

The Wikibase Ecosystem in DH and GLAM

David Lindemann¹, Gustavo Candela², Alessandro Marchetti³, Camillo Carlo Pellizzari di San Girolamo⁴, Ismael Olea⁵, Christos Varvantakis⁶, Tiago Assis⁷, Vera Moitinho de Almeida⁷, Ángel Obregón⁸, Ana Salgado⁷, Alice Santiago Faria⁹, Christof Schöch¹⁰, Luis Trigo⁷

¹ EHU University of the Basque Country

² University of Alicante and Wikimedia Spain

³ Wikimedia Switzerland and Università di Pisa

⁴ Scuola Normale Superiore, Pisa

⁵ Centro Documental del Cine de Almería and LaOficina Producciones Culturales

⁶ Wikimedia Germany

⁷ Universidade do Porto

⁸ Universidad Isabel I and Wikimedia Spain

⁹ Universidade NOVA de Lisboa

¹⁰ Trier Center for Digital Humanities, University of Trier

Transformations, A DARIAH Journal

Volume 2, 2026

<https://transformations.episciences.org/>

Dates

Received: 28/12/2025

Accepted: 01/07/2026

Published: 15/09/2026

DOI: [10.46298/transformations.17200](https://doi.org/10.46298/transformations.17200)

© The authors



Creative Commons Attribution 4.0 International

Abstract

DARIAH-EU Working Group DHwiki presents this white paper as a practical introduction to Wikibase and the broader Wikibase ecosystem for researchers, GLAM (gallery, library, archive and museum) professionals, contributors to Wikimedia-related platforms and software developers working with structured and reusable data. The group's work is documented through its Wikibase instance at <https://dhwiki.wikibase.cloud>. Although there is a great deal of documentation on Wikibase and Wikidata, it is spread across different platforms, varies in scope and is written for readers with different levels of technical knowledge. This can make it difficult, especially for newcomers in Digital Humanities (DH) and GLAM contexts, to identify the most relevant concepts, tools and workflows.

To help users navigate these platforms, this article brings together the main elements of the Wikibase ecosystem with a specific focus on DH and GLAM use cases. It explains the main structure and features of Wikibase, discusses key processes such as data round-tripping and reconciliation, and examines the role of Wikidata and Wikibase in managing identifiers, authority data and controlled vocabularies. By bringing these aspects together, the paper aims to support the use of Wikibase in FAIR-compliant and collaborative data workflows, while also drawing attention to wider issues such as sustainability, governance and reuse.

Keywords: Linked Data, FAIR data, Wikibase, Wikidata

Riassunto

Il gruppo di lavoro DARIAH-EU DHwiki presenta questo white paper come introduzione pratica a Wikibase e al più ampio ecosistema Wikibase per ricercatori, professionisti del settore GLAM (Gallerie, Biblioteche, Archivi e Musei), collaboratori delle piattaforme legate a Wikimedia e sviluppatori software che lavorano con dati strutturati e riutilizzabili. Il lavoro del gruppo è documentato attraverso la propria istanza Wikibase all'indirizzo <https://dhwiki.wikibase.cloud>. Sebbene esista una grande quantità di documentazione su Wikibase e Wikidata, essa è frammentata su diverse piattaforme, varia per ampiezza di contenuti ed è scritta per lettori con diversi livelli di competenze tecniche. Tutto ciò può rendere difficile, specialmente per i neofiti nel contesto delle Digital Humanities (DH) e del settore GLAM, identificare i concetti, gli strumenti e i flussi di lavoro più pertinenti. Per far fronte a questo scenario, l'articolo affronta assieme i principali elementi dell'ecosistema Wikibase con un focus specifico sui casi d'uso in ambito DH e GLAM. Vengono illustrate la struttura principale e le funzionalità di Wikibase, discussi processi chiave come il data round-tripping e la riconciliazione dei dati, ed esaminato il ruolo di Wikidata e Wikibase nella gestione di identificativi, dati di autorità (authority data) e vocabolari controllati. Con la sua trattazione organica di tutti questi aspetti, l'articolo mira a supportare l'uso di Wikibase in flussi di lavoro di dati collaborativi e conformi ai principi FAIR, richiamando al contempo l'attenzione su questioni più ampie quali la sostenibilità, la governance e il riutilizzo.

Parole chiave: dati connessi, dati FAIR, Wikibase, Wikidata

Acknowledgements

We would like to thank Meltem Dişli, Shani Evenstein and Tomás Saorín for reviewing the text.

