involving research and participatory curatorship, a "meeting of curatorial practices with digital culture" (Lima and Carvalho 2021).

With the partnership with Wikimedia's local affiliate, Wiki Movimento Brasil (Fontenelle; Peschanski 2021), the Paulista Museum took a step towards a more interactive way of making the documentary treatment of its collections available through GLAM-Wikis. Between 2017 and 2024, 33,562 images from the museum's collection were uploaded in Wikimedia Commons (and metadata in Wikidata), used on 21,762 pages on wiki platforms and viewed an average of 1.7 million times per month (GLAM Wiki Dashboard 2024). The Paulista Museum's GLAM-Wiki was the source for the Wikilivros project in Portuguese.

The wikibook *The photographs of Guilherme Gaensly in the collection of the Paulista Museum* was produced as part of the New Paulista Museum Wikipedia Initiative (2020) organized by Wiki Movimento Brasil and the Paulista Museum, in partnership with the Banco do Brasil Foundation and carried out by the University of São Paulo and the Foundation for the Support of the University of São Paulo (Wiki Movimento Brasil 2020c). It was launched in the live "São Paulo photographic edition marathon - Paulista Museum of USP" on YouTube on 8 May 2020 (Wiki Movimento Brasil 2020a). It presents a set of 140 photographs and postcards by Gaensly, especially records of São Paulo and portraits taken in the studio, and encourages readers to collaborate in describing the images (The photographs of Guilherme Gaensly in the collection of the Paulista Museum 2020b; 2020d) using digital tools. There are nine parts in Gaensly's wikibook: introduction; seven parts of image sets; and afterword. The introduction is divided into a cover and a presentation text entitled "The Guilherme Gaensly Collection of the Museu Paulista and the Wikipedia Initiative and the afterword contains a methodological text, "On the development of this semantic book".

2. Knowledge production

The wikibook is presented in its introduction as "the first semantic wikibook of the Wikimedia projects" (The photographs of Guilherme Gaensly [...], 2020b) and the content of the pages is generated by transposing structured data into natural language. The transposition takes place through presets using MBabel technology (Azzellini et al., 2019; Guilherme Gaensly's photographs [...], 2020e) and he information repository is Wikidata. So, a semantic wikibook could be define as a wikibook created collaboratively using web semantic technologies like RDF, OWL and SPARQL.

In the technical infrastructure of the wikibook about Gaensly, each page was created by just one person, but the information on it comes from edits made either directly on Wikidata or in applications that indirectly generate edits on Wikidata. The page created initially contains an image of Gaensly and its title, as well as predefined text and data sheets that either have gaps to be filled in ac-

cording to the specifics of each image, or are generated as information into Wikidata. The page also contains a section on participatory curation, a section on connections with other images and two sections in the footer with contextual references to Gaensly's work and the Paulista Museum's GLAM-Wiki.

Figure 1 – Clockwise from top to bottom, the cover, part of the table of contents and two clippings from an internal page in the wikibook.

Figure 1 shows two partial reproductions of a page with an image. Both the information in the caption and in the image, metadata are transposed from Wikidata (São Paulo-Lavandeiras 2018) and complete gaps in a predetermined textual structure. The same structure is used for all images. As new image de-

scriptors are added to the Wikidata item, they are automatically transposed onto the page. In the caption, information about what is depicted in the image is transposed, linking to Wikipedia pages when they exist (these are the underlined descriptors in the caption, such as Várzea do Carmo, which refers to the corresponding Wikipedia page). The image file uses Wikidata to reproduce the information from the Museu Paulista's own metadata publication. There are also additions, such as the reference to the image's identifier in Google's Arts and Culture project (Google Arts and Culture, 2024), which make this Wikibooks page a point of convergence for references to sources on this work.

The image curation section, next to the metadata card, makes a call to those reading the wikibook ("Participate in the curation of this image"), so that they can also contribute to the description of the image. This call, in a way, proposes a conversion from consumer to knowledge producer. The forms of contribution are facilitated by links to five apps, shown in Table 1, with the following information: name, link to the app and link to the example image "São Paulo – Lavandeiras", purpose, type of Wikidata edition and developer (Guilherme Gaensly's photographs in the Paulista Museum collection, 2020d). The links to the apps on each Wikibooks page lead directly to the means of contributing to Wikidata about the corresponding image.

The apps have two functions:

1) They offer a simpler way of editing in Wikimedia projects, especially Wikidata. Each item in Wikidata, in its basic mode, has the appearance of a form, with manual completion that can be considered tedious. Apps aim to improve the user experience.
2) They direct edits to certain gaps in the description of objects, in particular descriptors and their qualifiers, thus serving as a means of guidance for curators.

One of the challenges in curatorship based on decentralised applications is their maintenance. As each tool has their responsible developers, sometimes even volunteers like Wikidata Image Positions and Tabernacle, it is up to them to ensure that the tools work. There are ways of requesting adjustments to the applications in the eventual problem, but the order of priority depends on the availability and organisation of the developer's time.

Following on from the wikibook, there is a section with suggestions for other Gaensly images. The suggestions are based on the descriptors identified in each image and are dynamically updated by a robot as new descriptors are included in Wikidata.

There are two fixed sections at the bottom of each page. On the left, there are Wikidata queries that seek to place the image presented on each page in the context of the Paulista Museum's Guilherme Gaensly Collection, for example in the timeline of the artist's production. In the section on the right, there are general references to the museum's initiative page on Wikimedia and to the museum's institutional page.

Each page of the wikibook has become a space for curatorial organisation and dissemination. The information produced in a decentralised way about Gaensly's works (for example in apps or on the Wikidata page itself) is transposed onto the pages of the wikibook and contributes to semantic collaborative writing through natural language processing. The wikibook also offers a menu of editing possibilities that guide and facilitate collaboration and, as information about the photographs is included in Wikidata, it appears, for example, in tables and information boxes on Wikipedia and Wikimedia Commons, and can also be found in third-party projects, including the museum's own digital repository.

3. Results

The pages of the wikibook were viewed 19,734 times between its launch and February 2025 (Pageviews 2025a).

Figure 2 shows the flow of content production on Wikidata items related to the Guilherme Gaensly Collection. There is a continuous editorial engagement with the collection. There are two identifiable clusters—the publication of the collection on Wikidata in 2018 and the launch of the wikibook in 2020—and apart from these catalysing moments of collaboration, contributions to the collection's items appear to be self-motivated. Peaks in editing can be seen, for example in the first quarter of 2023, when there were 394 edits, possibly carried out in the context of coordinated initiatives to improve the content of this collection on Wikidata.

It is significant that the properties to be fed through the applications available on the wikibook are at the top of the number of edits made. The edits of the "Museu do Ipiranga: Povo Conta", "Wikidata Art Depiction Explorer" and "Wikidata Image Positions" apps occur in the "portrait" property, which had 778 edits by March 2024. The "description" property, which is written in potentially hundreds of languages, can be understood as a measure of international contribution or internationalisation of content. This property has been edited 692 times on items in the William Gaensly Collection up to March 2024 and the Tabernacle app allows it to be edited quickly.

4. Final considerations

This wikibook is based on an infrastructure for the semantic collaborative writing of the Gaensly catalogue of works. The first batch of contributions came from the museum itself, when the collection's description sheets were transferred to Wikidata. The wikibook became a kind of hub for various contents and resources that facilitate and guide collective curatorship, thus creating a complementary curatorial circuit between the museum and the interested public, with free digital socio-technical mediation, identified as an "advanced" platform for structuring the contribution of diverse audiences to open educational resources (Sigalov and Nachmias 2023). Each page of the wikibook is not only a means of enjoying Gaensly's photographs, but mainly a means of directly producing knowledge, with a high social impact. In addition, technology is used on the

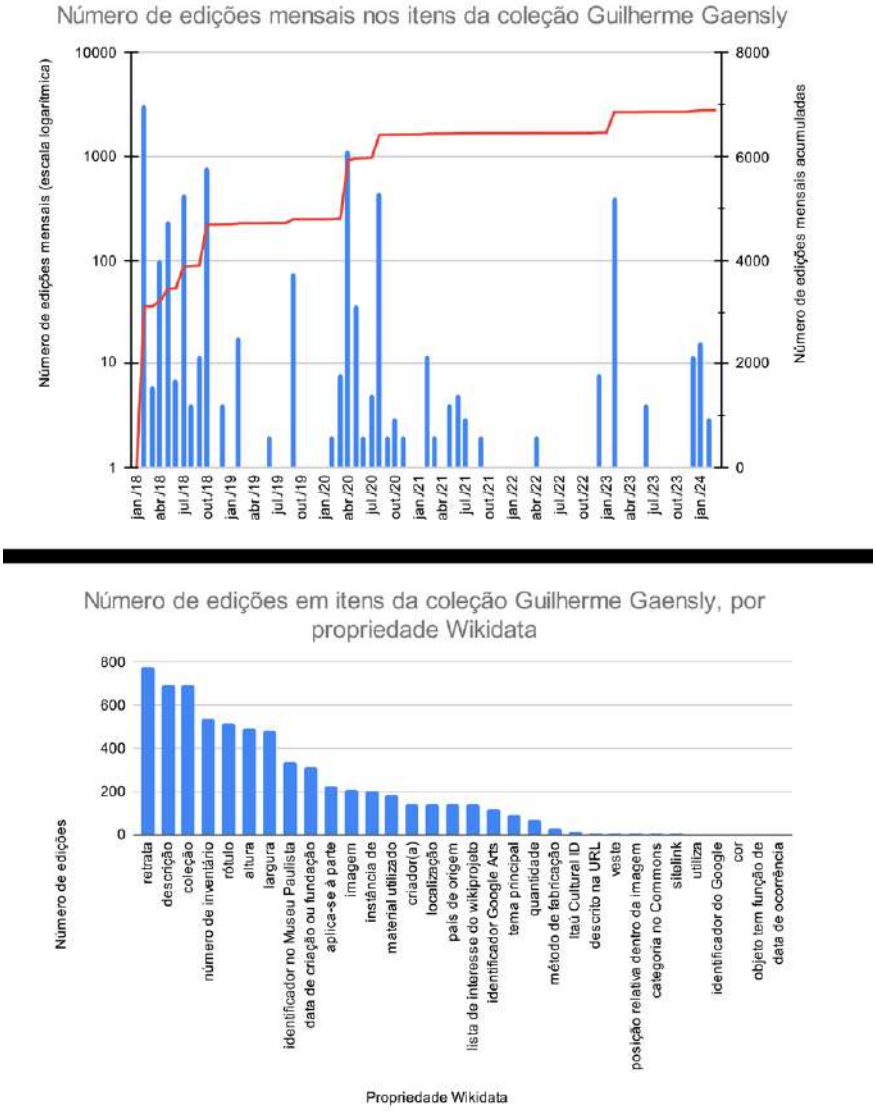


Figure 2 – Above, a graph showing the number of edits made each month (on a logarithmic scale) and the cumulative total of edits to Wikidata items related to the Guilherme Gaensly Collection, from the first quarter of 2018 to 2024. Below is a graph showing the most edited fields in Wikidata items related to the Guilherme Gaensly Collection, up to March 2024.

pages to direct the production of knowledge to specific contribution gaps (Sigalov and Nachmias 2023).

The experience reported here allows us to infer some important general principles related to a policy of producing open educational resources in the field of museums, namely:

- The need to integrate digital actions into strategic planning, with budget forecasting on a permanent basis. In other words, the digital maturity of a museum is associated with the integration of actions of this type into the daily lives of the teams, involving the resulting costs. The wikibook on Guilherme Gaensly's production, with its sophisticated socio-technical infrastructure and mobilising tools developed in a decentralised way, gives us a measure of the investment not only in equipment, but above all in training human resources.
- By making collections available on networks, the proposal for shared curation is expanded and outside institutional control. Therefore, it is necessary to invest in socio-technical infrastructure to guarantee the return of widely shared curation to the institution, integrating it into the museological documentation, which is constantly being updated.

5. Funding and acknowledgements

João Alexandre Peschanski's research is supported by FAPESP, project 2013/07699-0. We would like to thank Éder Porto and Lucas Belo for their technical support.

References

Alves, É. P., P. R. Burley, and J. A. Peschanski. 2021. "Structuring Bibliographic References: Taking the Journal Anais do Museu Paulista to Wikidata." In *Wikipedia and Academic Libraries: A Global Project*, 260–76. [s.l.]: Maize Books. https://doi.org/10.3998/mpub.11778416.ch17.en

Azzellini, É. C., J. A. Peschanski, and F. J. da Paixão. 2019. "The Potential of Structured Narratives for Computational Journalism: Journalistic Competences in the Elaboration of Texts Generated with Databases." *Texto Livre* 12 (1): 138–52. https://doi.org/10.17851/1983-3652.12.1.138-152.

Carvalho, V. C. D., P. C. G. Marins, and S. F. D. Lima. 2021. "Curatorship in History Museums." *Anais do Museu Paulista: História e Cultura Material* 29: e40.

Fontenelle, G., and J. A. Peschanski. 2021. "The Work of Art in the Age of Digital Convergence." Paper presented at the 41° Congresso Brasileiro de Ciências da Comunicação, Intercom – Sociedade Brasileira de Estudos Interdisciplinares da Comunicação, Joinville, SC.

GLAM Wiki Dashboard. 2024. *GLAM Wiki Dashboard*. https://glamwikidashboard.wmcloud.org/MPUSP.

"Guilherme Gaensly's Photographs in the Paulista Museum Collection." 2020a. *Wikibooks*. https://pt.wikibooks.org/wiki/As_fotografias_de_Guilherme_Gaensly_no_acervo_do_Museu_Paulista.

"Guilherme Gaensly's Photographs in the Collection of the Paulista Museum / The Guilherme Gaensly Collection of the Paulista Museum and the Wikipedia

Initiative." 2020b. *Wikibooks*. https://pt.wikibooks.org/wiki/As_fotografias_de_Guilherme_Gaensly_no_acervo_do_Museu_Paulista/A_Coleção_Guilherme_Gaensly_do_Museu_Paulista_e_a_Iniciativa_Wikipédia.

"Guilherme Gaensly's Photographs in the Paulista Museum Collection. Cover." 2020c. *Wikibooks*. https://pt.wikibooks.org/wiki/As_fotografias_de_Guilherme_Gaensly_no_acervo_do_Museu_Paulista/Capa.

"Guilherme Gaensly's Photographs in the Collection of the Museu Paulista / São Paulo – Lavandeiras." 2020d. *Wikibooks*. https://pt.wikibooks.org/wiki/As_fotografias_de_Guilherme_Gaensly_no_acervo_do_Museu_Paulista/São_Paulo_-_Lavandeiras.

"Guilherme Gaensly's Photographs in the Collection of the Paulista Museum / About the Development of This Semantic Book." 2020e. *Wikibooks*. https://pt.wikibooks.org/wiki/As_fotografias_de_Guilherme_Gaensly_no_acervo_do_Museu_Paulista/Sobre_o_desenvolvimento_deste_livro_semântico.

"Help:Image-Annotator." 2009. *Wikimedia Commons*. https://commons.wikimedia.org/wiki/Help:Image-Annotator.

Lima, S. F. de, and V. C. de Carvalho. 2021. "The Dynamics of Research with Collections in a University Museum." Paper presented at the ANPUH-Brasil – 31° Simpósio Nacional de História, Rio de Janeiro.

Pageviews. 2025. "Analysis of Page Views – Category:Book/The Photographs of Guilherme Gaensly in the Collection of the Museu Paulista." *Pageviews*. https://pageviews.wmcloud.org/massviews/?platform=all-access&agent=user&source=category&start=2020-01-01&end=2025-02-20&subjectpage=0&subcategories=0&sort=views&direction=1&view=list&target=https://pt.wikibooks.org/wiki/Categoria:Livro/As_fotografias_de_Guilherme_Gaensly_no_acervo_do_Museu_Paulista.

"São Paulo – Lavandeiras." 2018. *Wikidata*. https://www.wikidata.org/wiki/Q49903896.

Sigalov, S. E., and R. Nachmias. 2023. "Investigating the Potential of the Semantic Web for Education: Exploring Wikidata as a Learning Platform." *Education and Information Technologies* 28: 1–50.

Wiki Movimento Brasil. 2020a. "Editing Marathon 'São Paulo Fotográfica' – Museu Paulista da USP." YouTube video. https://www.youtube.com/watch?v=A3piYwkamyk.

Wiki Movimento Brasil. 2020c. "Iniciativa Wikipédia Novo Museu do Ipiranga." PDF file. https://commons.wikimedia.org/wiki/File:Iniciativa_Wikipédia_Novo_Museu_do_Ipiranga.pdf.

POSTER

Wikidata and Research: A Decadal View

Daniel Mietchen

Abstract: The breadth, depth and diversity of interactions between the research ecosystem and the ecosystem around Wikidata have evolved considerably over the years. These interactions include, e.g., research about Wikidata and Wikibase matters as well as research-related content, infrastructure or activities with a Wikidata or Wikibase component. This contribution is intended for researchers of any background as well as science communicators and other stakeholders in the research landscape. It aims to highlights key facets of these interactions and their evolution, considering developments within and beyond both ecosystems, such as the trends towards increased openness in research workflows or towards using Wikibase as a platform for collaborative and multilingual curation of structured data.

Keywords: Wikidata Status Updates, Wikidata and the research landscape, poster design, open science, metascience.

The poster "Wikidata And Research As Reflected in Wikidata Status Updates 2012-2025" is an attempt to capture both the multitude and the evolution of interactions between Wikidata and the research landscape and to portray them (using portrait format) in a way that could be presumed to facilitate discussions at the conference's poster session. While printed on paper and pinned to a poster wall, it was complemented by a QR code pointing to a digital version of the poster on Zenodo, where it is available under the terms of the CC0 1.0 Universal Public Domain Dedication via https://doi.org/10.5281/zenodo.15611921 . The same QR code was provided on four smaller additional sheets of paper, with the intent of their arrangement being that poster visitors would be able to see (and scan) at least one version of the code even if people would be standing in front of that poster wall, as is expected during poster sessions. The QR code was not included in the poster itself because it was assumed that most engagement with the poster would not take place in person during that poster session but online, where anyone can explore it at their own leisure and where assistance by QR codes is of little benefit.

The core content of the poster is a detailed yet condensed survey of the research-related content found within the weekly Wikidata Status Updates published from April 2012 (the project's inception) until early June 2025. The Status Updates themselves constitute a large, continuous corpus of project documen-

Daniel Mietchen, FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Germany, daniel.mietchen@wikipedia.de, 0000-0001-9488-1870

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Daniel Mietchen, *Wikidata and Research: A Decadal View,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.37, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 313-316, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

tation, serving as a regular but lightweight account of development, community activity, technical milestones, and external engagement. The poster thus chronicles how the nascent idea of Wikidata as a centralized, multilingual, and freely accessible knowledge base for Wikimedia projects gradually evolved into Wikidata serving an additional role as a recognized, indispensable tool and object of study for the global scientific and academic community, i.e. in the center of the conference's scope. By specifically focusing on the research-related elements within this sizeable collection, the poster represents an act of distillation.

Its design reflects the distillation process, with the poster's section structure closely following that of the Status Updates, with thematic divisions such as "Publications," "Data Integration," "Tool Development," or "Community Outreach". Within each section, the content is provided in chronological order, which allows the viewer to trace the developmental trajectory of Wikidata's impact on the research ecosystem.

In its earliest years, the research-related entries have been relatively few and far between, focusing on foundational scope and data modeling discussions, early tools and proof-of-concept projects, or initial academic papers announcing and describing the platform's potential. As the timeline progresses toward 2025, the volume and complexity of the entries has grown, reflecting the growing complexity of interactions between Wikidata and the research ecosystem, which also gave rise to the conference. This chronological progression charts the shift from early adoption to systemic integration: from initial collaborations involving a small number of individual researchers to major institutional partnerships, from a handful of publications citing Wikidata to its established role as a key data source for entire fields like digital humanities, biomedical research, and scientometrics, and from basic data ingestion to the development of sophisticated research tools and analytical queries built directly atop the evolving knowledge graph. The poster thus captures the essence of this period and the deepening integration between this community-driven platform and institutionalized academic inquiry.

A key feature of the poster is its design as an interactive document, where the paper version and the presence of the presenter allows (as usual) for immediate and direct in-person exchanges on-site, whereas further interactivity can be (and indeed was) leveraged on the spot through accessing the digital file itself, which is also possible at any other occasion. The digital version is full of hyperlinks that connect an entry in the poster's timeline with the corresponding Wikidata Status Update. These hyperlinks are rendered in blue and visible in the paper version as well, which enables the discussion in front of the physical poster to leverage both the breadth of the landscape overview provided by compiling a decade of activities into a one-page format with in-document search functionality and the depth of examining individual Status Update entries linked from the poster and linking, in turn, to yet more granular details. In order to enable others to build on this effort more readily, the digital version of the poster was not only shared in PDF format but also in form of the original LaTeX source file, which can be edited and updated as anyone may see fit.

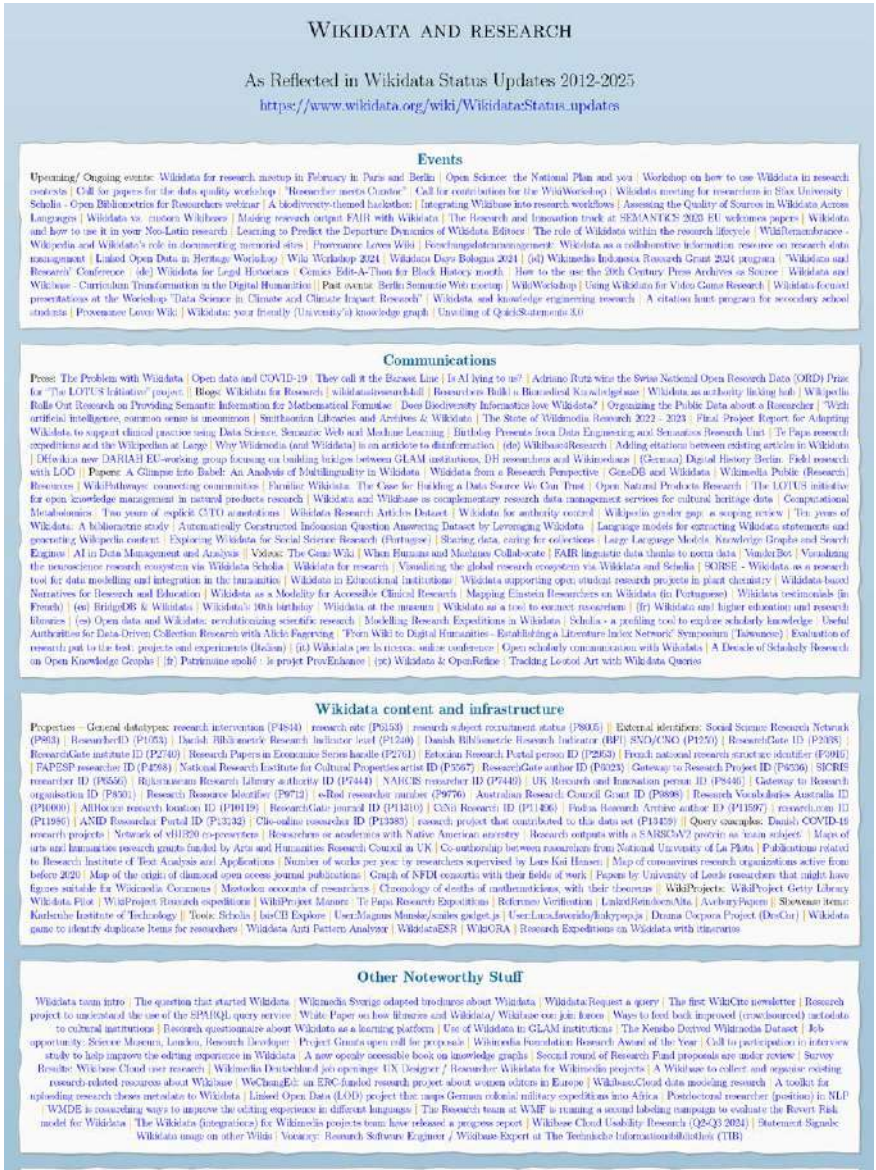


Figure 1 – Poster submitted at "Wikidata and Research". Conference Proceedings, Florence, 5-6 giugno 2025.

The granularity of sharing was the focus of another poster that was presented alongside, highlighting the other end of the "Wikidata and Research" landscape: "Open Scholarly Metadata for Sharing the Nuances of Research Processes". This one was originally designed for another conference held a week earlier in Bologna, which is available via https://doi.org/10.5281/zenodo.15512071 , with its abstract additionally available via https://doi.org/10.5281/zenodo.16367857 . Instead of the link-heavy approach outlined above, it took a more visual approach inspired by the BetterPoster recommendations, and it also stood out because it was the only poster (at both events) that was presented in landscape rather than portrait format. It made the case that research processes are often documented in a much more detailed fashion than what is finally shared in official publications, yet sharing more of these details - and with structured metadata - can have benefits in terms of enhancing various facets of the research landscape, from reproducibility to teaching, from reuse to community engagement and beyond. The layout of the poster emphasized in-person discussion, and the QR code pointing to the Zenodo version was embedded to facilitate follow-up. As with the other poster, hyperlinks were marked in blue, and some of these links went to Wikidata and other parts of the Wikimedia ecosystem, thereby closing the loop to the scope of the Wikidata Status Updates poster.

Taken together, the two posters represent perspectives from both ends of the "Wikidata and Research" landscape, and they illustrate different approaches to visual versus textual discussion primers, on-site or online reception, or the granularity of sharing.

References

Mietchen, D. 2025. "Wikidata And Research as Reflected in Wikidata Status Updates 2012-2025." <https://doi.org/10.5281/zenodo.15611921>.

The DAMEIP Project: Data and Metadata to Implement Peer Review

Rossana Morriello, Donatella Selva, Annantonia Martorano, Silvia Pezzoli, Lorenzo Sergi, Andrea Girometti

Abstract: The DAMEIP (Data and Metadata to Implement Peer Review) research project, funded by the University of Florence for the years 2025–26, will be carried out by an interdisciplinary team with expertise in library and information science as well as sociology. The project aims to examine the dynamics and practices of peer review in journals published by Florence University Press (FUP). With full respect for privacy and ethical considerations, it seeks to identify and analyze recurring patterns, procedures, quantitative indicators, and timelines of the peer review process, with particular attention to gender differences. Peer review is one of the core systems of research evaluation, operating at two stages of the research cycle: prior to the publication of research results in journals or other outlets, and subsequently within research assessment frameworks implemented by national agencies such as ANVUR. This model is applied predominantly in the humanities and social sciences (HSS), and to a lesser extent in STEM disciplines. As part of its activities, the project will also update the essential metadata of FUP journals in Wikidata, once again with a focus on gender representation. Wikidata is a key reference infrastructure for projects and applications that rely on metadata, ranging from the simplest to the most complex, and increasingly connected to artificial intelligence systems. Ensuring high-quality data in Wikidata is therefore essential for effective knowledge organization in the digital environment and for scholarly communication. Enriching publications with accurate metadata also enhances their global discoverability and facilitates the assignment of reviewers in internal journal workflows.

Keywords: Peer review, research evaluation, scholarly communication, metadata, gender representation.

Peer review is one of the fundamental activities in support of some of the structures on which science rests (publications, conferences, etc.). This is the characteristic that distinguishes a scientific journal and represents a quality filter, as well as a process aimed at the overall improvement of the journal and the advancement of science. Peer review differentiates a scientific journal, whose articles communicate the original results of a research, from "non-scientific" journals and other types of serial publications such as newspapers, newsletters, generic journals and journals with commercial content.

Rossana Morriello, University of Florence, Italy, rossana.morriello@unifi.it, 0000-0002-9990-9243
Donatella Selva, Luiss University, Italy, dselva@luiss.it , 0000-0002-9051-7876
Annantonia Martorano, University of Florence, Italy, annantonia.martorano@unifi.it, 0000-0001-8795-447X
Silvia Pezzoli, University of Florence, Italy, silvia.pezzoli@unifi.it, 0000-0002-3958-1450
Lorenzo Sergi, University of Tuscia, Italy, lorenzo.sergi@unitus.it, 0000-0001-5163-7392
Andrea Girometti, University of Florence, Italy, andrea.girometti@unifi.it, 0000-0002-2389-9085

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Rossana Morriello, Donatella Selva, Annantonia Martorano, Silvia Pezzoli, Lorenzo Sergi, Andrea Girometti, *The DAMEIP Project: Data and Metadata to Implement Peer Review,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.38, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 317-321, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

Subsequent research by other scientists is based on the results of research published in journals, after scrutiny aimed at certifying its scientific and integrity and therefore the peer review process is one of the pillars of the scientific method.

Peer review is undergoing many challenges, old and new: can be affected by bias and subjectivity (institutional or geographic bias, gender or racial bias, confirmation bias, editor influence bias), which can skew analyses, sometimes resulting in unfair or inconsistent evaluations; it slows down publication process; the peer review process can be time-consuming, leading to delays in the publication of important research; there is lack of transparency when peer review is not open; a lack of transparency and accountability in peer review may cause inconsistent feedback, impacting scientific reliability; more recently, artificial intelligence is impacting peer review. Moreover, journals often do not standardize review criteria, some focus on novelty, others on methodology, clarity, or significance.

Published studies show evidence of gender bias, which is often subtle and systemic, not always obvious in individual cases, but cumulative effects can significantly impact women's careers and scientific output. Some studies found that papers with female first or corresponding authors sometimes receive more critical reviews or lower acceptance rates compared to male authors; others that women's work is cited less frequently, which can influence reviewers' perceptions of "impact" or "novelty." In addition, gender bias is not uniform: it can vary by discipline, journal, and region.

The project aims to examine the dynamics and practices of peer review in journals published by Florence University Press (FUP) in order to strengthen it. Also, the project will explore in depth the role of university presses in scholarly publishing with the analysis of qualitative data collected through surveys.

The current phase is the analysis of data and the selection of samples to be studied. The selection was guided by the two main project objectives: analysis of data on journal refereeing/peer review; and uploading data related to journals, articles, and authors in Wikidata. For these analyses, the data must be as homogeneous as possible in terms of handling, collection methods, and completeness. From this perspective, the entirety of FUP periodicals—which can be traced back to 59 journals—presents an extremely varied characterization in terms of: distribution and presence in scientific databases (such as Scopus and Web of Science); refereeing methods; use of applications for the publication process; whether they belong to bibliometric sectors or not; the ANVUR scientific recognition process and reference areas/SSD; the history of the journal (year of establishment, evolution of its title and editorial context); periodicity; online or print publication; first publication with FUP (whether originally launched there or later acquired by the publisher); total number of issues, volumes, and articles; average trends in the number of authors and reviewers associated with individual contributions; and editorial and scientific boards (geographical and institutional origin, gender and numerical composition). This will allow to proceed in the analysis:

- With full respect of privacy and anonymity, map the dynamics and key characteristics of peer review, such as reviewer–author relationships (for example, who reviews whom?),
- Examining the data through a gender-informed lens,
- Enhancing the transparency of the process through adding data in Wikidata,
- Improve discoverability and findability of FUP publications through the work on Wikidata,
- Improve quality and reliable dataset in Wikidata for the use in scholarly community,
- Quality dataset feed also fair results in the applications of AI.

A first page for the project was created in Wikidata[1] where all significant information and data will be added during the project. The next steps will be as follows:

- For the 59 journals with corresponding information to be fully metadata-tagged for the creation of a record (one per journal) to be uploaded to Wikidata. Each entry should accurately reflect the key data previously collected and presented.
- Selecting a sample of periodicals with corresponding information to be metadata-tagged in detail for the creation of a record (one per article) to be uploaded to Wikidata. The selection will consider some main journals related to the two scientific areas most closely connected to the fields of interest of the DAMEIP project according to the area classification of disciplines of the Italian Ministry of University (Area 11 and Area 14).
- Selecting a sample of periodicals with corresponding information to be metadata-tagged in detail for the creation of a database on peer review practices, including specific information about referees (geographical and institutional affiliation, gender, and numerical data). This is the most complex to identify due to the presence of certain variables and complications. Data on refereeing/peer review are characterized by a degree of 'confidentiality'—especially in double-blind peer review, where the author does not know the identity of the reviewers and vice versa—and are therefore treated as 'sensitive data,' with a high ethical standard, by scientific committees and publishers. Consequently, three aspects appear necessary: the completeness of refereeing information, held by a specific and single publishing house; the quality of the data and the completeness of its processing; and the provision of the data, which is 'regulated' by a specific ethical commitment—ensuring the confidentiality of detailed information—on the part of the publisher.

[1] The DAMEIP Project, <https://www.wikidata.org/wiki/Q135485829>.

Figura 1 – Poster presentato al convegno Wikidata e la ricerca, Firenze, 5 giugno 2025.

Criteria for the selection of significant samples will be discussed in the research group to guarantee completeness of refereeing information, data quality and processing consistency, and regulated data access. Also, careful consideration must be related to the many challenges that adding data in Wikidata for peer review has, such as

- confidentiality: most peer reviews are anonymous, so adding reviewer-specific data is often not possible,
- granularity: Wikidata works best for structured facts (journal, article, type of review), less so for subjective outcomes (review quality, reviewer comments),
- property gaps: some relationships (e.g., "reviewed by") may need a proposal for a new property.

In essence, a very crucial challenge for the project team is that handling this type of data requires a careful balance between accessibility for analysis and adherence to ethical and confidentiality standards.

Wikidata and the Thesaurus Nuovo Soggettario: Together to Build a Domain Ontology in the Field of Photography

Silvia Bruni, Alida Daniele, Fabrizio Nunnari, Elena Cencetti, Elisabetta Viti, Francesca Palareti

Abstract: The project aims to build a domain ontology on photography in Wikidata, in dialogue with the Nuovo soggettario of the National Central Library of Florence, the official Italian subject indexing tool for resources of various kinds. In the Nuovo soggettario, terms are organized in a hierarchical structure (from the most general to the most specific), while in Wikidata, entities are defined through their properties and relationships. Aligning conceptual hierarchies and enriching both databases with new terms and links to other Knowledge Organization Systems (KOSs) highlights the value of participatory semantic modeling to enhance interoperability between tools and cultural resources.

Keywords: Nuovo soggettario, ontology, searchability, accessibility, interoperability, reusability of tools and information resources.

Since 2018, the Social Sciences Library has set up a Wikistation to promote Wikimedia projects within the University of Florence. A Wikidata working group has also been formed with the aim of deepening the study of ontology and its application in the field of research (Bruni 2025).

In 2024, the UNIFI Wikidata group began collaborating with the National Central Library in Florence to develop a discussion on methods and tools for organising knowledge, with a particular focus on thesauri, classifications and ontologies (Kless et al. 2012). A project was launched to refine an ontology in the domain of photography in Wikidata, starting from the terminology of the *Thesaurus Nuovo Soggettario* (ThNS). Both thesauri and ontologies are structured formal representations designed to organize information and data on a given subject, defining concepts expressed by terms or entities, their semantic relations, reference categories, and links. Both use a standardized, machine-

Silvia Bruni, University of Florence, Italy, silvia.bruni@unifi.it, 0000-0001-5461-8226
Alida Daniele, University of Florence, Italy, alida.daniele@unifi.it,
Fabrizio Nunnari, University of Florence, Italy, fabrizio.nunnari1@unifi.it,
Elena Cencetti, Ales S.p.A., Italy, elena.cencetti@unifi.it,
Elisabetta Viti, Central National Library of Florence, Italy, elisabetta.viti@cultura.gov.it, 0000-0002-7627-4062
Francesca Palareti, University of Florence, Italy, francesca.palareti@unifi.it

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Silvia Bruni, Alida Daniele, Fabrizio Nunnari, Elena Cencetti, Elisabetta Viti, Francesca Palareti, *Wikidata and the Thesaurus Nuovo Soggettario: Together to Build a Domain Ontology in the Field of Photography,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.39, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 323-329, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

readable language. While a thesaurus is created to index resources of various types (such as texts, images, websites, etc.) within catalogs and databases, ontologies aim to schematically represent knowledge domains for computational processing (ISO 2013; 2011; Studer et al. 1998, 184–88).

To make this description more concrete, let us take an example from the field of photography, which we have identified as the focus of the project: a photograph is a digital or physical object, taken with a photographic technique (albumen print, digital photograph, daguerreotype, etc.) by a given device, and may serve a particular function (advertising photograph, travel photograph, etc.).

Defining in a consistent and unambiguous way the terms representing these concepts, the categories to which they belong, and their relationships is fundamental for describing and querying data available in a database. In this case, the ThNS, compliant with ISO 25964, uses hierarchical, associative, and equivalence relationships.

For example, the ThNS models the concept of "advertising photography" (in Italian "*fotografia pubblicitaria*" as follows:

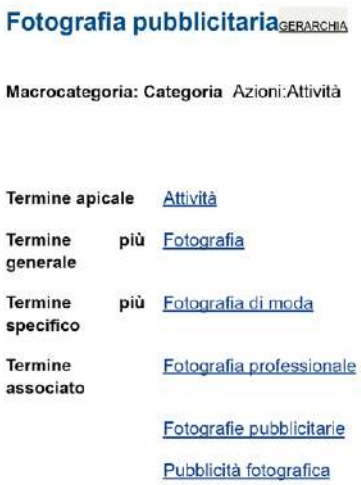


Figure 1 – Structure of the item, in Italian, "fotografia pubblicitaria" ("advertising photography") in the Nuovo soggettario, from the broader term down to the most specific term, along with the associated terms.

In an ontology these relationships are represented by functions and relations, logical rules that constrain meanings. Here is a very simplified representation:

The instances and relations of the ThNS are verified and monitored over time. For this reason, it is a valuable source for controlling and enriching instances, classes, attributes, and logical relations in Wikidata (Cencetti et al. 2025).

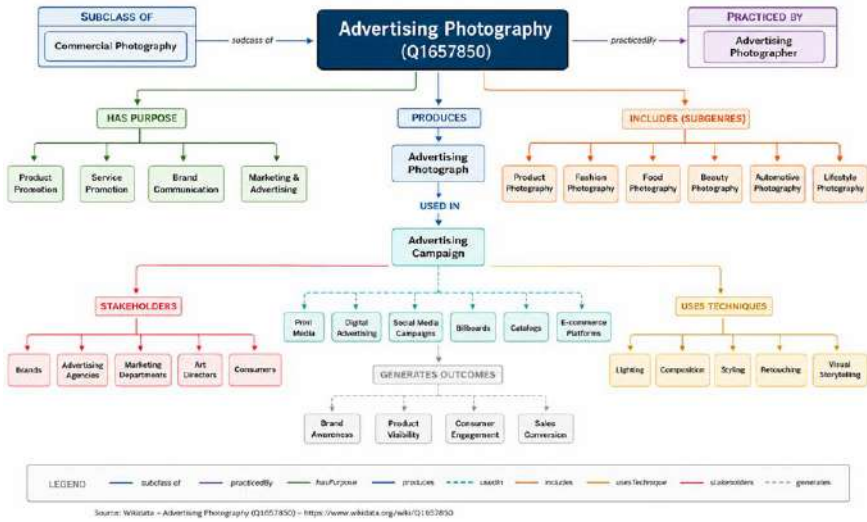


Figure 2 – Ontological model of Advertising Photography (Wikidata Q1657850). The diagram represents the conceptual structure of Advertising Photography as a subclass of Commercial Photography, highlighting its purposes, agents, outputs, application contexts, specialized subgenres, and technical components. Source: adapted from Wikidata Q1657850. Image created by the authors using ChatGPT 5.3 mini.

Figure 3 – Diagram of an ontology of "fotografia pubblicitaria" ("advertising photography"), showing the possible relationships of this item (subclass, context, objectives, equipment, connections).

1. The project

The project focused specifically on the semantic and technical dialogue between Wikidata and ThNS, with the following goals:

1) experimentation with the creation and refinement in Wikidata of a domain ontology on photography, in semantic and technical dialogue with the ThNS,
2) comparative study of methods and tools for importing data into Wikidata,
3) mutual enrichment of both databases with new terminology,
4) mutual enrichment of both databases with links to other KOSs,
5) compliance with FAIR principles (guide lines to make data findable, Accessible, Interoperable and Reusable).

The experimentation had a very practical character and was mainly based on the analysis and comparison, using an inductive method, of data structures, with particular attention to meanings and hierarchical relationships of a significant and representative set of terminology in the domain of photography (about 160 terms selected from the ThNS and and repositories in specific domain (Cencetti 2025; Corti and Gioffredi Superbi 2004).

We verified the correspondence in form and meaning of the terms in the ThNS and in Wikidata entities, in particular adding synonyms (UF/aliases), clarifying definition of concepts (through scope and definition notes; descriptions), reviewing the hierarchical structure; introducing new entities in Wikidata and new terms inTHNS, and finally increasing reciprocal links and links to other foreign national library thesauri (LCSH, RAMEAU, GND, EMBNE), to specialized thesauri (AAT), to archive and museum databases and to Wikipedia articles (Lanza 2022). One example is the ThNS term "*Procedimento all'albumina* (75362)", meaning albumen process, which corresponds in meaning, but not in form, to the Wikidata entity "*stampa all'albume* (Q107387614)", albumen print. In this case, we did not standardize the two entries because the Wikidata entity derives directly from the Wikipedia entry and, as a rule, the label should reflect the entry of the encyclopedia.

In other cases, we eliminated some aliases in Wikidata that seemed irrelevant or misleading. For example, "*ferrotipo"* (tintype) did not belong among the aliases of "*ferrotipia* (Q913381)," nor did "*macro obiettivo"* (macro lens) among those of "*macrofotografia (*Q272444)", macrophotography.

We also corrected some incorrect or redundant statements in Wikidata. For example, in the case of "*procedimento carbro (*Q1035568)", carbro process, which refers to a specific technique, we removed the subclass relationship with "*stampa fotografica* (56055236)", photographic print, because a hierarchical link between a process and an object could not exist.

Instead, in the case of "*ologramma (*Q3139490)", hologram, which was a subclass of both "*positivo* (Q1413646)", positive, and "*fotografia* (Q125191)", photograph, we removed the second link because it was redundant, since "positive" is already a subclass of "photograph."

Figure 3 – Poster submitted at "Wikidata and Research". Conference Proceedings, Florence, 5-6 giugno 2025.

To further enrich the network of connections, we leveraged the associative relations of the ThNS. For example, we linked activities to the person and profession practicing them: "*fotografia scientifica* (Q2586218)", scientific photography, and "*fotografo scientifico* (Q21550957)", scientific photographer, were connected through the statement "practiced by (P3095)" and "field of this occupation (P425)."

2. Conclusions

This analysis and comparison of the two tools has brought many benefits on several aspects. Firstly, it enabled the mutual enrichment of both KOSs with new terminology and links to other knowledge organisation systems (Hodge 2000, 1–22). Secondly, it also highlighted the need for analysis and revision of critical issues that emerged during the process, due to semantic misalignment, in order to optimize the interoperability of the two tools and to foster integrated access to information.