Funding

The work leading to the presented results has been supported by DARIAH-EU (2025-26 Theme Call funding).

Conflict of Interest

The authors do not have any conflict of interest to declare

Introduction

This article explores Wikibase,¹ the software behind Wikidata,² as a flexible and collaborative infrastructure to create, curate and connect structured knowledge bases according to Linked Open Data (LOD) principles. In an increasingly data-driven research landscape, individuals and institutions across research and cultural heritage sectors face the challenge of managing, connecting and disseminating large, complex and often interlinked datasets. In this wider context, where approaches for digital Cultural Heritage and other collections that support computational use are discussed, those approaches that follow the Linked Data paradigm are of particular interest. Deploying Semantic Web technologies (Berners-Lee, Hendler, and Lassila 2001) in Digital Humanities (DH) data collections offers the following advantages: i) resources are identified with persistent URIs, including at the most fine-grained level within the dataset; ii) resources are interoperable; and iii) data can easily be reused as LOD, avoiding data silos.

Wikibase offers a robust and flexible solution for working with such datasets. Compared to many other Linked Data environments, its distinctive contribution lies in the combination of structured semantic data with a collaborative, page-based editing interface inherited from MediaWiki.³ As data volumes grow and interdisciplinary collaboration becomes more important, FAIR-aligned workflows (Wilkinson et al. 2016) help ensure that datasets are not only technically accessible but also understandable and reusable across projects, institutions and communities (Candela et al. 2025). In this context, Wikidata and Wikibase are increasingly discussed as infrastructures that support collaborative and interoperable knowledge production (Waagmeester et al. 2019).

This paper is intended to serve as a practical and structured guide to assist new adopters who wish to understand how Wikibase can be employed. While Wikibase is broadly relevant across disciplines, this paper has a specific focus on DH and the GLAM sector, with the aim of supporting the adoption of Wikibase in diverse use cases and contributing to its further development. The main contribution of this article lies in bringing together technical, infrastructural and workflow-oriented perspectives into a single synthesis oriented toward DH and GLAM adoption. In addition, a Wikibase instance was created for DHwiki, along with a collection of reproducible Jupyter Notebooks to show how to access and edit the Wikibase instance.⁴

The remainder of this paper is structured as follows: following section , which is dedicated to related work, section introduces the main structure and features

1. <https://wikiba.se/>

2. <https://www.wikidata.org/>

3. <https://www.mediawiki.org/>

4. See the full version of this contribution at <https://dhwiki.wikibase.cloud/wiki/Documentation>

of Wikibase, with attention given to issues that are especially relevant in DH and GLAM implementations; section focuses on the concept of data round-tripping as a key workflow in the Wikibase ecosystem; and section explains the role of Wikidata as a hub for identifiers and controlled vocabularies. A closing section highlights some conclusions and an outlook towards the future activities of this Working Group.

Related Work

The publication of digital collections in the GLAM sector has transformed access to cultural heritage, enabling wider public engagement, interdisciplinary research and long-term preservation of historical materials. Data has become crucial for new AI-based applications and tools (Tasovac, Chambers, and Tóth-Czifra 2020), and new initiatives promote the publication of machine-actionable collections supporting computational use (Padilla et al. 2023; Candela et al. 2025). Data infrastructures such as the common data space for cultural heritage⁵ or the European Collaborative Cloud for Cultural Heritage (ECCCH) are in the position to influence data reuse in innovative and creative ways.⁶

Previous work employed Semantic Web technologies to make digital collections available through relevant institutions, such as GLAM, and research projects (Candela 2026; Dijkshoorn et al. 2017; Shieh 2025). External repositories such as VIAF and Wikidata were extensively employed in the DH and GLAM domains to enrich the catalogues (Candela et al. 2024; Zhao 2023).

Wikibase, by design, is particularly well suited to support FAIR principles, enabling structured data modelling with rich metadata and persistent identifiers, facilitating findability and cross-dataset interoperability, and enabling collaborative curation workflows. The use of Semantic Web standards such as the Resource Description Framework (RDF) (W3C 2014) and SPARQL endpoints (W3C 2013) aligns with accessibility and reusability goals, allowing integration into global LOD networks. Such features align with open science policies in Europe and globally, including principles like open data, transparency and ethical data governance, which are key concepts in publicly funded research and projects involving cultural heritage data.