Intervening in the ontology of Wikidata not only means defining more precisely the concepts represented by individual items, but also embedding them in a broader relational system, structured according to principles of semantic coherence and internal logic (Pellizzari di San Girolamo 2024). In this way, individual elements do not remain isolated entities, but become nodes in a dynamic network of connections that enables interoperability between different domains, fostering the construction of an informational fabric capable of organically and effectively linking areas of knowledge that may be very distant from one another (Gruber 1993; Morin 2012, 41–42).

References

Bruni, Silvia. 2025. "Realizzare una Wikistazione e vivere più felici." *Bibelot: notizie dalle biblioteche toscane* 31 (1).

Cencetti, Elena. 2025. "Thesaurus Nuovo Soggettario e Wikidata: analisi di un corpus terminologico relativo all'ambito della fotografia." Tesi di master in Organizzazione e gestione degli archivi, catalogazione e metadatazione di risorse manoscritte, stampate e digitali.

Cencetti, Elena, Camillo Carlo Pellizzari di San Girolamo, and Elisabetta Viti. 2025. "Termini, dati e collegamenti: conversazioni tra il Thesaurus del Nuovo Soggettario e Wikidata." *Imagines* 12: 85–94.

Corti, Laura, and Fiorella Gioffredi Superbi. 2004. *Glossario di terminologia fotografica.* Roma: AIB. https://www.aib.it/aib/lis/lpi13eg.htm.

Gruber, Thomas R. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 5 (2): 199–220. https://doi.org/10.1006/knac.1993.1008.

Hodge, Gail. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files.* Washington, DC: Digital Library Federation, Council on Library and Information Resources.

International Organization for Standardization. 2011. *ISO 25964-1:2011 – Information and Documentation: Thesauri and Interoperability with Other Vocabularies. Part 1: Thesauri for Information Retrieval.* Ginevra: ISO.

International Organization for Standardization. 2013. *ISO 25964-2:2013 – Information and Documentation: Thesauri and Interoperability with Other Vocabularies. Part 2: Interoperability with Other Vocabularies.* Ginevra: ISO.

Kless, Daniel, Ludger Jansen, Jutta Lindenthal, and Jens Wiebensohn. 2012. "A Method for Re-Engineering a Thesaurus into an Ontology." In *Frontiers in Artificial Intelligence and Applications,* edited by Giancarlo Guizzardi and Maureen Donnelly. Amsterdam: IOS Press.

Lanza, Claudia. 2022. *Semantic Control for the Cybersecurity Domain: Investigation on the Representativeness of a Domain-Specific Terminology Referring to Lexical Variation.* Boca Raton: CRC Press.

Morin, Edgar. 2012. *I sette saperi necessari all'educazione del futuro.* 14a rist. Milano: Cortina (Minima/Cortina; 59).

Pellizzari di San Girolamo, Camillo Carlo. 2024. "Conflations and Duplications in Wikidata Items: Causes, Detection, Solutions, and Issues." In *Proceedings of the Wikidata Workshop 2023 Co-Located with 22nd International Semantic Web Conference (ISWC 2023).* Athens: CEUR.

Studer, Rudi, V. Richard Benjamins, and Dieter Fensel. 1998. "Knowledge Engineering: Principles and Methods." *Data & Knowledge Engineering* 25 (1–2): 161–97. https://doi.org/10.1016/s0169-023x(97)00056-6.

The Mare Magnum: A Bibliographic Repertory through the Centuries

Valentina Sonzini

Abstract: The Marucelli's "Mare Magnum" is one of the most interesting and least studied bibliographic repertory in the world. Thanks to an agreement between the SAGAS Department of the University of Florence and the Biblioteca Marucelliana, which holds the eighteenth-century manuscript composed by 111 volumes, a series of researches by students of Bibliography and the History of the book has brought to light the authors and printers cited by Marucelli. These data, checked against Italian and international authority files, formed the basis of a massive data insertion in Wikidata, linking the Mare Magnum to its bibliographical content. The work is a part of the projects started by Gruppo Wikidata per Musei, Archivi e Biblioteche to improve Wikidata with valuable and authoritative data.

Keywords: Mare Magnum, bibliography, history of printing, Francesco Marucelli.

Nel 2022 il Dipartimento SAGAS dell'Università di Firenze ha siglato con la Biblioteca Marucelliana un Protocollo di intesa al fine di valorizzare il Repertorio bibliografico *Mare Magnum* utilizzando lo strumento *open access* di Wikidata. Nel 2003, in occasione del terzo centenario della morte del Marucelli[1], l'opera era stata completamente digitalizzata e resa disponibile per la fruizione da remoto direttamente dal sito della Biblioteca. Di fatto però, fino agli anni Duemila, il *Mare Magnum* non era stato interessato da campagne specifiche di studio e di ricerca se si eccettuano un paio di contributi: uno di carattere più generale e l'altro relativo alla storia postale (Laurentiis 2008; Pescini 1993).

Il Mare Magnum è una lista manoscritta di citazioni bibliografiche organizzata in categorie e quarantatre classi la cui stesura è iniziata nel 1670 e che non si struttura intorno agli interessi di un bibliofilo, né a quelli di un collezionista, né tantomeno a quelli di uno studioso di una precisa disciplina. Francesco Marucelli prima, poi il nipote Alessandro e infine Angelo Maria Bandini, hanno dato vita ad un insieme che, pur nella sua grandezza, non ha sollecitato nel tempo interventi strutturati in grado di valorizzare le peculiarità di questo tenta-

[1] *Mare Magnum*, Biblioteca Marucelliana, Firenze, <https://www.internetculturale.it/it/41/collezioni-digitali/26172/mare-magnum>.

Valentina Sonzini, University of Florence, Italy, valentina.sonzini@unifi.it, 0000-0002-7142-9045

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

tivo settecentesco di sistematizzazione del sapere, forse anche a causa appunto della mole poderosa.

Fra l'altro, in esso, ad una prima analisi, non si rinvengono edizioni rare o di pregio (pochissimi erano gli incunaboli posseduti dal Marucelli), ma bensì edizioni coeve al compilatore, a riprova del fatto che il Repertorio non doveva presentarsi come il risultato dell'accumulazione di un dotto appassionato di libri, ma come strumento di organizzazione del sapere, utile per districarsi nelle complessità di una raccolta quantitativamente rilevante che l'abate aveva costituito nel corso della propria vita.

Per presentare brevemente la storia del Repertorio, è necessario ricordare che sia Francesco Marucelli, sia Alessandro nel suo testamento del 1751, non fornirono precise indicazioni sul futuro del manoscritto[2], e solo dopo il sostanziale intervento del bibliotecario Bandini, un Indice generale viene redatto nel 1888 da Guido Biagi (Biagi 1888). Da allora, i volumi del Mare Magnum riposano presso la Biblioteca Marucelliana a disposizione di ricercatori ed utenti che, come detto precedentemente, si sono però accostati al Repertorio solo occasionalmente.

La sfida è stata quindi quella di dare inizio ad una corposa opera di emersione dell'insieme che partisse dalla registrazione dei singoli dati bibliografici contenuti nei volumi manoscritti mai dati alle stampe. In primis, si è resa necessaria l'interpretazione degli item, sia nell'espressione grafica, sia nel contenuto (dato che alcune delle registrazioni appaiono come appunti da riformulare in citazioni più formali, come fanno intuire, in alcuni casi, l'assenza di uno o più dati bibliografici).

Volendo procedere con l'inserimento in Wikidata di autori e tipografi (si è infatti preferito, almeno per ora, omettere le citazioni bibliografiche complete prediligendo la valorizzazione degli authority file dei nomi), è stato innanzitutto necessario trascrivere ogni item, cioè ogni citazione raccolta nei singoli volumi attenzionati, inserendolo in una tabella excel organizzata a partire dai dati bibliografici che si volevano porre in evidenza, quali: autore, titolo, data e luogo di stampa, nome del tipografo/editore. Va sottolineato, come precedentemente rilevato, che non tutte le citazioni elencate riportano le informazioni bibliografiche utili all'identificazione dell'item, poiché più spesso si trovano veloci annotazioni. Inoltre, benché dall'archivio della biblioteca parrebbe che il lavoro di trascrizione del repertorio sia stato affidato al solo Antonio Ristori, il progetto ha evidenziato l'intervento di più persone che hanno copiato non tanto singole voci quanto piuttosto singoli fascicoli: quando infatti una voce occupa più fascicoli spesso si nota vistosamente il cambio di mano. Inoltre, i vari copisti, hanno effettuato qua e là aggiunte integrative postume.

A partire quindi da quanto è stato possibile rilevare in ogni item, si è proceduto con l'individuazione di ogni singola edizione all'interno di diversi OPAC (Online Public Access Catalogue), e, principalmente, in quelli nazionali curati dall'ICCU (Istituto Centrale per il Catalogo Unico): SBN, ed EDIT16 per le

[2] Solo il frontespizio venne stampato nel 1701 dal tipografo marchigiano Gaetano Zenobi.

Cinquecentine italiane. Tuttavia, in alcuni casi, soprattutto per quanto riguarda le edizioni stampate Oltralpe, i cataloghi citati non hanno restituito alcuna descrizione e ci si è quindi avvalsi di altri cataloghi, quali quelli della: BNF-Bibliothèque Nationale de France; BNE-Biblioteca Nacional de España; o del meta-OPAC internazionale KVK-Karlsruher Virtueller Katalog. ISTC-Incunabula Short Title Catalogue è stato invece utilizzato nei rari casi in cui è stato rilevato un incunabolo.

Reperita l'edizione, si sono potute estrapolare le voci di autorità a essa afferenti; non solo, quindi, l'autore principale, quasi sempre espresso nella citazione manoscritta, ma eventuali co-autori, incisori, contributori a vario titolo, tipografo e/o editore. Per rendere efficace e univoca l'identificazione, ad ognuna di queste voci è stato attribuito l'identificato fornito da SBN. Pertanto, a record processato, la tabella excel riportava non solo la trascrizione, ma anche gli identificativi di autore, tipografo e edizione (VID e BID).

L'implementazione dello schema di sintesi ha comportato alcune scelte frutto delle problematiche che, progressivamente, emergevano sia durante la trascrizione, sia durante l'identificazione. Per esempio, in alcuni casi, in fase di individuazione negli OPAC di una data edizione, quando luogo e data di stampa corrispondevano si è scelto di operare con elasticità nel trattare la non corrispondenza tra formati simili (quali l'ottavo e il dodicesimo) e rilevando la discrepanza nelle note, ma utilizzando comunque il codice identificativo dell'edizione censita. Non è infatti possibile determinare quanti di questi riferimenti bibliografici siano di prima mano, tratti cioè dalla collezione privata del Marucelli, o il risultato dell'osservazione diretta di esemplari conservati in altre biblioteche, ovvero quanto la conoscenza di alcune edizioni derivasse da una bibliografia consultata (magari in parte errata) o da indicazioni di seconda mano non verificabili direttamente a partire da una copia.

Va inoltre rilevato che il corpus è costituito, sebbene in una parte assolutamente minima, ma non irrilevante, da citazioni di manoscritti, evidentemente non identificabili con certezza, e da riferimenti bibliografici di edizioni non individuate nei vari OPAC consultati. Quest'ultimo è un dato interessante che restituisce una limitata porzione di quelle che possono, in tutta evidenza, considerarsi edizioni disperse, senza cioè una localizzazione attestata, citate in antichi repertori, ma non rintracciabili. Nella tabella excel di lavoro si è resa necessaria l'implementazione di un campo note per dar conto di come si è proceduto nei casi che hanno appunto presentato criticità.

Realizzato questo schema articolato di dati composto quindi dai rilevamenti bibliografici a cui è stato associato un identificativo specifico, si è proceduto con il riversamento in Wikidata[3].

[3] L'iniziativa che sta riguardando il MM è parte di un progetto più ampio inserito nell'attività promossa dal Gruppo Wikidata per Musei, Archivi e Biblioteche, <https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche>.

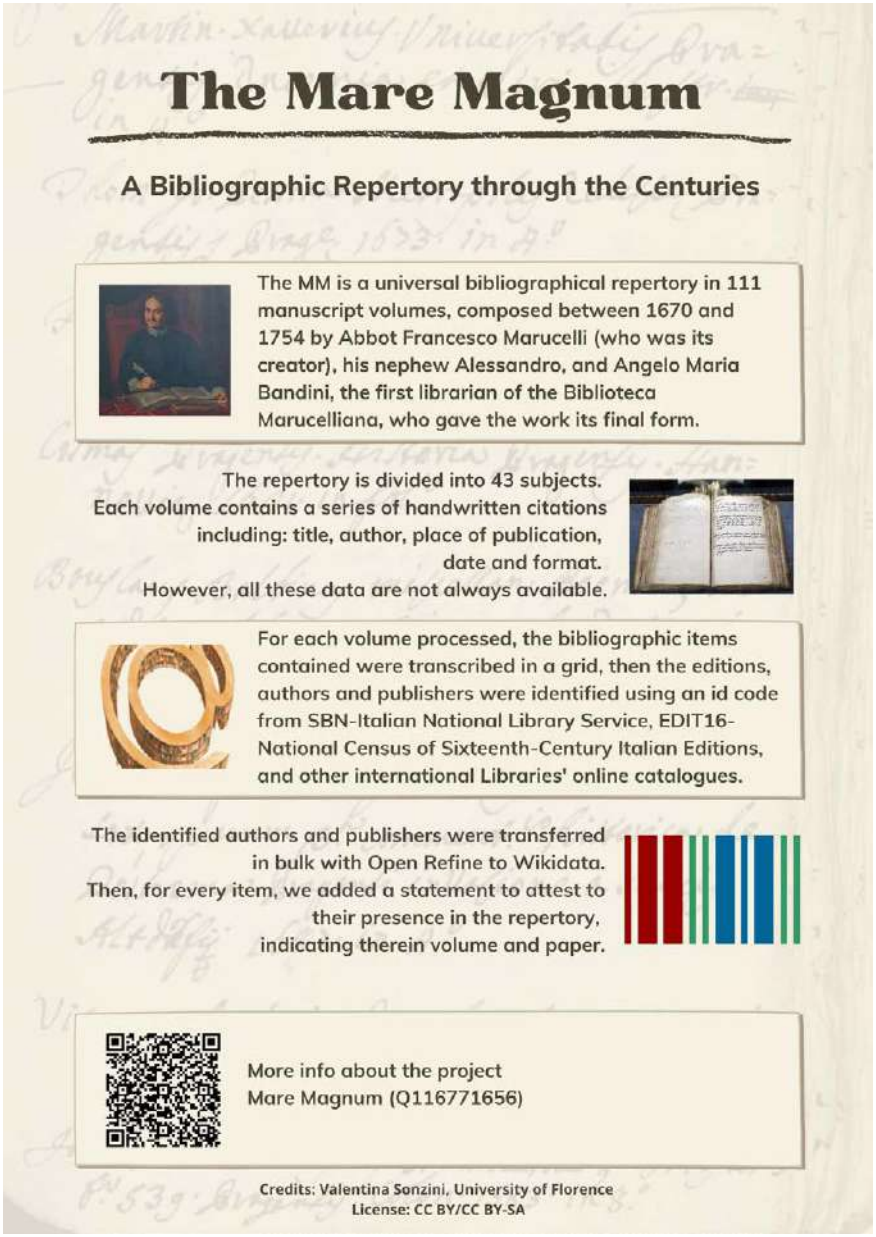


Figura 1 – Poster presentato al convegno "Wikidata e la ricerca", Firenze, 5 giugno 2025.

In primis, nel sistema è stata creata una scheda relativa al Repertorio del Marucelli oggetto dello studio[4]. Quindi, attraverso una procedura massiva, sono stati inseriti nel database tutti i codici identificativi SBN delle varie voci di autorità. A questo punto si sono verificate due eventualità: la voce di autorità era già presente (il caso, per esempio, di Aristotele, Q868); oppure non era ancora presente e andava creata. In entrambe le situazioni sono state associate le dichiarazioni e le proprietà fondamentali, nonché il codice identificativo SBN (se mancante). Infine, è stato creato lo *statement* «attestato in» (attested in) con proprietà «Mare Magnum vol. …» specificando la carta (o le carte) in cui la voce di autorità è contenuta. All'interno della dichiarazione Wikidata è stato inserito come riferimento l'URL che rinvia alla digitalizzazione dell'opera[5].

Il lavoro di valorizzazione digitale si è posto l'obiettivo di implementare le voci di autorità in Wikidata e, soprattutto, di creare collegamenti fra queste e il Repertorio del Marucelli rilasciando nel web dati ad accesso libero che potranno essere adeguatamente interrogati dagli utenti. I risultati ottenuti, visualizzati attraverso grafici prodotti da software di data visualization quali, per esempio, Tableau, potranno risultare utili per fornire una panoramica complessiva sui singoli volumi del Repertorio per quanto concerne gli autori e editori rilevati, il loro periodo di attività, le edizioni e i luoghi di pubblicazione.

Attraverso questi strumenti digitali in open access, il Mare Magnum è quindi in grado di vivere una nuova stagione, entrando nell'infosfera del web e rendendo il suo straordinario patrimonio bibliografico finalmente fruibile da parte della comunità scientifica internazionale.

Riferimenti bibliografici

Biagi, Guido. 1888. *Indice del Mare Magnum di Francesco Marucelli*. Roma: Ministero della Pubblica Istruzione.

De Laurentiis, Rossano. 2008. "Mare Magnum di Francesco Marucelli: un catalogo e la sua ricezione". In *Navigare nei mari dell'umano sapere, Biblioteche e circolazione libraria nel Trentino e nell'Italia del XVIII secolo*, a cura di Giancarlo Petrella. Trento: Provincia autonoma di Trento, Soprintendenza per i beni librari e archivistici 101-26.

Pescini, Ilaria. 1993. *Dal Mare Magnum dell'abate Marucelli. La più antica bibliografia di storia postale*. Prato: Istituto di studi storici postali.

[4] Mare Magnum (Q116771656), <https://www.wikidata.org/wiki/Q116771656>.

[5] Attualmente, la collezione digitale Mare Magnum della Biblioteca Marucelliana (<https://marucelliana.cultura.gov.it/scopri-la-biblioteca/collezioni-digitali>) non è ancora fruibile da remoto a causa di un attacco hacker che l'ha interessata nel 2024. Il Repertorio è consultabile solo dall'interno della biblioteca attraverso il link <http://cat.maru.firenze.sbn.it>. Quindi, persistendo tale situazione, gli URL "agganciati" allo *statement* non sono momentaneamente attivi.

The Journal of Open Humanities Data: Bridging open data and Wikidata for the Humanities

Andrea Farina, Barbara McGillivray

Abstract: This paper explores the intersection between open humanities research and Wikidata, focusing on data papers published in the *Journal of Open Humanities Data* (JOHD). We analyse 13 JOHD articles to assess how Wikidata is integrated, cited, or proposed for future use, identifying trends in disciplines, data formats, and periods covered. Our findings reveal a growing engagement with Wikidata, particularly in History and Literary Studies, and highlight its potential to enhance semantic richness, interoperability, and reuse. We argue for broader adoption through improved tools and community support. The conference also marks the launch of JOHD's new special collection, *Wikidata across the humanities: datasets, methodologies, reuse,* aimed at fostering further research and collaboration.

Keywords: Journal of Open Humanities Data (JOHD), datasets, data papers, open research, interoperability.

The intersection of open data infrastructures and humanities research practices has grown increasingly significant in recent years. At the heart of this development lies a dual trend the expansion of community-driven platforms like Wikidata and the formalisation of academic publishing through venues such as the *Journal of Open Humanities Data* (JOHD[1]). Wikidata serves as a critical tool for enriching and interconnecting datasets, enabling researchers to explore relationships across diverse domains (Farda-Sarbas and Müller-Birn 2019; Neubert 2017). By offering a centralised platform for integrating identifiers, metadata, and semantic links, Wikidata facilitates the creation of interoperable and reusable datasets that support advanced analysis and interdisciplinary research. Established in 2015 by Ubiquity Press, JOHD is the largest interdisciplinary data journal dedicated specifically to humanities disciplines. JOHD complements existing infrastructures and serves as a peer-reviewed outlet for data papers, i.e., scholarly articles dedicated to the description of research datasets in the humanities, with a focus on their creation, structure, and reuse potential. These

[1] <https://openhumanitiesdata.metajnl.com/> (last accessed date: 12 June 2026).

Andrea Farina, King's College London, United Kingdom, andrea.farina@kcl.ac.uk, 0000-0002-1948-9008
Barbara McGillivray, King's College London, United Kingdom, barbara.mcgillivray@kcl.ac.uk, 0000-0003-3426-8200

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Andrea Farina, Barbara McGillivray, *The Journal of Open Humanities Data: Bridging open data and Wikidata for the Humanities*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.41, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 337-341, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

papers not only document data collection and management practices, but also contribute to scholarly credit systems for data work, reinforcing the principles of openness, interoperability, and reproducibility (McGillivray et al. 2022; Farina et al. 2025). JOHD embraces the potential of platforms like Wikidata to amplify the impact of open data, and shares a mission rooted in transparency, accessibility, and collaboration. Together, initiatives like Wikidata and JOHD are elevating the role of data in advancing both scholarly inquiry and public engagement across the humanities (Wigdorowitz et al. 2024).

In this study, we examine the extent to which JOHD data papers engage with Wikidata, either through direct integration or conceptual alignment. The investigation is motivated by two key questions: (i) to what degree is Wikidata being used as a technical and semantic resource within published humanities datasets, and (ii) what potential remains untapped for its broader adoption? Our analysis is based on a comprehensive review of all JOHD publications available as of April 2025, identifying a total of 13 data papers that either reference, integrate, or discuss Wikidata in relation to their datasets. Overall, the use and visibility of Wikidata in JOHD papers has increased over time, with a concentration of contributions focusing on Early Modern (from the end of the Medieval Period around 1500 CE to the late 18th century) and Modern periods (from the late 18th century to the end of World War II in 1945). Most of these datasets are large-scale (i.e. they contain between 100,000 and 10 million records) and typically formatted in CSV or XML, reflecting current technical standards in humanities data publication. History and Literary Studies emerge as the fields most actively experimenting with Wikidata integration, which suggests a disciplinary alignment that may not yet be equally developed across the broader humanities spectrum.

A close reading of these papers reveals three distinct patterns of engagement with Wikidata. Several data papers demonstrate a direct and functional use of Wikidata, where identifiers and lexicographical resources from the platform are actively embedded within the datasets. These examples (e.g., Besnier and Mattingly 2021; Coll Ardanuy et al. 2022; Giovannini and Gallucci 2024; Röttgermann 2024; Malínek et al. 2024) illustrate how Wikidata facilitates entity disambiguation, enriches metadata structures, and enhances semantic interoperability. For instance, in their dataset on Early Modern Italian Drama, Giovannini and Gallucci (2024, 3) report aligning multiple categories (author, city, location, publisher, composer) with corresponding Wikidata items to improve data coherence and machine readability. Similarly, Röttgermann (2024) uses Wikidata to enrich metadata in her collection of 18th century French novels, drawing on the platform to assign gender information to authors alongside automated methods based on gender-specific titles. These cases demonstrate the practical role Wikidata can play in standardising and enhancing humanities datasets through semantic linking.

A second group of papers reference Wikidata as a conceptual or infrastructural resource without directly integrating it into their dataset, though they acknowledge its relevance for future reuse and interoperability. These instances signal an emerging alignment between humanities datasets and the broader

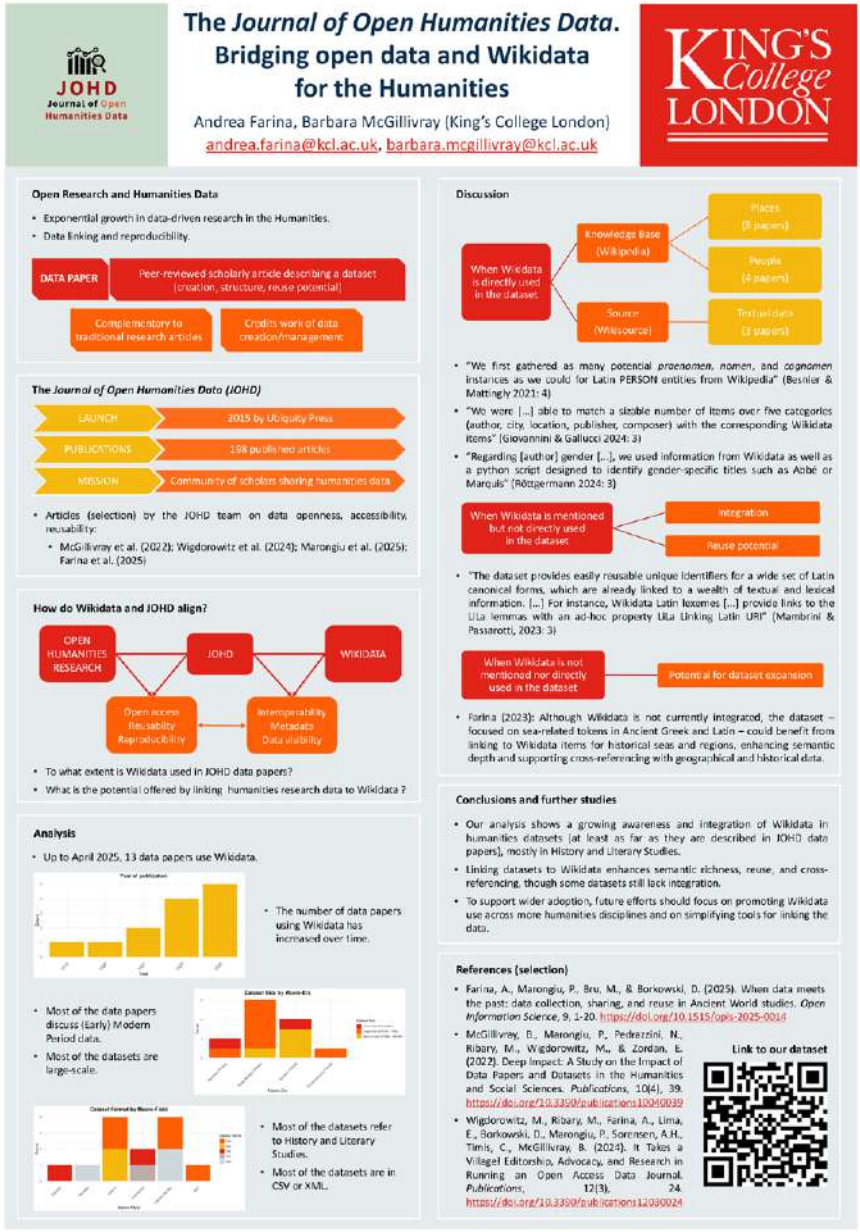


Figure 1 – Poster presented at "Wikidata and Research". Conference Proceedings, Florence, 5-6 giugno 2025.

Linked Open Data (LOD) ecosystem. For example, Mambrini and Passarotti (2023, 3) emphasise that their Latin Lemma Bank provides stable, reusable URIs for Latin canonical forms based on the OntoLex model and note that "Wikidata Latin lexemes [...] provide links to the LiLa lemmas with an ad-hoc property LiLa Linking Latin URI". While their dataset itself does not link out to Wikidata, they highlight that Wikidata has already begun linking it, and that federated queries across SPARQL endpoints (including Wikidata's) can enhance future research applications. This demonstrates a growing infrastructural integration, even if it is not yet fully realised at the dataset level.

The third category includes papers that do not currently engage with Wikidata, but identify its potential utility, although this is not overtly stated in the reuse potential section of the paper. The authors of these papers recognise opportunities to enhance semantic depth, link to historical and geographical identifiers, and improve cross-referencing with established authority data. Farina (2023), for instance, discusses a dataset on maritime terminology in Ancient Greek and Latin and proposes that expanding it to Wikidata entries for historical seas and regions could significantly improve its contextual richness and interoperability.

The findings of this analysis point to a growing convergence between the goals of open humanities research and the affordances of Wikidata. When integrated effectively, Wikidata enhances data transparency, semantic clarity, and reuse potential. Our review of JOHD publications confirms this emerging alignment: while only 13 data papers to date explicitly reference or integrate Wikidata, the diversity and depth of these engagements suggest that LOD practices are gaining traction across the humanities. Nonetheless, broader adoption remains uneven. Challenges include the technical barrier posed by data modelling and linking procedures, the limited availability of domain-specific tools, and a lack of widespread training or incentives to adopt Linked Open Data workflows. To address these gaps and foster broader uptake, future initiatives should prioritise community-driven tool development, training programs, and interdisciplinary dialogue between data creators and Wikidata contributors. In this spirit, and building on the momentum observed in our study, the JOHD special collection *Wikidata across the humanities: datasets, methodologies, reuse*[2] provides a dedicated venue for exploring how Wikidata is being applied across humanities disciplines, with a focus on practical integration, innovative modelling approaches, and examples of semantic enrichment and reuse. By foregrounding this work, the collection supports the continued expansion of Wikidata-enabled research and reinforces the platform's role as a foundational infrastructure for open knowledge in the humanities. This conference has provided a valuable opportunity to engage with the active Wikidata research community, and we hope this initiative will further promote collaboration, visibility, and innovation at the intersection of data publication and semantic web technologies.

[2] <https://openhumanitiesdata.metajnl.com/collections/wikidata_across_the_humanities> (last accessed date: 12 June 2026).

References

Besnier, Clément, and William Mattingly. 2021. "Named-Entity Dataset for Medieval Latin, Middle High German and Old Norse." *Journal of Open Humanities Data* 7: 23. https://doi.org/10.5334/johd.36.

Coll Ardanuy, Mariona, David Beavan, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, Katherine McDonough, Federico Nanni, Daniel van Strien, and Daniel C. S. Wilson. 2022. "A Dataset for Toponym Resolution in Nineteenth-Century English Newspapers." *Journal of Open Humanities Data* 8 (1): 3. https://doi.org/10.5334/johd.56.

Farda-Sarbas, Miriam, and Claudia Müller-Birn. 2019. "Wikidata from a Research Perspective: A Systematic Mapping Study of Wikidata." arXiv, abs/1908.11153. https://doi.org/10.48550/arXiv.1908.11153.

Farina, Andrea. 2023. "Lost at Sea: A Dataset of 25+ SEA Words Morpho-Semantically Annotated in Ancient Greek and Latin." *Journal of Open Humanities Data* 9: 24, 1–7. https://doi.org/10.5334/johd.139.

Farina, Andrea, Paola Marongiu, Mathilde Bru, and Daniele Borkowski. 2025. "When Data Meets the Past: Data Collection, Sharing, and Reuse in Ancient World Studies." *Open Information Science* 9 (1): 20250014. https://doi.org/10.1515/opis-2025-0014.

Giovannini, Luca, and Giorgia Gallucci. 2024. "Allacci Digitale: A Historical Dataset for Early Modern Italian Drama." *Journal of Open Humanities Data* 10 (1): 55. https://doi.org/10.5334/johd.250.

Malínek, Vojtěch, Tomasz Umerle, Edward Gray, Ivan Heibi, Péter Király, Christiane Klaes, Przemysław Korytkowski, David Lindemann, Arianna Moretti, Charlotte Panušková, Róbert Péter, Mikko Tolonen, Aldona Tomczyńska, and Ondřej Vimr. 2024. "Open Bibliographical Data Workflows and the Multilinguality Challenge." *Journal of Open Humanities Data* 10: 27, 1–14. https://doi.org/10.5334/johd.190.

Mambrini, Francesco, and Marco Carlo Passarotti. 2023. "The Lila Lemma Bank: A Knowledge Base of Latin Canonical Forms." *Journal of Open Humanities Data* 9 (1): 28. https://doi.org/10.5334/johd.145.

McGillivray, Barbara, Paola Marongiu, Nilo Pedrazzini, Marton Ribary, Mandy Wigdorowitz, and Eleonora Zordan. 2022. "Deep Impact: A Study on the Impact of Data Papers and Datasets in the Humanities and Social Sciences." *Publications* 10 (4): 39. https://doi.org/10.3390/publications10040039.

Neubert, Joachim. 2017. "Wikidata as a Linking Hub for Knowledge Organization Systems? Integrating an Authority Mapping into Wikidata and Learning Lessons for KOS Mappings." *NKOS@TPDL*: 1–12.

Röttgermann, Julia. 2024. "The Collection of Eighteenth-Century French Novels 1751–1800." *Journal of Open Humanities Data* 10 (1): 31. https://doi.org/10.5334/johd.201.

Wigdorowitz, Mandy, Marton Ribary, Andrea Farina, Eleonora Lima, Daniele Borkowski, Paola Marongiu, Amanda H. Sorensen, Christelle Timis, and Barbara McGillivray. 2024. "It Takes a Village! Editorship, Advocacy, and Research in Running an Open Access Data Journal." *Publications* 12 (3): 24. https://doi.org/10.3390/publications12030024.

Using Wikidata in the European Literary Bibliography: A Reproducible Approach

Gustavo Candela, Cezary Rosiński, Arkadiusz Margraf

Abstract: The European Literary Bibliography intends to open bibliographic data for literary studies at the European level holding resources from several institutions. This work provides a reproducible framework for publishing and reusing digital collections based on literary bibliographies made available by GLAM institutions. It also presents a collection of DH research scenarios to show how data can be explored. The results are available in the form of reproducible code. This work advances the publishing of digital collections in computationally usable forms describing how Wikidata can be used to explore new ways of analysis of the data. Future research directions include extending and implementing the research scenarios, and applying and adapting the framework to other domains.

Keywords: Data Publishing Framework, Collections as Data, Linked Open Data, European Literary Bibliography (ELB), GLAM.

GLAM institutions (Galleries, Libraries, Archives, and Museums) have been exploring new ways to make available their digital collections. Wikidata has emerged as a leading approach with which to enrich their digital collections (Candela 2024). In parallel, new trends such as Labs and Collections as data promote the publication of digital collections suitable for computational use (Candela et al. 2023) as well as the use of reproducible code in the form of Jupyter Notebooks (Candela, Chambers and Sherrat 2023).

The European Literary Bibliography (ELB) is a project of the Institute of Czech Literature (Czech Academy of Sciences) and the Institute for Literary Research (Polish Academy of Sciences). It intends to open bibliographic data for literary studies at the European level holding resources from several institutions.

Following the Collections as data principles and focusing on the ELB, this work provides a reproducible framework including several steps for publishing and reusing digital collections based on literary bibliographies made available by GLAM institutions (Candela, Rosiński and Margraf 2024). It also presents a collection of Digital Humanities (DH) research scenarios to show how data can be explored and reused. This work is the result of an ATRIUM Transna-

Gustavo Candela, University of Alicante, Spain, gcandela@ua.es, 0000-0001-6122-0777
Cezary Rosiński, University of Extremadura, Spain, cezary.rosinski@ibl.waw.pl, 0000-0002-6136-7186
Arkadiusz Margraf, University of València, Spain, margraf@man.poznan.pl, 0009-0002-8608-5991

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Gustavo Candela, Cezary Rosiński, Arkadiusz Margraf, *Using Wikidata in the European Literary Bibliography: A Reproducible Approach,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.42, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 343-346, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

tional Access Scheme Grant. The results are available in the form of a repository of reproducible code[1].

1. A reproducible framework to transform bibliographic metadata to Collections as data

This section presents the framework to publish and reuse digital collections in the form of Collections as data (Candela, Rosiński and Margraf 2024), as shown in Figure 1.

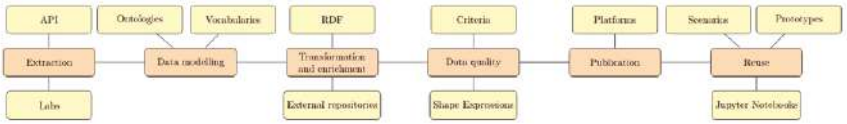


Figure 1 – Reproducible framework for publishing and reusing collections as data.

Data extraction refers to the selection of data relevant to a specific topic (e.g., author, organization or theme). *Data modelling* aims at ensuring machine-readable bibliographic metadata, using ontologies and vocabularies. The *transformation and enrichment* step refers to the transformation of the data into Linked Open Data (LOD) using RDF to describe metadata as triples as well as the use of Wikidata to enrich the metadata. The *data quality* step ensures the high quality of the RDF data. The publication requires the inclusion of additional documentation including aspects such as provenance and licensing. Finally, the published datasets can be *reused* in various ways (e.g., prototypes or research scenarios defined by DH scholars).

2. Defining research scenarios

After applying the proposed framework to the ELB and exploring new uses of the data, a selection of research scenarios were defined to illustrate data reuse and integration using Wikidata as a main repository with which to enrich the metadata: i) comparative analysis of provincial vampire novels in Spain; ii) republican writers who emigrated during the Spanish Civil War; and iii) geographical distribution of publications about specific Spanish writers. Limitations were identified in terms of scope and completeness to meet researchers' needs.

3. Conclusions

This work advances the publishing of digital collections in computationally usable forms describing how Wikidata can be used to explore new ways of analysis of the data. Future research directions include extending and implementing the research scenarios, and applying and adapting the framework to other domains.

[1] <https://gitlab.pcss.pl/dl-team/atrium/tna01/> (accessed May 13, 2026).

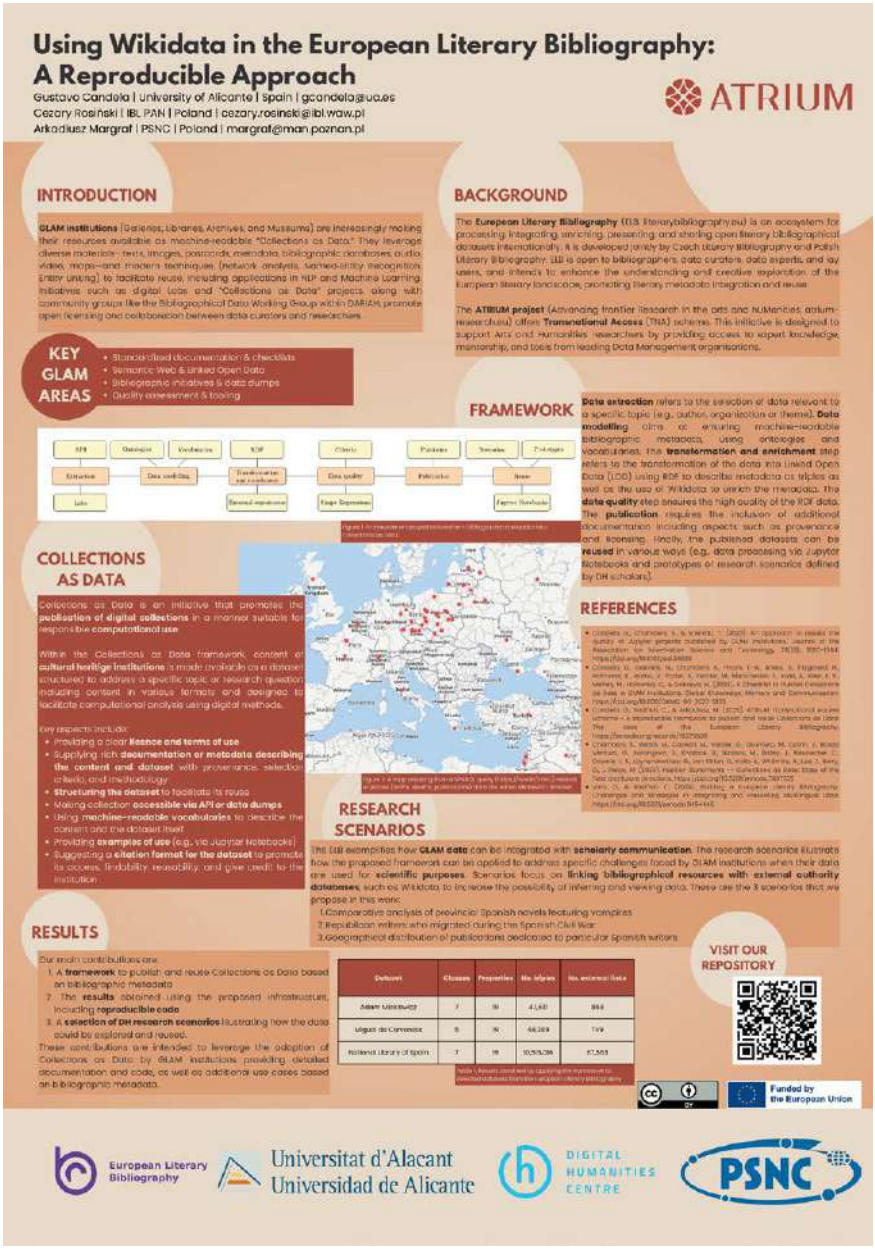


Figure 2 – Poster submitted at "Wikidata and Research". Conference Proceedings, Florence, 5-6 giugno 2025.

References

Candela, Gustavo, Sally Chambers, and Tim Sherratt. 2023. "An Approach to Assess the Quality of Jupyter Projects Published by GLAM Institutions." *Journal of the Association for Information Science and Technology* 74 (13): 1550–64. https://doi.org/10.1002/asi.24835.

Candela, Gustavo, Mirjam Cuper, Olga Holownia, Nele Gabriëls, Milena Dobreva, and Mahendra Mahey. 2024. "A Systematic Review of Wikidata in GLAM Institutions: A Labs Approach." *Lecture Notes in Computer Science*: 34–50. https://doi.org/10.1007/978-3-031-72440-4_4.

Candela, Gustavo, Nele Gabriëls, Sally Chambers, et al. 2023. "A Checklist to Publish Collections as Data in GLAM Institutions." *Global Knowledge, Memory and Communication* 74 (5-6): 1323–55. https://doi.org/10.1108/gkmc-06-2023-0195.

Candela, Gustavo, Cezary Rosiński, and Arkadiusz Margraf. 2024. "A Reproducible Framework to Publish and Reuse Collections as Data: The Case of the European Literary Bibliography." Zenodo. https://doi.org/10.5281/zenodo.14106707.

From Data to Images: OpenRefine for Wikidata and Wikimedia Commons

Elena Martellotta

Abstract: Many cultural institutions are joining the open data movement by making their collections and archives publicly available, in line with Wikimedia's mission to provide free access to human knowledge. Through Wikidata and Wikimedia Commons, they share structured data on artworks, events, and historical figures, fostering research and collaboration. In this context, the pilot project Progetto Dati Lombardia was created, starting from a public-domain dataset on cultural heritage buildings in Lombardy. Using OpenRefine and Quickstatements, the data were processed and uploaded to Wikidata, minimizing information loss. At the same time, the Egyptian Museum of Turin released both data and images, linking them to Wikidata and Wikimedia Commons. These initiatives show how institutions can enhance and disseminate global cultural heritage.

Keywords: Opendata, OpenRefine, Wikidata, cultural heritage.

Sempre più istituzioni culturali stanno aderendo all'*open data movement,* estendendo l'accessibilità della fruizione, nonché dello studio delle proprie collezioni, sia in termini di dati che di immagini, a tutti i tipi di pubblico. Questo approccio risponde alla mission del movimento Wikimedia, che mira a fornire libero accesso alla conoscenza umana. Attraverso piattaforme come Wikidata e Wikimedia Commons, le istituzioni possono condividere dati strutturati relativi a opere d'arte e documenti, rendendoli così interoperabili. 'Liberando' dati e immagini per il loro uso e riuso digitale, viene agevolata e incoraggiata non solo la ricerca, ma anche la collaborazione finalizzata ad un accrescimento della conoscenza, coinvolgendo diversi ambiti, dalla ricerca accademica, alla didattica, fino alla divulgazione.

Da queste premesse, è nato il progetto pilota sui Dati relativi ai beni culturali della Regione Lombardia, sviluppato con l'obiettivo di approfondire pratiche, metodi e possibilità d'uso delle estensioni di OpenRefine. Dal dataset degli immobili dichiarati di interesse culturale insistenti sul territorio regionale, il quale è rilasciato in pubblico dominio, i dati sono stati inizialmente elaborati e preparati tramite OpenRefine, verificati con Quickstatements e poi caricati su Wikidata grazie all'estensione presente su OpenRefine. La trasformazione dei record in valori di proprietà per Wikidata ha permesso di minimizzare la per-

Elena Martellotta, elenamartellotta@gmail.com

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Elena Martellotta, *From Data to Images: OpenRefine for Wikidata and Wikimedia Commons,* © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.43, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 347-351, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

dita di informazioni. In tal modo, il Progetto Dati Lombardia non solo ha reso accessibili al pubblico una serie di informazioni preziose sul patrimonio architettonico e culturale della regione, ma ha anche definito un modello metodologico replicabile per altre istituzioni e territori.

Successivamente, parallelamente all'implementazione specifica di OpenRefine per Wikimedia Commons, il Museo Egizio di Torino ha rilasciato non solo i dati, ma anche le immagini della propria collezione, già disponibili sul suo sito web, dimostrando di essere tra le istituzioni museali più virtuose e disponibili all'apertura dei dati e delle immagini.

Il Progetto Dati Lombardia e l'iniziativa del Museo Egizio di Torino dimostrano che la convergenza tra open data, tecnologie semantiche e collaborazione comunitaria non è solo possibile, ma estremamente feconda.

1. Upload dati

La prima parte del progetto si è occupata dei dati, ed è quindi stata analoga a quanto già svolto per il Progetto Dati Lombardia. Per le immagini, poi, è stato necessario seguire un iter diverso per collegare la pagina dell'elemento su Wikidata a ciascuna immagine, in modo che i metadati fossero già strutturati prima del caricamento su Wikimedia Commons.

Il data wrangling preliminare è stato finalizzato alla successiva importazione dei record come valori delle proprietà di Wikidata, eliminando incongruenze, duplicati ed errori tipografici, senza perdere alcun tipo di informazione riportata nel dataset di partenza. Successivamente, i dati sono stati verificati tramite Quickstatements, uno strumento che permette di caricare grandi quantità di informazioni su Wikidata in modo automatico ma controllato.

La prima operazione è stata quella di individuare gli item già esistenti su Wikidata a cui erano stati aggiunti i numeri di inventario, dato univoco al fine di evitare la creazione di duplicati. A seguito delle difficoltà di matching, oltre che dall'analisi dell'identificativo stesso, si è intervenuti sull'espressione regolare del formato dell'identificativo, così da includere tutte le tipologie di sigle alfanumeriche di inventario.

Il lavoro successivo ha puntato alla normalizzazione dei record per la tipologia di valori delle diverse proprietà. L'istanza di (P13), ad esempio, è stata popolata dal tipo di categoria assegnato al reperto archeologico. La colonna delle descrizioni è confluita nelle etichette, sia in inglese che in italiano. Nel caso del materiale e della tecnica, per meglio rappresentare i manufatti polimaterici, sono state generate più colonne con separatore per distinguere i diversi materiali costituenti dei manufatti. Per l'inserimento delle dimensioni, grazie all'uso di espressioni, dai record sono stati ottenuti valori di tipo numerico. Sfruttando i separatori presenti, i valori sono stati suddivisi in tre nuove colonne, rispettivamente per altezza, spessore e larghezza, per poi essere riconciliati sulla proprietà di Wikidata per grandezze fisiche. In questo caso, laddove presente, è stato preservato anche il peso. Proseguendo, gli estremi cronologici presentano date sia avanti che dopo Cristo, motivo per cui i valori numerici delle celle sono stati

riconciliati su 'anno a.C.' e 'anno'. Per questioni di praticità, oltre che di maggiore affinità per l'appartenenza ad una collezione, si preferisce questa soluzione alla data di fondazione o creazione (P571) in cui si inserisce il millennio di riferimento, sebbene anche questa presenti l'opportunità di inserire i limiti inferiori (P1319) e superiori (P1326). Inoltre, per maggiori riferimenti cronologici si importano le colonne con il periodo di riferimento, della dinastia e del regno. Si è reso necessario, inoltre, trasformare il contenuto delle celle relative alla collocazione *Luogo (P276),* creando un'ulteriore colonna in cui inserire la sotto stringa che specifica il numero della sala, facendo attenzione inoltre ad eliminare lo zero che precede i numeri delle prime nove sale. Segue dunque la riconciliazione complessa sul tipo 'stanza' che abbia il Luogo (ME) corrispondente alla proprietà *Parte di (P361),* essendo state precedentemente create le sale corrispondenti. Per il link del Museo Egizio, normalizzato anche questo, si è mantenuta sia la versione in lingua italiana che in lingua inglese dell'url in quanto valore della proprietà *Descritto nell'URL (P973).*

Si è ritenuto inoltre necessario creare nuove colonne per sottolineare l'appartenenza degli oggetti museali alla collezione del Museo Egizio (*Collezione - P195),* oltre allo *Status del Copyright (P6216).* Quest'ultimo specifica che i dati inseriti sono rilasciati in pubblico dominio, compatibilmente con la licenza di WikiData.

2. Upload immagini

Questa seconda fase del progetto ha riguardato interamente l'associazione dei dati strutturati alle immagini dei reperti archeologi del Museo Egizio di Torino. Per ciascun reperto si è creato un nuovo progetto su OpenRefine popolato dai valori, poi riconciliati con Wikimedia commons, con la seguente struttura:

1) url,
2) filename,
3) depicts,
4) wikitext.

Per predisporre la creazione di nuovi item su Wikimedia Commons che corrispondono alle immagini, si effettua la riconciliazione sull'*url* del cloud in cui erano stoccate le stesse. Nelle celle relative a *depicts,* si inserisce il Codice univoco di Wikidata, successivamente riconciliato in Wikidata. Grazie a questa operazione, diventa quindi possibile popolare le celle del *filename* grazie alle espressioni di OpenRefine. Infine, si crea la colonna del *wikitext* con il tipo di oggetto, la fonte, la licenza e la categoria di Wikimedia commons Media form Museo Egizio (Turin). Per esportare le immagini, è necessario creare lo *schema* in cui si collocano i valori degli url nel *filepath,* oltre al *filename* e al *wikitext.* Si prosegue aggiungendo le dichiarazioni del *copyright status* e del c*opyright license,* oltre che della *digital representation, mail subject* e *depicts.* Queste ultime tre proprietà presentano lo stesso valore, ovvero l'item di Wikidata con cui è stato riconciliato il valore di *depicts.* Dopo il controllo dei problemi, si caricano finalmente le modifiche su Wikibase.

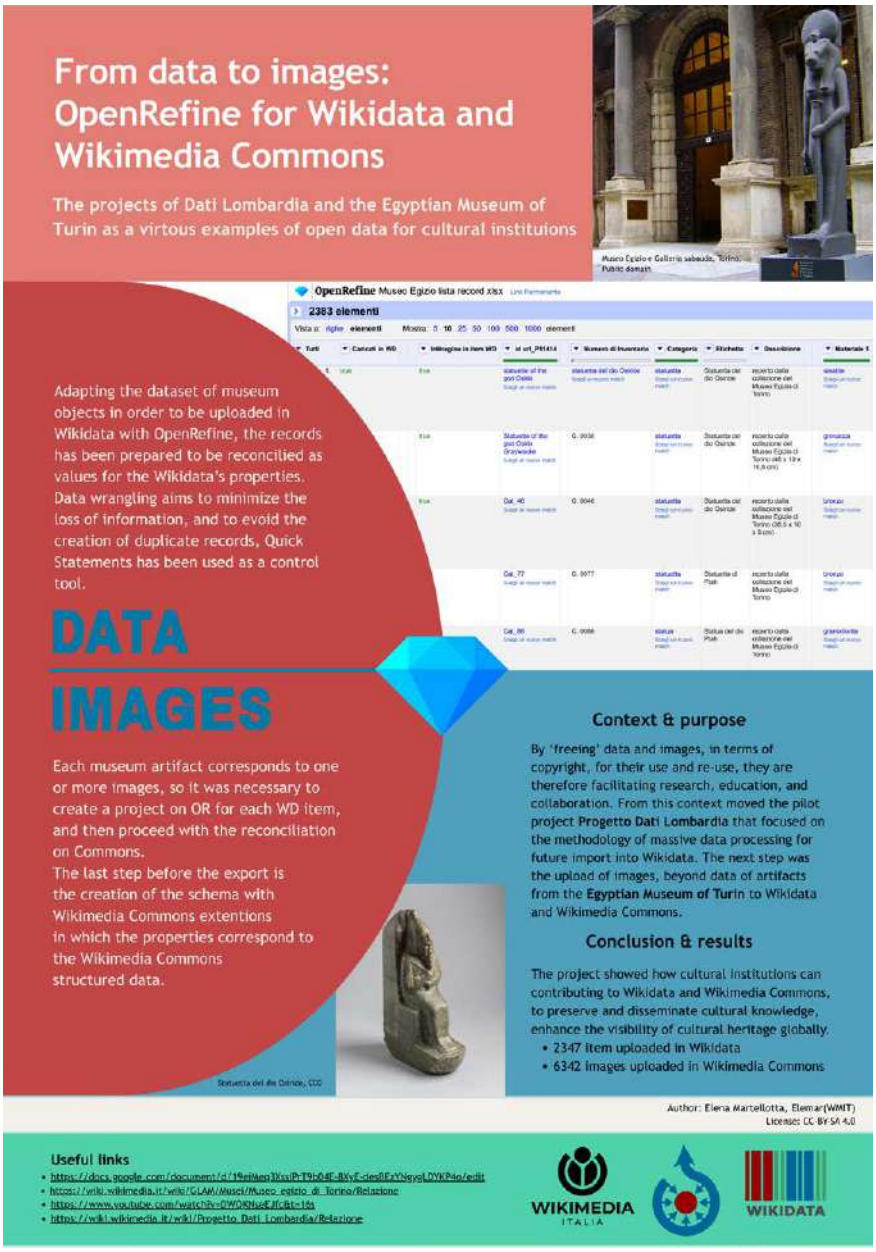


Figure 1 – Poster presented at "Wikidata and Research". Conference Proceedings, Florence, 5-6 giugno 2025.

Sono stati quindi caricati 2347 nuovi item su Wikidata e 6342 immagini su Wikimedia Commons. In entrambi i progetti è stato messo in evidenza il ruolo cruciale che il rilascio da parte delle istituzioni culturali di dati, e in particolar modo di immagini, possano contribuire a progetti come Wikidata e Wikimedia Commons. Attraverso software open source e l'impiego di dati strutturati che possano essere interoperabili, è possibile contribuire alla diffusione della conoscenza. Difatti, migliorando la visibilità e l'accessibilità del patrimonio culturale, non solo si costruisce una rete di scambio sostenibile tra le istituzioni, ma si partecipa alla trasformazione del modo di concepire la cultura nell'era digitale.

HQWiki: A Universal Pipeline for High-Quality Monolingual Dataset Creation

Iaroslav Chelombitko, Aleksey Komissarov

Abstract: Creating high-quality monolingual datasets poses significant challenges due to cross-language contamination, metadata noise, and inconsistent quality control, especially for less-resourced languages. We present HQWiki, a pipeline for extracting and validating clean monolingual content from Wikipedia. The framework introduces universal approaches combining character set validation, context-aware filtering, and adaptive language detection. Evaluation across multiple language families demonstrated significant noise reduction while maintaining high accuracy in language identification. The modular architecture allows researchers to adapt filtering criteria for specific requirements, providing a foundation for systematic dataset creation across different language families.

Keywords: dataset creation, monolingual corpora, Wikipedia processing, language identification, less-resourced languages.

1. Introduction

High-quality monolingual text corpora serve as the foundation for modern natural language processing systems, from language models to specialized applications like optical character recognition and morphological analysis. While Wikipedia represents one of the most comprehensive multilingual resources available, its widespread use masks significant quality challenges that compromise downstream NLP applications. These challenges include pervasive cross-lingual contamination, where articles contain embedded phrases or entire sentences in languages other than the document's primary language; disruptive metadata noise from XML tags, HTML comments, and template artifacts; and complex intra-sentence code-switching that disrupts linguistic consistency.

These quality issues disproportionately affect less-resourced languages (LRLs), where limited data availability makes every training example critical. Standard corpus preparation methods, primarily designed for high-resource languages like English, often prove inadequate for LRLs due to insufficient tooling support and different linguistic characteristics. Our recent work on specialized tokenization for Uralic languages (Chelombitko and Komissarov 2024) demonstrated how data quality directly impacts tokenizer performance and model efficiency. The development of monolingual BPE tokenizers from clean Wiki-

Iaroslav Chelombitko, Neapolis University Pafos, Cyprus, les4sixstm@gmail.com, 0009-0003-6843-0453
Aleksey Komissarov, aglabx, Cyprus, ad3002@gmail.com

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Iaroslav Chelombitko, Aleksey Komissarov, *HQWiki: A Universal Pipeline for High-Quality Monolingual Dataset Creation*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1010-2.44, in Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, pp. 353-358, 2026, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

pedia data showed substantial improvements over general-purpose tokenizers, reducing compression ratios significantly for underrepresented languages.

To address corpus quality challenges systematically, we present HQWiki, a flexible and extensible pipeline designed to extract and validate clean monolingual content from Wikipedia. By employing specialized filtering techniques grounded in script-level analysis and adaptive validation, HQWiki establishes a more reliable foundation for diverse NLP applications while significantly improving data quality for less-resourced languages.

2. Methodology

Our pipeline employs a multi-stage process designed to rigorously extract, clean, and filter Wikipedia content while addressing common data quality issues systematically. The approach prioritizes transparency and reproducibility through modular architecture that allows researchers to adapt filtering criteria for specific requirements.

The process begins with data acquisition from Wikipedia ZIM archives, which provide compressed, offline-accessible dumps of Wikipedia content. We prioritize recent "nopic" versions that exclude images, focusing computational resources on textual content. Initial extraction parses HTML from valid articles, targeting paragraphs within the main content area while performing basic cleaning operations including tag removal and whitespace normalization.

A critical innovation in our approach involves identifying the dominant script for each language's dataset. Rather than relying on language identification tools that often fail for LRLs or short text segments, we analyze character-level script distribution using Unicode properties across a representative sample. The system determines the single most frequent script—such as Latin, Cyrillic, or Devanagari—as the dominant script for subsequent filtering operations.

The core cleaning mechanism implements word-level script conformance filtering. For each word in the text, the pipeline normalizes by removing leading and trailing punctuation, then validates that every alphabetic character matches the dominant script. This approach effectively removes mixed-script words, foreign terminology, and template artifacts while preserving legitimate linguistic content. Words failing this validation are discarded, creating a monolingual corpus with high script consistency.

To eliminate low-value content, the pipeline filters articles based on valid word counts after script filtering. Articles containing fewer than ten words are discarded entirely, removing stubs, disambiguation pages, and other content unsuitable for training data. This threshold balances corpus size with content quality, ensuring that retained articles contain sufficient context for meaningful language modeling.

We evaluated several advanced language identification tools including CLD2, CLD3, FastText, and langid.py for final verification stages. However, these tools demonstrated significant limitations for our use case: poor support for many less-resourced languages, unreliable performance on short text segments, and inconsistent behavior on text containing technical terminology or proper nouns. Our evaluation

framework (Chelombitko, Safronov, and Komissarov 2024) provided systematic assessment of these tools' multilingual capabilities, confirming that script-based filtering offers more robust and predictable results across diverse language families.

The modular architecture facilitates adaptation and extension. Researchers can adjust filtering thresholds, incorporate additional cleaning stages, or integrate domain-specific validation rules. Both the framework code and resulting corpora are released as open-source resources, ensuring transparency and enabling reproducible research across the community.

3. Results

Application of the HQWiki pipeline to Wikipedia dumps demonstrated its effectiveness across diverse linguistic contexts. We processed data from 321 distinct languages, representing major language families and writing systems worldwide. Dominant script analysis confirmed broad coverage: Latin script predominated at 68,85% of processed languages, followed by Cyrillic (11,53%), Arabic (4,05%), Devanagari (3,12%), and various other scripts collectively representing approximately 10% of the dataset.

Quantitative analysis of major European languages illustrated the pipeline's impact on data quality and volume. For English, the system processed 54.1 million raw paragraphs, applying rigorous filtering to retain 40.9 million high-quality paragraphs—a 75% retention rate yielding 2.24 billion words. German showed similar effectiveness with 22.4 million raw paragraphs producing 16.7 million clean paragraphs (74% retention) totaling 820.9 million words. French demonstrated even higher retention at 78%, processing 20.5 million paragraphs to produce 16.0 million clean paragraphs and 739.3 million words.

Languages with different script characteristics showed comparable quality improvements. Spanish achieved 84% paragraph retention, the highest among evaluated languages, processing 14.1 million raw paragraphs into 11.9 million clean paragraphs. Ukrainian, representing Cyrillic script languages, maintained 61% retention despite the additional challenges of distinguishing legitimate borrowings from contamination in a language with substantial historical multilingualism.

The pipeline proved particularly effective for Uralic languages, where our previous tokenization work required clean monolingual data. For Finnish, the system processed 3.1 million raw paragraphs into 2.1 million clean paragraphs (68% retention), producing 74.1 million words of high-quality training data. This clean corpus enabled development of specialized tokenizers that reduced compression ratios from 3.41 (LLaMA-2) to 1.18, demonstrating the direct impact of corpus quality on downstream performance.

4. Applications and impact

The high-quality monolingual corpora generated by HQWiki provide immediate practical value for multiple NLP applications. For language model training, cleaner data reduces the risk of models learning spurious patterns from

noise and contamination. This proves particularly valuable for less-resourced languages where limited data availability means every training example significantly impacts model quality.

Our work on specialized tokenization (Chelombitko and Komissarov 2024) directly benefited from HQWiki-processed data. Training BPE tokenizers on clean monolingual corpora produced vocabulary distributions that better represent actual language usage, with most tokens unique to each language and meaningful overlap only between closely related languages. The resulting tokenizers substantially outperformed general-purpose alternatives, improving both model efficiency and performance.

The systematic removal of cross-lingual contamination and metadata artifacts enables models to capture genuine language features more accurately. This facilitates deeper linguistic analysis, supporting research into morphology, syntax, and semantic patterns. For languages with complex morphological systems, the improved data quality proves essential for developing accurate analysis tools and understanding linguistic structure.

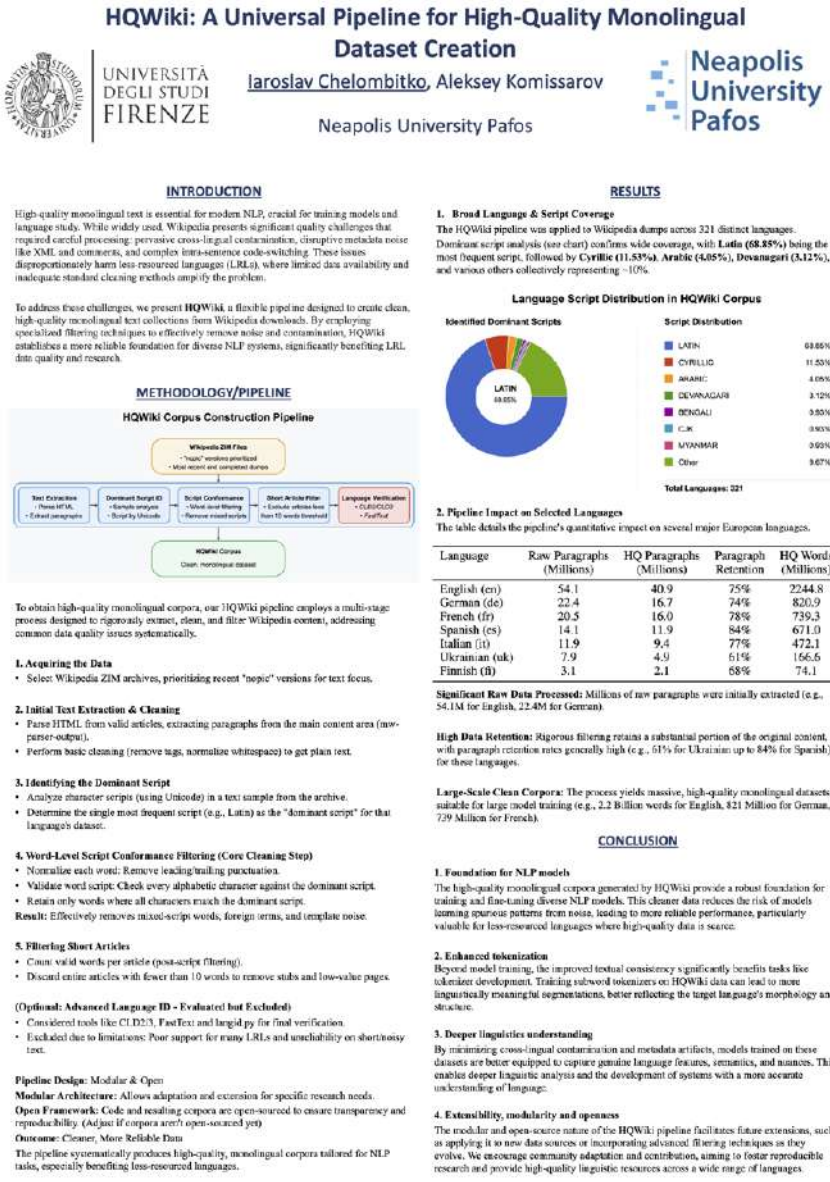
Quality assessment frameworks (Chelombitko, Safronov, and Komissarov 2024) benefit from standardized, clean corpora for evaluating tokenizer performance across languages. The consistent data quality provided by HQWiki enables fair comparisons and reliable benchmarking, advancing research in multilingual NLP systems.

The modular, open-source design facilitates future extensions and community contributions. Researchers can adapt the pipeline to new data sources beyond Wikipedia, incorporate emerging filtering techniques, or customize validation criteria for specific research requirements. By releasing both the framework and resulting corpora openly, we enable reproducible research and provide foundational resources across diverse language families.

5. Conclusion

HQWiki addresses fundamental challenges in monolingual corpus creation through systematic, script-based filtering and modular architecture. Processing Wikipedia data across 321 languages, the pipeline demonstrates that thoughtful engineering of data preparation workflows significantly improves corpus quality while maintaining high data retention. The resulting datasets provide robust foundations for language model training, specialized NLP tool development, and linguistic research, particularly benefiting less-resourced languages where data quality critically impacts all downstream applications.

Our work on tokenization and quality evaluation demonstrates the practical value of clean monolingual corpora for advancing multilingual NLP. We encourage community adaptation and contribution to the open-source framework, aiming to foster reproducible research and provide high-quality linguistic resources across the world's languages. Future work will explore application to additional data sources and integration of complementary quality validation techniques as they emerge.



HQWiki: A Universal Pipeline for High-Quality Monolingual Dataset Creation

UNIVERSITÀ DEGLI STUDI FIRENZE

Iaroslav Chelombitko, Aleksey Komissarov

Neapolis University Pafos

Neapolis University Pafos

INTRODUCTION

High-quality monolingual text is essential for modern NLP, crucial for training models and language study. While widely used, Wikipedia presents significant quality challenges that required careful processing: pervasive cross-lingual contamination, disruptive metadata noise like XML and comments, and complex intra-sentence code-switching. These issues disproportionately harm less-resourced languages (LRLs), where limited data availability and inadequate standard cleaning methods amplify the problem.

To address these challenges, we present **HQWiki**, a flexible pipeline designed to create clean, high-quality monolingual text collections from Wikipedia downloads. By employing specialized filtering techniques to effectively remove noise and contamination, HQWiki establishes a more reliable foundation for diverse NLP systems, significantly benefiting LRL data quality and research.

METHODOLOGY/PIPELINE

HQWiki Corpus Construction Pipeline

To obtain high-quality monolingual corpora, our HQWiki pipeline employs a multi-stage process designed to rigorously extract, clean, and filter Wikipedia content, addressing common data quality issues systematically.

1. Acquiring the Data

- Select Wikipedia ZIM archives, prioritizing recent "nopic" versions for text focus.

2. Initial Text Extraction & Cleaning

- Parse HTML from valid articles, extracting paragraphs from the main content area (mw-parser-output).
- Perform basic cleaning (remove tags, normalize whitespace) to get plain text.

3. Identifying the Dominant Script

- Analyze character scripts (using Unicode) in a text sample from the archive.
- Determine the single most frequent script (e.g., Latin) as the "dominant script" for that language's dataset.

4. Word-Level Script Conformance Filtering (Core Cleaning Step)

- Normalize each word: Remove leading/trailing punctuation.
- Validate word script: Check every alphabetic character against the dominant script.
- Retain only words where all characters match the dominant script.

Result: Effectively removes mixed-script words, foreign terms, and template noise.

5. Filtering Short Articles

- Count valid words per article (post-script filtering).
- Discard entire articles with fewer than 10 words to remove stubs and low-value pages.

(Optional: Advanced Language ID - Evaluated but Excluded)

- Considered tools like CLD2/3, FastText and langid.py for final verification.
- Excluded due to limitations: Poor support for many LRLs and unreliability on short/noisy text.

Pipeline Design: Modular & Open

Modular Architecture: Allows adaptation and extension for specific research needs.

Open Framework: Code and resulting corpora are open-sourced to ensure transparency and reproducibility. (Adjust if corpora aren't open-sourced yet)

Outcome: Cleaner, More Reliable Data

The pipeline systematically produces high-quality, monolingual corpora tailored for NLP tasks, especially benefiting less-resourced languages.

RESULTS

1. Broad Language & Script Coverage

The HQWiki pipeline was applied to Wikipedia dumps across 321 distinct languages.

Dominant script analysis (see chart) confirms wide coverage, with **Latin (68.85%)** being the most frequent script, followed by **Cyrillic (11.53%)**, **Arabic (4.05%)**, **Devanagari (3.12%)**, and various others collectively representing ~10%.

Language Script Distribution in HQWiki Corpus

2. Pipeline Impact on Selected Languages

The table details the pipeline's quantitative impact on several major European languages.

Language	Raw Paragraphs (Millions)	HQ Paragraphs (Millions)	Paragraph Retention	HQ Words (Millions)
English (en)	54.1	40.9	75%	2244.8
German (de)	22.4	16.7	74%	820.9
French (fr)	20.5	16.0	78%	739.3
Spanish (es)	14.1	11.9	84%	671.0
Italian (it)	11.9	9.4	77%	472.1
Ukrainian (uk)	7.9	4.9	61%	166.6
Finnish (fi)	3.1	2.1	68%	74.1

Significant Raw Data Processed: Millions of raw paragraphs were initially extracted (e.g., 54.1M for English, 22.4M for German).

High Data Retention: Rigorous filtering retains a substantial portion of the original content, with paragraph retention rates generally high (e.g., 61% for Ukrainian up to 84% for Spanish) for these languages.

Large-Scale Clean Corpora: The process yields massive, high-quality monolingual datasets suitable for large model training (e.g., 2.2 Billion words for English, 821 Million for German, 739 Million for French).

CONCLUSION

1. Foundation for NLP models

The high-quality monolingual corpora generated by HQWiki provide a robust foundation for training and fine-tuning diverse NLP models. This cleaner data reduces the risk of models learning spurious patterns from noise, leading to more reliable performance, particularly valuable for less-resourced languages where high-quality data is scarce.

2. Enhanced tokenization

Beyond model training, the improved textual consistency significantly benefits tasks like tokenizer development. Training subword tokenizers on HQWiki data can lead to more linguistically meaningful segmentations, better reflecting the target language's morphology and structure.

3. Deeper linguistics understanding

By minimizing cross-lingual contamination and metadata artifacts, models trained on these datasets are better equipped to capture genuine language features, semantics, and nuances. This enables deeper linguistic analysis and the development of systems with a more accurate understanding of language.

4. Extensibility, modularity and openness

The modular and open-source nature of the HQWiki pipeline facilitates future extensions, such as applying it to new data sources or incorporating advanced filtering techniques as they evolve. We encourage community adaptation and contribution, aiming to foster reproducible research and provide high-quality linguistic resources across a wide range of languages.

Figure 1 – Poster presented at "Wikidata and Research". Conference Proceedings, Florence, 5-6 giugno 2025.

References

Chelombitko, Iaroslav, and Aleksey Komissarov. 2024. "Specialized Monolingual BPE Tokenizers for Uralic Languages Representation in Large Language Models." In *Proceedings of the 8th International Workshop on Computational Linguistics for Uralic Languages*, 115–25. Stroudsburg, PA: Association for Computational Linguistics. https://aclanthology.org/2024.iwclul-1.11/

Chelombitko, Iaroslav, Egor Safronov, and Aleksey Komissarov. 2024. "Qtok: A Comprehensive Framework for Evaluating Multilingual Tokenizer Quality in Large Language Models." *arXiv preprint* arXiv:2410.12989. https://arxiv.org/abs/2410.12989

Authors

Kholoud Saad Alghamdi is an assistant professor at the University of Jeddah. Her research interests include knowledge graphs, distributed AI and socio-technical systems.

Gabriel Maia Rocha Amaral is a research fellow and PhD candidate in Computer Science at King's College London. His research interests include artificial intelligence, knowledge graphs, fact verification, and machine learning.

Elena Avellino is a librarian at the Ecole française de Rome. She is particularly involved in authority management and projects related to linked data and open science.

Stefano Bargioni is deputy director of the Library at the Pontificia Università della Santa Croce (Rome). Since 2011 he has adopted Koha as an open-source ILS and integrated it with Wikidata starting in 2019. He is a founding member of the Koha Italian Group and the Wikidata Group for Museums, Archives, and Libraries (GWMAB) and contributes to research on authority control, metadata, and Linked Open Data.

Lily Baumeister is a researcher in Art History and Digital Humanities. Her work focuses on the digital modelling, analysis, and curation of art-historical data and collaborative research infrastructures. Her research has been recognized with the Heinrich Wölfflin Prize.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Elena Marangoni, Camillo Carlo Pellizzari di San Girolamo (edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*, © 2026 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-1010-2, DOI 10.36253/979-12-215-1010-2

Andrea Benedetti holds a PhD in design from Politecnico of Milan and is a researcher with the DensityDesign Lab. His work focuses on the intersection of data visualization, creative programming, and communication design to raise awareness of how data is produced online by users. Since 2018, he has been part of the DensityDesign Lab, contributing to teaching and research projects.

Giovanni Bergamin is a librarian. He currently leads Digital Humanities research at Logos-RI. He is involved in the LIS field as a teacher, author, and seminar speaker. Previously, he served as Head of Information Technology Services at the Biblioteca Nazionale Centrale Firenze (1990–2017) and was a Board Member of the Associazione Italiana Biblioteche (AIB) from 2017 to 2023.

Ángeles Briones holds a PhD in design from Politecnico of Milan. Former member of the DensityDesign Lab, she is an adjunct professor at the Design School of Politecnico of Milan. She is interested in the intersection of data visualization and data activism with a focus on human rights. She addressed this areas in her doctoral thesis *Disclose to Tell: A Data Design Framework for Alternative Narratives* (2019).

Silvia Bruni works at the Social Sciences Library of the University of Florence. She deals in particular with collaboration projects with Wikimedia (Wikistazione), activities in prison (Penitentiary University Centre), projects on writing, qualitative research, involvement of students in the activities of cultural institutions. She is part of the Wikidata working group of the University Library System.

Gustavo Candela holds a PhD in Computer Science from the University of Alicante where he is a Lecturer in Computer Science. He has been actively involved in the integration and quality of Linked Open Data in libraries.

Marianna Cappellina is the head of conservation at the National Museum of Science and Technology in Milan.

Elena Cencetti works for Ales S.p.A. in the Semantic Indexing Tools and Searches sector of the BNCF. She is completing her master's degree in Organization and management of archives, cataloging and metadating at UNIFI, with a thesis on a domain ontology in Wikidata, interoperable with the New Subjectory Thesaurus. She is part of the Wikidata-UNIFI Group for the Photography Ontology Project.

Iaroslav Chelombitko is a researcher at Neapolis University Pafos and member of aglabx independent laboratory, specializing in computational linguistics with focus on tokenization methods for low-resourced languages. His research develops machine learning models for multilingual text recognition and morphological analysis.

Marco Chemello is a Wikimedian since 2004 and administrator of Italian Wikipedia (2005–). He has served as Wikimedian in Residence for multiple Italian GLAMs, including the National Museum of Science and Technology in Milan, ICAR, BEIC and the University of Padua, and works as GLAM specialist at Wikimedia Italia.

Dario Crespi has been a volunteer editor of the Italian Wikipedia since 2007, serving as an administrator since 2013 and as an oversighter since 2024. Since 2013, he has also contributed as an editor for Wikidata and OpenStreetMap. After a period as a library assistant, in 2022 he joined the staff of Wikimedia Italia, where he supports and organizes national projects and volunteers.

Tiago Trindade Cruz holds a PhD in heritage studies from Faculdade de Letras da Universidade do Porto (2022). He has been a senior researcher in the project ID-SCAPES since September 2024. Since the 2021–22 academic year, he has been a visiting assistant professor at FLUP, where he teaches history of contemporary architecture. He has been an integrated researcher at CITCEM/FLUP and a collaborator at CEAU/FAUP since 2023.

Leonardo D'Addario is a PhD candidate in Ancient Greek Language, Literature, and History at Leipzig University. Under the supervision of PD Monica Berti and Prof. Stephan Schorn, he is currently working on the MECANO PhD Project "Detecting and Retrieving Lost Historians", which investigates fragmentary Greek historians and their reception and use in Polybius, also with the aid of digital tools.

Alida Daniele works at the Biomedical Library of the University of Florence. She deals in particular with support for Open Access and research evaluation, courses for users and assistance with bibliographic research. She is part of the Wikidata working group of the University Library System.

Tommaso Elli holds a PhD in design focused on visualization for literary studies. He is adjunct professor at Politecnico di Milano's Design School and taught at other design schools including POLI.Design, ISIA Urbino, and Domus Academy. Since 2023, he has contributed to the MUSA project (NRRP) exploring fashion retail, sustainability and circularity. Since 2016 he has collaborated with DensityDesign on research and teaching in information design and data visualization. From September 2026 he joins the University of Bern's Bit Philology project in Digital Humanities.

Alice Santiago Faria is an auxiliary researcher at CHAM – Centre for the Humanities, FCSH, Universidade NOVA de Lisboa, where she coordinates the Digital Paradigm thematic line. Her research focuses on the history of the colonial built environment in the Portuguese Empire during the long 19th century.

Andrea Farina is post-doctoral research associate in Latin Linguistics in the Department of Digital Humanities at King's College London, where he leads King's Linguistics Network. His research focuses on quantitative analyses in historical languages. He is senior editor of the Journal of Open Humanities Data.

Lennart Finke is a master's student of Mathematics and Statistics at Cambridge, ETH and the University of Göttingen having previously earned his Bachelor's degree in Mathematics from the University of Göttingen.

Claudio Forziati is a librarian at the University of Naples Federico II. He serves as the contact person for the SHARE Catalogue technical team and is a founding member of the Wikidata Group for Museums, Archives, and Libraries (GWMAB). Since 2017, he has published several works exploring the interaction between Wikimedia projects and libraries.

Chiara Franceschini is professor of Early Modern Art History at Ludwig-Maximilians-Universität Munich and holds a concurrent professorship at the University of Pisa. Her research focuses on Renaissance and early modern visual culture. She leads the ERC Advanced Grant project ARCHIATER on the art and architecture of early modern hospitals.

Nathan Schneider Gavenski is a PhD candidate in Computer Science. His research interests are imitation learning, multi-agent systems and machine learning.

Donatella Gavrilovich is associate professor of History of Theatre and Digital Technologies at University of Rome "Tor Vergata" and PI of the HYPERSTAGE project, funded by Next Generation EU. She is the director of the IDOS_ARTS Research Centre and the ASPA International Scientific Journal.

Andrea Girometti is a research fellow at the University of Florence and an adjunct professor at the University of Urbino. His research interests and main publications focus on political sociology, sociological theory (with particular attention to the Bourdieusian approach) and the history of political thought.

Janek Große is a master's student of Mathematics and Statistics at Cambridge, ETH and the University of Göttingen having previously earned his Bachelor's degree in Mathematics from the University of Göttingen.

Maxime Guénette is a PhD candidate in History at the University of Montréal. His research mainly focuses on sacred spaces in Roman Britain, digital gazetteers and Linked Open Data. Over the years, he also worked on several digital humanities project, such as the Anthologia Graeca project.

Stefan Haas is professor of Historical Science at the University of Göttingen.

Alessio Ionna holds a PhD in Humanism and Technology from the University of Macerata, with a thesis entitled "Semantic Patronage: Representing the Artistic Patronage of the Buonaccorsi Family of Macerata in a Web 3.0 Collaborative Environment".

Jongmo Kim is a research associate at King's College London. His research interests include ontology engineering, and knowledge graphs.

Aleksey Komissarov is a scientist at aglabx independent laboratory, working as bioinformatician and AI researcher. His expertise spans artificial intelligence applications, dataset quality improvement, and natural language processing for less-resourced languages.

Maximilian Kristen is a researcher at Ludwig-Maximilians-Universität Munich working at the intersection of art history and computer science. His research explores machine learning, computer vision, and computational methods for the analysis of large-scale art-historical image corpora.

Frieder Leipold is an art historian and researcher at Ludwig-Maximilians-Universität Munich. His work focuses on premodern architecture and art history as well as digital research infrastructures. He develops data-driven methods and knowledge bases for art-historical research within the digital humanities.

Valentina Lepore works at the Humanistic Library of Florence. She is a Wikibase expert and she is part of the Wikidata-UNIFI Group.

Solange Ferraz de Lima is an associate professor at the Museu Paulista da Universidade de São Paulo. She is also a curator at the Museu Paulista/University of São Paulo, specialising in material and visual culture. She was director of the Museu Paulista da Universidade de São Paulo (2016–20). She is coordinator of the Digital Strategies Group at Museu Paulista/USP.

Isabella Limmer is a researcher in art history with a focus on digital methods. Her work examines the structured documentation, annotation, and analysis of visual and textual sources for art-historical research. Her research has been recognized with the Heinrich Wölfflin Prize.

Luca Lukacevic is a master's student of Mathematics and Statistics at Cambridge, ETH and the University of Göttingen having previously earned his Bachelor's degree in Mathematics from the University of Göttingen.

Tania Maio is a librarian at the University of Salerno; she is a member of the AIB National Commission for Special Libraries, Archives, and Personal Collections. She participates in the Wikidata Group for Museums, Archives, and

Libraries (GWMAB) and collaborates with Wikimedia projects to share open bibliographic data and promote free, collaborative contents.

Arkadiusz Margraf works at Poznań Supercomputing and Networking Center (Poland). He is involved in digitisation and digital humanities projects such as ATRIUM.

Elena Martellotta is an archivist with a background in art-history. She works for a UNESCO-heritage site and previously collaborated with Wikimedia Italia on two open data projects tied to cultural heritage, in addition to contributing to the Wiki Loves Monuments initiative.

Annantonia Martorano is associate professor of Archival Science at the University of Florence. She is co-editor of the journal JLIS.it. Her research interests includes public and private archives and their impact on society.

Michele Mauri is an associate professor at the Politecnico of Milan's Design Department and co-director of the DensityDesign Lab. Within the laboratory, he coordinates research, design, and the development of projects related to the visual communication of data and information, particularly those involving born-digital data and digital methods.

Barbara McGillivray is a senior lecturer in Digital and Computational Humanities and Cultural Computation at King's College London, where she leads the Computational Humanities Research Group. She is editor-in-chief of the Journal of Open Humanities Data and PI of the project COALA, successfully evaluated by the ERC.

Sidh Losa Mendiratta is an assistant professor at the Faculty of Architecture of the University of Porto, and an Integrated Researcher at the Centre for Studies in Architecture and Urbanism. Since September 2024 he is the Coordinator of the research project ID-SCAPES funded by a Consolidator Grant of the European Research Council (2024–29). He holds a PhD in Architecture from the University of Coimbra (2012).

Sabine von Mering is a botanist and data scientist at the Museum für Naturkunde Berlin, where she leads the Data for Nature research group. Her work focuses on opening up, linking and contextualising collection data. She uses Wikidata as a research tool and studies networks of collection agents and research expeditions.

Albert Meroño-Peñuela is a senior lecturer in Computer Science at King's College London. His research interests include knowledge graphs, the semantic web, linked data APIs, and cultural AI.

Daniel Mietchen is a biophysicist and data scientist working on improving collaborative workflows for knowledge sharing.

Alessia Minnella is a PhD candidate at the Department of Environmental Science and Policy, University of Milan.

Alessandra Moi is a librarian at University Milano-Bicocca and functional analyst at @CULT software house. She is the coordinator of the AIB Study Group TBID. She is a founding member of the Wikidata Group for Museums, Archives, and Libraries (GWMAB).

Rossana Morriello is associate professor in Library and Information Science at the University of Florence. She holds a PhD in Library Science from Sapienza University of Rome. Her research interests include scholarly publishing, digital libraries, data librarianship, and gender issues in librarianship and publishing.

Fabrizio Nunnari works at the Science Library (BS) of the University of Florence. In particular, he deals with the management of periodicals, the promotion of library services and assistance with bibliographic research. He is part of the University Library System Working Group on Collection Management and of the Wikidata Working Group of the University Library System.

Ioanna Papazoglou is a research software engineer at IT-Gruppe Geisteswissenschaften, Ludwig-Maximilians-Universität Munich. She develops Linked Open Data pipelines for the Amateur Theatre Wiki: Wikidata reconciliation, FAIR data modeling, MediaWiki integrations.

Valeria Paraninfi is currently a research fellow in the Hyperstage PRIN 2022 project at the University of Rome "Tor Vergata". From 2014 to 2016, she collaborated with a research group at the same university on the design of the ontology for a knowledge base on the Performing Arts. She is also a choreographer and teaches dance history in the CSEN training courses since 2019.

Manuela Parrilli is a research fellow at the Department of History, Archaeology, Geography, Art and Entertainment (SAGAS) at the University of Florence, working on the project "The Archival and Library Collection of Giacomo Caputo: Archaeology and Architectural Restoration in an Author's Library".

Iolanda Pensa is senior researcher and Head of Culture and Territory at the Institute of Design, research and innovation, Department of Environment, Constructions and Design, SUPSI University of Applied Sciences and Arts of Southern Switzerland.

João Alexandre Peschanski is the executive director of Wiki Movimento Brasil. He holds a PhD in sociology from the University of Wisconsin-Madison. He is an associate researcher of the Research, Innovation and Dissemination Center for Neuromathematics (NeuroMat) funded by the São Paulo Research Foundation at the University of São Paulo.

Silvia Pezzoli is associate professor of Sociology of Cultural and Communication Processes at the University of Florence. She is the director of the master's program in Strategies of Public and Political Communication and teaches Sociology of the Media and Consumption and Society. Her main research interests concern communicative processes related to both collective and individual representations.

Sarah Sophie Pohl is a master's student of Mathematics and Statistics at Cambridge, ETH and the University of Göttingen having previously earned her Bachelor's degree in Mathematics from the University of Göttingen.

Miriam Redi is a research manager at Wikimedia Foundation. She is interested in intelligent multimedia understanding, multimedia indexing and machine learning.

Giuseppe Resta is an integrated researcher at CEAU (Centro de Estudos de Arquitectura e Urbanismo), Faculdade de Arquitectura da Universidade do Porto, Portugal. He has been a senior researcher in the project ID-SCAPES since October 2024. Resta received his PhD in architecture from University of Roma Tre.

Odinaldo Rodrigues is a reader in Artificial Intelligence at King's College London. His research focuses on knowledge representation and reasoning.

Cezary Rosiński holds a PhD; he is an assistant professor, documentation and data specialist at IBL PAN. Interested in the relationship between space and literature, data processing for humanities research and Linked Open Data. Involved in digital humanities projects (GRAPHIA, FASCA).

Elisa Saltetto is a librarian at the Ecole française de Rome. She is particularly involved in user service and collection management, actively supporting projects related to the development of Linked Open Data and open science.

Toni Sant is associate professor of Digital Curation at the University of Salford in the UK. He has been active in the global Wikimedia movement since 2010, and currently holds the role of Chair of the Steering Committee for the Wikimedia CEE Hub, operating in the Central and Eastern European region.

Donatella Selva is an assistant professor at the University of Florence. She holds a PhD in Sociology of Cultural and Communicative Processes. Her research interests include the interplay between technology and society, the political impact of technology and the study of gender dynamics. She is also research fellow at the Observatory on Gender, Inclusion and Democracy at LUISS University.

Lorenzo Sergi is researcher at the University of Tuscia and was research fellow at the University of Florence. His interests focus on archives and analog or digi-

tal tools for representing cultural heritage. He participates, as a section editor of the JLIS.it journal, in the project of the Wikidata Group for Museums, Archives, and Libraries (GWMAB) regarding Italian LIS journals.

Elena Simperl is professor of Computer Science at King's College London. Her research interests are at the intersection between AI and social computing.

Valentina Sonzini is associate professor in Bibliology, History of Libraries and History of Book at the University of Florence. Her field of research is the History of printing and publishing in relation to the modern period, with a particular focus on Italian women printers, and with some forays into the contemporary sphere as regards the presence of women in the book industry.

Enrique Tabone is an artist and PhD candidate at the University of Salford, where she works as a data scientist with the Digital Curation Lab at MediaCityUK.

Francesca Tornari is a PhD candidate at the Department of Environmental Science and Policy of the University of Milan.

Stefano Pierpaolo Marcello Trasatti is associate professor at the Department of Environmental Science and Policy of the University of Milan.

Mathilde Verstraete is a PhD candidate in literature and Digital Humanities at the University of Montréal. With a background in classics, she specializes in Digital Classics within the Research Laboratory on Digital Textualities. Her doctoral research explores the production of critical editions of Greek texts in digital environments. She also coordinates the Anthologia Graeca project.

Marcello Vitali-Rosati is a philosopher and specialist in digital publishing. He is professor in the Department of French Literatures at the University of Montréal where he leads the Research Laboratory on Digital Textualities; he also leads the Chair of excellence in digital publishing at the University of Rouen. He develops a philosophical reflection on what the world becomes in the era of digital technologies.

Elisabetta Viti is library officer at BNCF in the Semantic Indexing Tools and Searches sector. Member of the Technical Committee UNI/CT 014/SC04 "Automation and documentation", representing the Ministry of Culture. Member of the Cataloging, indexing, Linked Open Data and semantic web (CILW) study group of the Italian Library Association since 2020. Member of the Wikidata-UNIFI Group for the Photography Ontology Project.

Meike Wagner is professor of Theatre Studies at Ludwig-Maximilians-Universität Munich. Her research interests include theatre history (18th–19th century), contemporary performance, puppet theatre and object performance, theatre and media, curatorial dramaturgy.

Yiwen Xing is a research assistant in Data Visualization at the University of Oxford. Her research focuses on information visualization, human-computer interaction, and data science, with an emphasis on interdisciplinary applications.

Bohui Zhang is a PhD candidate in Computer Science at King's College London. His research interests include knowledge graphs, ontology engineering, explainable AI and natural language processing.

Yihang Zhao is a PhD candidate in Computer Science at King's College London. His research interests include Human–Computer Interaction (HCI), generative AI (GenAI), knowledge graphs (KG), and ontology engineering.

STUDI E SAGGI

TITOLI PUBBLICATI

ARCHITETTURA, STORIA DELL'ARTE E ARCHEOLOGIA

Acciai Serena, *Sedad Hakki Eldem. An aristocratic architect and more*

Bartoli Maria Teresa, Lusoli Monica (a cura di), *Le teorie, le tecniche, i repertori figurativi nella prospettiva d'architettura tra il '400 e il '700. Dall'acquisizione alla lettura del dato*

Bartoli Maria Teresa, Lusoli Monica (a cura di), *Diminuzioni e accrescimenti. Le misure dei maestri di prospettiva*

Benelli Elisabetta, *Archetipi e citazioni nel fashion design*

Benzi Sara, Bertuzzi Luca, *Il Palagio di Parte Guelfa a Firenze. Documenti, immagini e percorsi multimediali*

Betti Marco, Brovadan Carlotta Paola (a cura di), *Donum. Studi di storia della pittura, della scultura e del collezionismo a Firenze dal Cinquecento al Settecento*

Biagini Carlo (a cura di), *L'Ospedale degli Infermi di Faenza. Studi per una lettura tipo-morfologica dell'edilizia ospedaliera storica*

Bologna Alberto, *Pier Luigi Nervi negli Stati Uniti. 1952-1979. Master Builder of the Modern Age*

Bologna Roberto, Lambertini Anna, Solari Luca (edited by), *Nature and City. An Integrated Approach to Urban Biodiversity*

Eccheli Maria Grazia, Cavallo Claudia (a cura di), *Il progetto nei borghi abbandonati*

Eccheli Maria Grazia, Pireddu Alberto (a cura di), *Oltre l'Apocalisse. Arte, Architettura, Abbandono*

Fischer von Erlach Johann Bernhard, *Progetto di un'architettura istorica. Entwurff einer Historischen Architectur*, a cura di Rakowitz Gundula

Frati Marco, *"De bonis lapidibus conciis": la costruzione di Firenze ai tempi di Arnolfo di Cambio. Strumenti, tecniche e maestranze nei cantieri fra XIII e XIV secolo*

Gregotti Vittorio, *Una lezione di architettura. Rappresentazione, globalizzazione, interdisciplinarità*

Gulli Riccardo, *Figure. Ars e ratio nel progetto di architettura*

Gulli Riccardo, *Tempo e materia. Per un'etica del costruire*

Laurìa Antonio, Benesperi Beatrice, Costa Paolo, Valli Fabio, *Designing Autonomy at home. The ADA Project. An Interdisciplinary Strategy for Adaptation of the Homes of Disabled Persons*

Laurìa Antonio, Flora Valbona, Guza Kamela, *Five Albanian Villages. Guidelines for a Sustainable Tourism Development through the Enhancement of the Cultural Heritage*

Lisini Caterina, *Lezione di sguardi. Edoardo Detti fotografo*

Maggiora Giuliano, *Sulla retorica dell'architettura*

Mantese Eleonora (a cura di), *House and Site. Rudofsky, Lewerentz, Zanuso, Sert, Rainer*

Mazza Barbara, *Le Corbusier e la fotografia. La vérité blanche*

Mazzoni Stefania (a cura di), *Studi di Archeologia del Vicino Oriente. Scritti degli allievi fiorentini per Paolo Emilio Pecorella*

Méndez Baiges Maite, *Les Demoiselles d'Avignon and Modernism*

Messina Maria Grazia, *Paul Gauguin. Un esotismo controverso*

Paolucci Fabrizio (a cura di), *Epigrafia tra erudizione antiquaria e scienza storica. Ad honorem Detlef Heikamp*

Pezone Maria Gabriella, Convertini Angela Michela (a cura di), *NeaVia La villa napoletana. Antichità e natura tra Rinascimento e Barocco. Atti del Convegno Nazionale di studi*

Pireddu Alberto, *In limine. Between Earth and Architecture*

Pireddu Alberto, *In abstracto. Sull'architettura di Giuseppe Terragni*

Pireddu Alberto, *The Solitude of Places. Journeys and Architecture on the Edges*

Rakowitz Gundula, *Tradizione, traduzione, tradimento in Johann Bernhard Fischer von Erlach*

Tonelli Maria Cristina, *Industrial design: latitudine e longitudine. Una prima lezione*

Tonelli Maria Cristina (a cura di), *Giovanni Klaus Koenig. Un fiorentino nel dibattito nazionale su architettura e design (1924-1989)*

CULTURAL STUDIES

Candotti Maria Piera, *Interprétations du discours métalinguistique. La fortune du sutra A 1 1 68 chez*

Patañjali et Bhartrhari

Castorina Miriam, *In the garden of the world. Italy to a young 19th century Chinese traveler*

Castorina Miriam, Cucinelli Diego (edited by), *Food issues* 雲路. *Interdisciplinary Studies on Food in Modern and Contemporary East Asia*

Cucinelli Diego, Scibetta Andrea (edited by), *Tracing Pathways* 雲路. *Interdisciplinary Studies on Modern and Contemporary East Asia*

Graziani Michela, Casetti Lapo, Vuelta García Salomé (a cura di), *Nel segno di Magellano tra terra e cielo. Il viaggio nelle arti umanistiche e scientifiche di lingua portoghese e di altre culture europee in un'ottica interculturale*

Nesti Arnaldo, *Qual è la religione degli italiani?. Religioni civili, mondo cattolico, ateismo devoto, fede, laicità*

Nesti Arnaldo, *Per una mappa delle religioni mondiali*

Pedone Valentina, *A Journey to the West. Observations on the Chinese Migration to Italy*

Pedone Valentina, Sagiyama Ikuko (edited by), *Transcending Borders. Selected papers in East Asian studies*

Pedone Valentina, Castorina Miriam (edited by), *Words and visions around/about Chinese transnational mobilities* 流动

Rigopoulos Antonio, *The Mahanubhavs*

Sagiyama Ikuko, Castorina Miriam (edited by), *Trajectories. selected papers in East Asian studies* 軌跡

Sagiyama Ikuko, Pedone Valentina (edited by), *Perspectives on East Asia*

Squarcini Federico (edited by), *Boundaries, Dynamics and Construction of Traditions in South Asia*

Vanoli Alessandro, *Il mondo musulmano e i volti della guerra. Conflitti, politica e comunicazione nella storia dell'islam*

Vinci Renata (edited by), *Navigating the Mediterranean Through the Chinese Lens. Transcultural Narratives of the Sea Among Lands*

DIRITTO

Allegretti Umberto (a cura di), *Democrazia partecipativa. Esperienze e prospettive in Italia e in Europa*

Campus Mauro, Dorigo Stefano, Federico Veronica, Lazzerini Nicole (a cura di), *Pago, dunque sono (cittadino europeo). Il futuro dell'UE tra responsabilità fiscale, solidarietà e nuova cittadinanza europea*

Chiaromonte William, Vallauri Maria Luisa (a cura di), *Trasformazioni, valori e regole del lavoro. Scritti per Riccardo Del Punta*, vol. III, 2024

Cingari Francesco (a cura di), *Corruzione: strategie di contrasto. (legge 190/2012)*

Curreri Salvatore, *Democrazia e rappresentanza politica. Dal divieto di mandato al mandato di partito*

Curreri Salvatore, *Partiti e gruppi parlamentari nell'ordinamento spagnolo*

Del Punta Riccardo, *Trasformazioni, valori e regole del lavoro. Scritti scelti* sul *Diritto del lavoro,* vol. 1, a cura di William Chiaromonte e Maria Luisa Vallauri

Del Punta Riccardo, *Trasformazioni, valori e regole del lavoro. Scritti scelti* di *diritto del lavoro*, vol. 2, a cura di William Chiaromonte e Maria Luisa Vallauri

Federico Veronica, Fusaro Carlo (edited by), *Constitutionalism and democratic transitions. Lessons from South Africa*

Ferrara Leonardo, Sorace Domenico, Cavallo Perin Roberto, Police Aristide, Saitta Fabio (a cura di), *A 150 anni dell'unificazione amministrativa italiana. Vol. I. L'organizzazione delle pubbliche amministrazioni tra Stato nazionale e integrazione europea*

Ferrara Leonardo, Sorace Domenico, De Giorgi Cezzi Gabriella, Portaluri Pier Luigi (a cura di), *A 150 anni dall'unificazione amministrativa italiana. Vol. II. La coesione politico-territoriale*

Ferrara Leonardo, Sorace Domenico, Marchetti Barbara, Renna Mauro (a cura di), *A 150 anni dall'unificazione amministrativa italiana. Vol. III. La giuridificazione*

Ferrara Leonardo, Sorace Domenico, Civitarese Matteucci Stefano, Torchia Luisa (a cura di), *A 150 anni dall'unificazione amministrativa italiana. Vol. IV. La tecnificazione*

Ferrara Leonardo, Sorace Domenico, Cafagno Maurizio, Manganaro Francesco (a cura di), *A 150 anni dall'unificazione amministrativa italiana. Vol. V. L'intervento pubblico nell'economia*

Ferrara Leonardo, Sorace Domenico, Chiti Edoardo, Gardini Gianluca, Sandulli Aldo (a cura di), *A*

150 anni dall'unificazione amministrativa italiana. Vol. VI. Unità e pluralismo culturale

Ferrara Leonardo, Sorace Domenico, Comporti Gian Domenico (a cura di), *A 150 anni dall'unificazione amministrativa italiana. Vol. VII. La giustizia amministrativa come servizio (tra effettività ed efficienza)*

Ferrara Leonardo, Sorace Domenico, Bartolini Antonio, Pioggia Alessandra (a cura di), *A 150 anni dall'unificazione amministrativa italiana. Vol. VIII. Cittadinanze amministrative*

Fiorita Nicola, *L'Islam spiegato ai miei studenti. Otto lezioni su Islam e diritto*

Fiorita Nicola, *L'Islam spiegato ai miei studenti. Undici lezioni sul diritto islamico. II edizione riveduta e ampliata*

Fossum John Erik, Menendez Agustin José, *La peculiare costituzione dell'Unione Europea*

Gregorio Massimiliano, *Le dottrine costituzionali del partito politico. L'Italia liberale*

Lucarelli Paola (a cura di), *Giustizia sostenibile. Sfide organizzative e tecnologiche per una nuova professionalità*

Palazzo Francesco, Bartoli Roberto (a cura di), *La mediazione penale nel diritto italiano e internazionale*

Ragno Francesca, *Il rispetto del principio di pari opportunità. L'annullamento della composizione delle giunte regionali e degli enti locali*

Sorace Domenico (a cura di), *Discipline processuali differenziate nei diritti amministrativi europei*

Trocker Nicolò, De Luca Alessandra (a cura di), *La mediazione civile alla luce della direttiva 2008/52/CE*

Urso Elena (a cura di), *Le ragioni degli altri. Mediazione e famiglia tra conflitto e dialogo: una prospettiva comparatistica ed interdisciplinare*

Urso Elena, *La mediazione familiare. Modelli, principi, obiettivi*

ECONOMIA

Ammannati Francesco, *Per filo e per segno. L'arte della lana a Firenze nel Cinquecento*

Bardazzi Rossella (edited by), *Economic multisectoral modelling between past and future. A tribute to Maurizio Grassini and a selection of his writings*

Bardazzi Rossella, Ghezzi Leonardo (edited by), *Macroeconomic modelling for policy analysis*

Barucci Piero, Bini Piero, Conigliello Lucilla (a cura di), *Economia e Diritto in Italia durante il Fascismo. Approfondimenti, biografie, nuovi percorsi di ricerca*

Barucci Piero, Bini Piero, Conigliello Lucilla (a cura di), *Il Corporativismo nell'Italia di Mussolini. Dal declino delle istituzioni liberali alla Costituzione repubblicana*

Barucci Piero, Bini Piero, Conigliello Lucilla (a cura di), *Intellettuali e uomini di regime nell'Italia fascista*

Barucci Piero, Bini Piero, Conigliello Lucilla (a cura di), *I mille volti del regime. Opposizione e consenso nella cultura giuridica, economica e politica italiana tra le due guerre*

Barucci Piero, Bini Piero, Conigliello Lucilla (a cura di), *Le sirene del corporativismo e l'isolamento dei dissidenti durante il fascismo*

Bellanca Nicolò, Pardi Luca, *O la capra o i cavoli. La biosfera, l'economia e il futuro da inventare*

Bellanca Nicolò, *La forza delle comunità locali. Giacomo Becattini e la teoria della cultura sociale*

Bini Piero, Magliulo Antonio, Pagliai Letizia (a cura di), *L'Italia repubblicana in cammino. Ricostruzione, crescita, instabilità*

Cecchi Amos, *Paul M. Sweezy. Monopolio e finanza nella crisi del capitalismo*

Ciampi Francesco, *Come la consulenza direzionale crea conoscenza. Prospettive di convergenza tra scienza e consulenza*

Ciampi Francesco, *Knowing Through Consulting in Action. Meta-consulting Knowledge Creation Pathways*

Ciappei Cristiano (a cura di), *La valorizzazione economica delle tipicità rurali tra localismo e globalizzazione*

Ciappei Cristiano, Sani Azzurra, *Strategie di internazionalizzazione e grande distribuzione nel settore dell'abbigliamento. Focus sulla realtà fiorentina*

Ciappei Cristiano, Citti Paolo, Bacci Niccolò, Campatelli Gianni, *La metodologia Sei Sigma nei servizi. Un'applicazione ai modelli di gestione finanziaria*

Ciappei Cristiano, Mininni Giacomo, *A Religious Foundation for Global Business Ethics. CSR and*

Religious Ethics in an Age of Globalization

Garofalo Giuseppe (a cura di), *Capitalismo distrettuale, localismi d'impresa, globalizzazione*

Laureti Tiziana, *L'efficienza rispetto alla frontiera delle possibilità produttive. Modelli teorici ed analisi empiriche*

Lazzeretti Luciana, Cinti Tommaso, *La valorizzazione economica del patrimonio artistico delle città d'arte. Il restauro artistico a Firenze*

Lazzeretti Luciana, *Nascita ed evoluzione del distretto orafo di Arezzo, 1947-2001. Primo studio in una prospettiva ecology based*

Lazzeretti Luciana (edited by), *Art Cities, Cultural Districts and Museums. An economic and managerial study of the culture sector in Florence*

Lazzeretti Luciana (a cura di), *I sistemi museali in Toscana. Primi risultati di una ricerca sul campo*

Mastronardi Luigi, Romagnoli Luca (a cura di), *Metodologie, percorsi operativi e strumenti per lo sviluppo delle cooperative di comunità nelle aree interne italiane*

Meade Douglas S. (edited by), *In Quest of the Craft. Economic Modeling for the 21st Century*

Perrotta Cosimo, *Il capitalismo è ancora progressivo?*

Simoni Christian, *Approccio strategico alla produzione. Oltre la produzione snella*

Simoni Christian, *Mastering the dynamics of apparel innovation*

Stefani Gianluca, Cecchetti Maria Chiara, Bucelli Andrea, Martellozzo Federico, Pazienza Maria Grazia, Vecchio Bruno, *La proprietà fondiaria nelle aree interne. Un'indagine sulla Montagna Fiorentina e la Val Bisenzio*

FILOSOFIA

Baldi Massimo, Desideri Fabrizio (a cura di), *Paul Celan. La poesia come frontiera filosofica*

Barale Alice, *La malinconia dell'immagine. Rappresentazione e significato in Walter Benjamin e Aby Warburg*

Berni Stefano, Fadini Ubaldo, *Linee di fuga. Nietzsche, Foucault, Deleuze*

Borsari Andrea, *Schopenhauer educatore? Storia e crisi di un'idea tra filosofia morale, estetica e antropologia*

Brunkhorst Hauke, *Habermas*

Cambi Franco, Mari Giovanni (a cura di), *Giulio Preti. Intellettuale critico e filosofo attuale*

Cambi Franco, *Pensiero e tempo. Ricerche sullo storicismo critico: figure, modelli, attualità*

Casalini Brunella, Cini Lorenzo, *Giustizia, uguaglianza e differenza. Una guida alla lettura della filosofia politica contemporanea*

Desideri Fabrizio, Matteucci Giovanni (a cura di), *Dall'oggetto estetico all'oggetto artistico*

Desideri Fabrizio, Matteucci Giovanni (a cura di), *Estetiche della percezione*

Di Stasio Margherita, *Alvin Plantinga: conoscenza religiosa e naturalizzazione epistemologica*

Giovagnoli Raffaela, *Autonomy: a Matter of Content*

Honneth Axel, *Capitalismo e riconoscimento*, a cura di Solinas Marco

Michelini Luca, *Il nazional-fascismo economico del giovane Franco Modigliani*

Mindus Patricia, *Cittadini e no. Forme e funzioni dell'inclusione e dell'esclusione*

Perni Romina, *Pubblicità, educazione e diritto in Kant*

Sandrini Maria Grazia, *La filosofia di R. Carnap tra empirismo e trascendentalismo. In appendice:R. Carnap Sugli enunciati protocollariTraduzione e commento di E. Palombi*

Solinas Marco, *Psiche: Platone e Freud. Desiderio, sogno, mania, eros*

Trentin Bruno, *La città del lavoro. Sinistra e crisi del fordismo*, a cura di Ariemma Iginio

Valle Gianluca, *La vita individuale. L'estetica sociologica di Georg Simmel*

FISICA

Arecchi Fortunato Tito, *Cognizione e realtà*

Pelosi Giuseppe, Selleri Stefano, *The Roots of Maxwell's A Dynamical Theory of the Electromagnetic Field. Scotland and Tuscany, 'twinned by science'*

LETTERATURA, FILOLOGIA E LINGUISTICA

Antonucci Fausta, Vuelta García Salomé (a cura di), *Ricerche sul teatro classico spagnolo in Italia e oltralpe (secoli XVI-XVIII)*

Bastianini Guido, Lapini Walter, Tulli Mauro (a cura di), *Harmonia. Scritti di filologia classica in onore di Angelo Casanova*

Battistin Sebastiani Breno, Ferreira Leão Delfim (edited by), *Crises (Staseis) and Changes (Metabolai). Athenian Democracy in the Making*

Berté Monica (a cura di), *Intorno a Boccaccio/Boccaccio e dintorni 2021. Atti del Seminario internazionale di studi (Certaldo Alta, Casa di Giovanni Boccaccio, 9-10 settembre 2021)*

Bilenchi Romano, *The Conservatory of Santa Teresa*, edited by Klopp Charles, Nelson Melinda

Bresciani Califano Mimma (Vincenza), *Piccole zone di simmetria. Scrittori del Novecento*

Caracchini Cristina, Minardi Enrico (a cura di), *Il pensiero della poesia. Da Leopardi ai contemporanei. Letture dal mondo di poeti italiani*

Cauchi Santoro Roberta, *Beyond the Suffering of Being: Desire in Giacomo Leopardi and Samuel Beckett*

Colucci Dalila, *L'Eleganza è frigida e L'Empire des signes. Un sogno fatto in Giappone*

Dei Luigi (a cura di), *Voci dal mondo per Primo Levi. In memoria, per la memoria*

Fanucchi Sonia, Virga Anita (edited by), *A South African Convivio with Dante. Born Frees' Interpretations of the Commedia*

Ferrara Enrica Maria, *Il realismo teatrale nella narrativa del Novecento: Vittorini, Pasolini, Calvino*

Ferrone Siro, *Visioni critiche. Recensioni teatrali da «l'Unità-Toscana» (1975-1983)*, a cura di Megale Teresa, Simoncini Francesca

Francese Joseph, *Antonio Gramsci. Vita morale e fede socialista*

Francese Joseph, *Leonardo Sciascia e la funzione sociale degli intellettuali*

Francese Joseph, *Vincenzo Consolo: gli anni de «l'Unità» (1992-2012), ovvero la poetica della colpa-espiazione*

Franchini Silvia, *Diventare grandi con il «Pioniere» (1950-1962). Politica, progetti di vita e identità di genere nella piccola posta di un giornalino di sinistra*

Francovich Onesti Nicoletta, *I nomi degli Ostrogoti*

Frau Ombretta, Gragnani Cristina, *Sottoboschi letterari. Sei "case studies" fra Otto e Novecento. Mara Antelling, Emma Boghen Conigliani, Evelyn, Anna Franchi, Jolanda, Flavia Steno*

Frosini Giovanna, Zamponi Stefano (a cura di), *Intorno a Boccaccio/Boccaccio e dintorni. Atti del Seminario internazionale di studi (Certaldo Alta, Casa di Giovanni Boccaccio, 25 giugno 2014)*

Frosini Giovanna (a cura di), *Intorno a Boccaccio / Boccaccio e dintorni 2020. Atti del Seminario internazionale di studi (Certaldo Alta, Casa di Giovanni Boccaccio, 10-11 settembre 2020)*

Frosini Giovanna (a cura di), *Intorno a Boccaccio / Boccaccio e dintorni 2019. Atti del Seminario internazionale di studi (Certaldo Alta, Casa di Giovanni Boccaccio, 12-13 settembre 2019)*

Galigani Giuseppe, *Salomè, mostruosa fanciulla*

Gigli Daria, Magnelli Enrico (a cura di), *Studi di poesia greca tardoantica. Atti della Giornata di Studi Università degli Studi di Firenze, 4 ottobre 2012*

Giuliani Luigi, Pineda Victoria (edited by), *La edición del diálogo teatral (siglos XVI-XVII)*

Gori Barbara, *La grammatica dei clitici portoghesi. Aspetti sincronici e diacronici*

Gorman Michael, *I nostri valori, rivisti. La biblioteconomia in un mondo in trasformazione*, a cura di Guerrini Mauro

Graziani Michela (a cura di), *Un incontro lusofono plurale di lingue, letterature, storie, culture*

Graziani Michela, *Il Settecento portoghese e lusofono*

Graziani Michela, Abbati Orietta, Gori Barbara (a cura di), *La spugna è la mia anima. Omaggio a Piero Ceccucci*

Guerrini Mauro, Mari Giovanni (a cura di), *Via verde e via d'oro. Le politiche open access dell'Università di Firenze*

Guerrini Mauro, *De bibliothecariis. Persone, idee, linguaggi*, a cura di Stagi Tiziana

Keidan Artemij, Alfieri Luca (a cura di), *Deissi, riferimento, metafora. Questioni classiche di linguistica e filosofia del linguaggio*

López Castro Cruz Hilda, *America Latina aportes lexicos al italiano contemporaneo*

Mario Anna, *Italo Calvino. Quale autore laggiù attende la fine?*

Masciandaro Franco, *The Stranger as Friend: The Poetics of Friendship in Homer, Dante, and Boccaccio*

Nosilia Viviana, Prandoni Marco (a cura di), *Trame controluce. Il patriarca 'protestante' Cirillo Loukaris / Backlighting Plots. The 'Protestant' Patriarch Cyril Loukaris*

Pagliaro Annamaria, Zuccala Brian (edited by), *Luigi Capuana: Experimental Fiction and Cultural Mediation in Post-Risorgimento Italy*

Pestelli Corrado, *Carlo Antici e l'ideologia della Restaurazione in Italia*

Rosengarten Frank, *Through Partisan Eyes. My Friendships, Literary Education, and Political Encounters in Italy (1956-2013). With Sidelights on My Experiences in the United States, France, and the Soviet Union*

Ross Silvia, Honess Claire (edited by), *Identity and Conflict in Tuscany*

Totaro Luigi, *Ragioni d'amore. Le donne nel Decameron*

Turbanti Simona, *Bibliometria e scienze del libro: internazionalizzazione e vitalità degli studi italiani*

Vicente Filipa Lowndes, *Altri orientalismi. L'India a Firenze 1860-1900*

Virga Anita, *Subalternità siciliana nella scrittura di Luigi Capuana e Giovanni Verga*

Zamponi Stefano (a cura di), *Intorno a Boccaccio / Boccaccio e dintorni 2015. Atti del Seminario internazionale di studi (Certaldo Alta, Casa di Giovanni Boccaccio, 9 settembre 2015)*

Zamponi Stefano (a cura di), *Intorno a Boccaccio / Boccaccio e dintorni 2018. Atti del Seminario internazionale di studi (Certaldo Alta, Casa di Giovanni Boccaccio, 6-7 settembre 2018)*

Zamponi Stefano (a cura di), *Intorno a Boccaccio / Boccaccio e dintorni 2016. Atti del Seminario internazionale di studi (Certaldo Alta, Casa di Giovanni Boccaccio, 9 settembre 2016)*

Zamponi Stefano (a cura di), *Intorno a Boccaccio / Boccaccio e dintorni 2017. Atti del Seminario internazionale di studi (Certaldo Alta, Casa di Giovanni Boccaccio, 16 settembre 2017)*

MATEMATICA

De Bartolomeis Paolo, *Matematica. Passione e conoscenza. Scritti (1975-2016)*, a cura di Battaglia Fiammetta, Nannicini Antonella, Tomassini Adriano

MEDICINA

Mannaioni Pierfrancesco, Mannaioni Guido, Masini Emanuela, *Club drugs. Cosa sono e cosa fanno*

Saint Sanjay, Krein Sarah, Stock Robert W., *La prevenzione delle infezioni correlate all'assistenza. Problemi reali, soluzioni pratiche*, a cura di Bartoloni Alessandro, Gensini Gian Franco, Moro Maria Luisa, Rossolini Gian Maria

Saint Sanjay, Chopra Vineet, *Le 30 regole per la leadership in sanità*, a cura di Bartoloni Alessandro, Boddi Maria, Damone Rocco Donato, Giusti Betti, Mechi Maria Teresa, Rossolini Gian Maria

PEDAGOGIA

Bandini Gianfranco, Oliviero Stefano (a cura di), *Public History of Education: riflessioni, testimonianze, esperienze*

Mariani Alessandro (a cura di), *L'orientamento e la formazione degli insegnanti del futuro*

Nardi Andrea, *Il lettore 'distratto'. Leggere e comprendere nell'epoca degli schermi digitali*

Ranieri Maria, Luzzi Damiana, Cuomo Stefano (a cura di), *Il video a 360° nella didattica universitaria. Modelli ed esperienze*

POLITICA

Attinà Fulvio, Bozzo Luciano, Cesa Marco, Lucarelli Sonia (a cura di), *Eirene e Atena. Studi di politica internazionale in onore di Umberto Gori*

Bulli Giorgia, Tonini Alberto (a cura di), *Migrazioni in Italia: oltre la sfida. Per un approccio interdisciplinare allo studio delle migrazioni*

Caruso Sergio, *"Homo oeconomicus". Paradigma, critiche, revisioni*

Cipriani Alberto, Gramolati Alessio, Mari Giovanni (a cura di), *Il lavoro 4.0. La Quarta Rivoluzione industriale e le trasformazioni delle attività lavorative*

Cipriani Alberto (a cura di), *Partecipazione creativa dei lavoratori nella 'fabbrica intelligente'. Atti del Seminario di Roma, 13 ottobre 2017*

Cipriani Alberto, Ponzellini Anna Maria (a cura di), *Colletti bianchi. Una ricerca nell'industria e la discussione dei suoi risultati*

Corsi Cecilia (a cura di), *Felicità e benessere. Una ricognizione critica*

Corsi Cecilia, Magnier Annick (a cura di), *L'Università allo specchio. Questioni e prospettive*

Cruciani Sante, Del Rossi Maria Paola (a cura di), *Diritti, Europa, Federalismo. Bruno Trentin in*

prospettiva transnazionale (1988-2007)

De Boni Claudio, *Descrivere il futuro. Scienza e utopia in Francia nell'età del positivismo*

De Boni Claudio (a cura di), *Lo stato sociale nel pensiero politico contemporaneo. 1. L'Ottocento*

De Boni Claudio, *Lo stato sociale nel pensiero politico contemporaneo. Il Novecento. Parte prima: Da inizio secolo alla seconda guerra mondiale*

De Boni Claudio (a cura di), *Lo stato sociale nel pensiero politico contemporaneo. II Novecento. Parte seconda: dal dopoguerra a oggi*

Del Punta Riccardo (a cura di), *Valori e tecniche nel diritto del lavoro*

Gramolati Alessio, Mari Giovanni (a cura di), *Bruno Trentin. Lavoro, libertà, conoscenza*

Gramolati Alessio, Mari Giovanni (a cura di), *Il lavoro dopo il Novecento: da produttori ad attori sociali. La città del lavoro di Bruno Trentin per un' «altra sinistra»*

Grassi Stefano, Morisi Massimo (a cura di), *La cittadinanza tra giustizia e democrazia. Atti della giornata di Studi in memoria di Sergio Caruso*

Lombardi Mauro, *Transizione ecologica e universo fisico-cibernetico. Soggetti, strategie, lavoro*

Lombardi Mauro, *Fabbrica 4.0: I processi innovativi nel Multiverso fisico-digitale*

Marasco Vincenzo, *Coworking. Senso ed esperienze di una forma di lavoro*

Mari Giovanni, Ammannati Francesco, Brogi Stefano, Faitini Tiziana, Fermani Arianna, Seghezzi Francesco, Tonarelli Annalisa (a cura di), *Idee di lavoro e di ozio per la nostra civiltà*

Molteni Tagliabue Giovanni, *Rationalized and Extended Democracy. Inserting Public Scientists into the Legislative/Executive Framework, Reinforcing Citizens' Participation*

Nacci Michela (a cura di), *Nazioni come individui. Il carattere nazionale fra passato e presente*

Renda Francesco, Ricciuti Roberto, *Tra economia e politica: l'internazionalizzazione di Finmeccanica, Eni ed Enel*

Spini Debora, Fontanella Margherita (a cura di), *Il sogno e la politica da Roosevelt a Obama. Il futuro dell'America nella comunicazione politica dei democrats*

Spinoso Giovanni, Turrini Claudio, *Giorgio La Pira: i capitoli di una vita*

Tonini Alberto, Simoni Marcella (a cura di), *Realtà e memoria di una disfatta. Il Medio Oriente dopo la guerra dei Sei Giorni*

Trentin Bruno, *La libertà viene prima. La libertà come posta in gioco nel conflitto sociale. Nuova edizione con pagine inedite dei Diari e altri scritti*, a cura di Cruciani Sante

Zolo Danilo, *Tramonto globale. La fame, il patibolo, la guerra*

PSICOLOGIA

Aprile Luigi (a cura di), *Psicologia dello sviluppo cognitivo-linguistico: tra teoria e intervento*

Luccio Riccardo, Salvadori Emilia, Bachmann Christina, *La verifica della significatività dell'ipotesi nulla in psicologia*

SCIENZE E TECNOLOGIE AGRARIE

Surico Giuseppe, *Lampedusa: dall'agricoltura, alla pesca, al turismo*

SCIENZE NATURALI

Bessi Franca Vittoria, Clauser Marina, *Le rose in fila. Rose selvatiche e coltivate: una storia che parte da lontano*

Friis Ib, Demissew Sebsebe, Weber Odile, van Breugel Paulo, *Plants and vegetation of NW Ethiopia. A new look at Rodolfo E.G. Pichi Sermolli's results from the 'Missione di Studio al Lago Tana', 1937*

Sánchez Marcelo, *Embrioni nel tempo profondo. Il registro paleontologico dell'evoluzione biologica*

SOCIOLOGIA

Alacevich Franca, *Promuovere il dialogo sociale. Le conseguenze dell'Europa sulla regolazione del lavoro*

Alacevich Franca, Bellini Andrea, Tonarelli Annalisa, *Una professione plurale. Il caso dell'avvocatura fiorentina*

Battiston Simone, Mascitelli Bruno, *Il voto italiano all'estero. Riflessioni, esperienze e risultati di un'indagine in Australia*

Becucci Stefano (a cura di), *Oltre gli stereotipi. La ricerca-azione di Renzo Rastrelli sull'immigrazione*

cinese in Italia
Becucci Stefano, Desideri Agnese, Landucci Sandro, Pezzoli Silvia, *L'università su misura. Governance, valutazione e ranking in quattro paesi europei*
Becucci Stefano, Garosi Eleonora, *Corpi globali. La prostituzione in Italia*
Bettin Lattes Gianfranco (a cura di), *Giovani Jeunes Jovenes. Rapporto di ricerca sulle nuove generazioni e la politica nell'Europa del sud*
Bettin Lattes Gianfranco (a cura di), *Per leggere la società*
Bettin Lattes Gianfranco, Turi Paolo (a cura di), *La sociologia di Luciano Cavalli*
Burroni Luigi, Piselli Fortunata, Ramella Francesco, Trigilia Carlo (a cura di), *Città metropolitane e politiche urbane*
Catarsi Enzo (a cura di), *Autobiografie scolastiche e scelta universitaria*
Leonardi Laura (edited by), *Opening the european box. Towards a new Sociology of Europe*
Virginia Miller, *Child Sexual Abuse Inquiries and the Catholic Church: Reassessing the Evidence*
Virginia Miller, *Child Sexual Abuse Inquiries and the Catholic Church: Reassessing the Evidence. Second Edition, Revised and Expanded*
Nuvolati Giampaolo (a cura di), *Sviluppo urbano e politiche per la qualità della vita*
Nuvolati Giampaolo, *L'interpretazione dei luoghi. Flânerie come esperienza di vita*
Nuvolati Giampaolo, *Mobilità quotidiana e complessità urbana*
Ramella Francesco, Trigilia Carlo (a cura di), *Reti sociali e innovazione. I sistemi locali dell'informatica*
Rondinone Antonella, *Donne mancanti. Un'analisi geografica del disequilibrio di genere in India*

STATISTICA E DEMOGRAFIA
Salvini Maria Silvana, *Globalizzazione: e la popolazione?. Le relazioni fra demografia e mondo globalizzato*

STORIA E SOCIOLOGIA DELLA SCIENZA
Angotti Franco, Pelosi Giuseppe, Soldani Simonetta (a cura di), *Alle radici della moderna ingegneria. Competenze e opportunità nella Firenze dell'Ottocento*
Cabras Pier Luigi, Chiti Silvia, Lippi Donatella (a cura di), *Joseph Guillaume Desmaisons Dupallans. La Francia alla ricerca del modello e l'Italia dei manicomi nel 1840*
Califano Salvatore, Schettino Vincenzo, *La nascita della meccanica quantistica*
Cartocci Alice, *La matematica degli Egizi. I papiri matematici del Medio Regno*
Fontani Marco, Orna Mary Virginia, Costa Mariagrazia, *Chimica e chimici a Firenze. Dall'ultimo de' Medici al padre del Centro Europeo di Risonanze Magnetiche*
Guatelli Fulvio (a cura di), *Scienza e opinione pubblica. Una relazione da ridefinire*
Massai Veronica, *Angelo Gatti (1724-1798). Un medico toscano in terra di Francia*
Meurig Thomas John, *Michael Faraday. La storia romantica di un genio*
Schettino Vincenzo, *Scienza e arte. chimica, arti figurative e letteratura*
Teixeira Anderson Vichinkeski, *Constitutionnalisme transnational. Histoire, ontologie et épistémologie*

STUDI DI BIBLIOTECONOMIA
Marangoni Elena, Pellizzari di San Girolamo Camillo Carlo (a cura di / edited by), *Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025*

STUDI DI BIOETICA
Baldini Gianni, Soldano Monica (a cura di), *Tecnologie riproduttive e tutela della persona. Verso un comune diritto europeo per la bioetica*
Baldini Gianni, Soldano Monica (a cura di), *Nascere e morire: quando decido io? Italia ed Europa a confronto*
Baldini Gianni (a cura di), *Persona e famiglia nell'era del Biodiritto. Verso un diritto comune europeo per la bioetica*
Bucelli Andrea (a cura di), *Produrre uomini. Procreazione assistita: un'indagine multidisciplinare*
Costa Giovanni, *Scelte procreative e responsabilità. Genetica, giustizia, obblighi verso le generazioni future*

Galletti Matteo, Zullo Silvia (a cura di), *La vita prima della fine. Lo stato vegetativo tra etica, religione e diritto*

Galletti Matteo, *Decidere per chi non può. Approcci filosofici all'eutanasia non volontaria*

STUDI EUROPEI

Bosco Andrea, Guderzo Massimiliano (edited by), *A Monetary Hope for Europe. The Euro and the Struggle for the Creation of a New Global Currency*

Scalise Gemma, *Il mercato non basta. Attori, istituzioni e identità dell'Europa in tempo di crisi*

Wikidata e la ricerca. Condividere esperienze. Atti del Convegno, Firenze, 5-6 giugno 2025 / Wikidata and Research. Sharing Experiences. Conference Proceedings, Florence, 5-6 giugno 2025. Il volume raccoglie gli atti del convegno internazionale "Wikidata e la ricerca", organizzato dall'Università di Firenze e Wikimedia Italia (5-6 giugno 2025), frutto della collaborazione tra mondo accademico, volontari dei progetti Wikimedia, bibliotecari e professionisti dei dati. Esplorando le sinergie tra ricerca e Wikidata e Wikibase, i contributi coprono diverse aree disciplinari (*digital humanities*, studi classici, storico-territoriali, storico-artistici e scientifici) e affrontano temi quali il *data modeling*, le tecniche di visualizzazione e il *data retrieval*, attraverso ricerche e progetti che adottano i principi della scienza aperta e l'uso di licenze e formati aperti.

Elena Marangoni è bibliotecaria presso l'Università di Torino, fa parte del gruppo di lavoro "Authority file SBN nomi di persona" ed è una volontaria sui progetti Wikimedia dal 2014, attiva in particolare su Wikidata. È attualmente membro del consiglio direttivo di Wikimedia Italia.

Camillo Carlo Pellizzari di San Girolamo è dottorando in Scienze dell'Antichità presso la Scuola Normale Superiore; dal 2012 modifica i progetti Wikimedia da volontario ed è attivo soprattutto su Wikidata, dove si occupa in particolare dei legami con gli authority file delle biblioteche.

ISSN 2704-6478 (print)
ISSN 2704-5919 (online)
ISBN 979-12-215-1009-6 (Print)
ISBN 979-12-215-1010-2 (PDF)
ISBN 979-12-215-1011-9 (XML)
DOI 10.36253/979-12-215-1010-2

22,90 €

www.fupress.com