Concerning the use of Wikibase, a preliminary search in the DBLP Computer Science Bibliography returned only 26 articles, and only a few of those focused on DH and GLAM.⁷ Some of these articles combine traditional ontology modelling with Wikibase, including manuscript metadata harmonisation and library catalogues (Koho et al. 2023; Zapounidou et al. 2024). These examples

5. <https://www.dataspace-culturalheritage.eu>

6. <https://www.echoes-eccch.eu/>

7. Results retrieved in May 2026 at <https://dblp.org/search?q=wikibase>

Feature	Description
MediaWiki extensions	Extensions can be combined within the same Wikibase instance to expand functionality.
Multilingual support	More than 600 languages and variants supported; full UTF-8 for all scripts.
Interface customisation	Customisation via PHP extensions and JavaScript gadgets.
APIs	Multiple APIs available for programmatic access and automation.
Constraint systems	Built-in QualityConstraints and Shape Expressions (ShEx) for validation and quality control.
Active development	Ongoing development by Wikimedia Deutschland and contributors.
Open Source licensing	Fully Free/Libre/Open Source; no intellectual-property lock-in risks.

Figure 1: Wikibase features

provide an overview for particular use cases of how Semantic Web and Wikibase can be integrated.

However, despite all this work, none of the previous approaches provides a comprehensive analysis of Wikibase for the DH and GLAM sectors. A detailed description of the services, tools and examples of use could facilitate Wikibase adoption by the community, and in so doing, challenges and new opportunities can be identified.

Structure and features of Wikibase for DH

Wikibase can be understood as an integrated environment that combines a linked data store, a human-readable editing interface, a query interface and a set of [Application Programming Interfaces](#) based on open standards. This combination makes it especially relevant for DH and GLAM projects that require both machine-readable structured data and workflows that remain accessible to editors, curators and domain experts. Many Wikibase features are inherited from the underlying MediaWiki framework (see [Figure 1](#)).

Documentation for Wikibase users is already available in several disparate platforms and communities; topics affecting all users, such as hosting, installation, maintenance and backup of Wikibase instances, user management, interface customisation through user scripts, etc., are covered to a certain extent. We aim to produce an overview of the available information sources as an orientation for new users as well as to assist advanced users with use-case specific requirements.

Wikibase hosting solutions

A Wikibase installation is a complex choreography of software artefacts. Installing and maintaining a Wikibase instance can be technically demanding, so several solutions have emerged to simplify deployment and hosting.

Wikibase Cloud⁸ is a software as a service (SaaS) solution, based on [WBStack](#),⁹ offered free of charge by Wikimedia Deutschland. It provides cloud-hosted

8. <https://wikibase.cloud>

9. <https://www.wbstack.com/>

Wikibase instances, including a user dashboard for the setup. Users can create an account with up to six Wikibase instances. Wikibase Cloud instances include a standard set of extensions.

Wikibase Suite¹⁰ is an open-source software Docker orchestration provided by Wikimedia Deutschland that bundles all necessary components to deploy a full-featured Wikibase environment on a user's servers. It includes MediaWiki, Wikibase, Blazegraph (the graph database),¹¹ QuickStatements¹² and other key services. It is ideal for institutions or technical users who want full control over their infrastructure (including additional extensions), their data, the instance's configuration and security backups.

Advanced users can choose their own implementation from source code repositories or Docker images. This allows them to configure the environment according to their requirements. This approach is suitable for teams with great technical expertise on systems integration, and it is valid for both standalone Wikibase and [WBStack](#) implementations.

Third-party hosting providers are an additional option. Some organisations offer Wikibase as a managed service, including setup, maintenance and support.¹³ This can be a good option for institutions lacking in-house technical capacity.

Taken together, these hosting options show that Wikibase can support very different adoption scenarios, from lightweight experimentation to institutionally managed infrastructures. This flexibility is one of the reasons why Wikibase has become attractive across a wide range of DH and GLAM contexts.

Wikibase Ontology, the backbone data model

The Wikibase Ontology is the core data model hardcoded in Wikibase, which guarantees interoperability between different Wikibase instances on a basic level.¹⁴ As described in detail in the Wikibase data model documentation,¹⁵ data about entities is structured in *statements*, in addition to a standardised way of representing basic metadata in *labels*, *aliases* and *descriptions*; in the edit interface, each of these three has its own place on the entity page. Each entity has its own identifier in the `/entity/` namespace of the Wikibase instance (e.g. `/entity/Q1`, `/entity/P1`), and one corresponding Page URL, where the entity data is displayed and can be edited by a user, e.g. `/wiki/Item:Q1`, or `/wiki/Property:P1`. The main entity types in the Wikibase are:

10. <https://wikiba.se>

11. <https://blazegraph.com/>

12. <https://quickstatements.toolforge.org/>

13. https://meta.wikimedia.org/wiki/Wikibase/Consultants_and_Support_Providers

14. <https://wikiba.se/ontology/>

15. <https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer>

- **Items:** Any concrete or abstract thing (apart from words, i.e. *lexemes*) to be described in the Wikibase, using properties, such as a subject in a semantic triple, or as a value of a property with a datatype item. Items have numeric identifiers preceded by the letter Q.
- **Properties:** Used as predicates in semantic triples. Each property has one datatype¹⁶ and a numeric identifier preceded by the letter P.
- **Lexemes:** Lexeme entities are used to describe words following Ontolex-Lemon, a W3C standard model for representing lexical data (Lindemann 2025). They have a numeric identifier preceded by the letter L.
- **EntitySchemes:**¹⁷ Shape expressions (ShEx) are used to perform data consistency checks and as a template for the form-based editing tool Cradle, etc. They have a numeric identifier preceded by the letter E.

Wikibase structured data about entities is stored as MediaWiki page data in **JSON** format; each entity stores the assertions made about it (*labels*, *descriptions* and *statements* involving other entities or data values) on its own entity data page, as described in detail in the Wikibase documentation.¹⁸ This JSON representation is the format required to upload or modify entity data using the MediaWiki API.

The Wikibase Ontology defines a number of classes and properties; some of these stem from W3C-recommended standard RDF vocabularies. Some parts of the data describing an entity are represented using these properties (e.g., `rdfs:label`), while others are represented using properties defined by the Wikibase user (numeric identifiers preceded by the letter P).

Wikibase *statements* are the core unit of structured data in Wikibase. Each statement consists of a property and a value and is made about an entity – which can be an *item*, *lexeme*, *sense*, *form*, or *property* – defined within the same instance. Each of these entity types has its own URI in the main entity namespace of the Wikibase instance. Statements include by default a mechanism for further qualification: a *rank* (*normal*, *preferred* or *deprecated*) indicating the relative priority of competing statements for the same property; *qualifiers* providing additional context for the main *claim*; and *references* documenting the source of the information. Figure 2 illustrates the RDF representation of a Wikibase statement. The blue-colored *value* nodes, depending on the property datatype, represent either entities in the same Wikibase or data values of various kinds, including strings, external identifiers, date-time objects and globe coordinates.

16. https://www.wikidata.org/wiki/Help:Data_type

17. https://www.wikidata.org/wiki/Wikidata:WikiProject_Schemas

18. https://doc.wikimedia.org/Wikibase/master/php/docs_topics_json.html

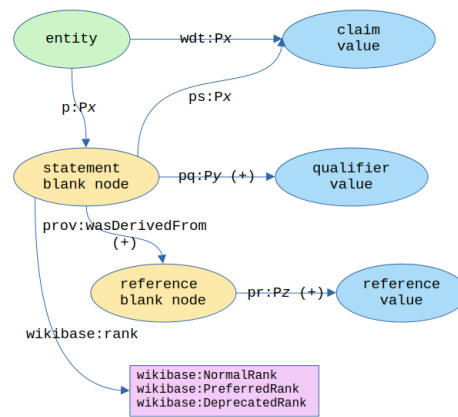


Figure 2: Wikibase Statement RDF model

SPARQL queries

The RDF representation described in Figure 2 is stored in the graph database of the Wikibase instance and can be queried using SPARQL, either through a graphical query interface, such as a query service,¹⁹ or through the SPARQL endpoint. Guidance on the graphical query service is provided on the Wikidata help pages.²⁰

Federated queries are SPARQL queries that involve more than one Wikibase instance, or other RDF databases, at the same time.²¹ Federations allow users to combine data from different source databases into a single query without physically merging those datasets. This is achieved through the SERVICE keyword in SPARQL, which instructs the query engine to retrieve data from one or more specified remote endpoints.

For DH and GLAM projects, the primary benefit of federation is to eventually overcome (redundant) data silos, creating a more interconnected and comprehensive knowledge graph. Federation enhances data analysis, entity linking and contextual enrichment, supporting scalable knowledge graph applications while retaining the autonomy of each dataset. In many use cases, it is beneficial to federate a local Wikibase instance with Wikidata. For example, details about individuals change over time, and federating with Wikidata as a central authority source avoids the need to manually synchronise updates across databases.

19. For example, the query service at DHwiki Wikibase is available at <https://dhwiki.wikibase.cloud/query/>.

20. https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service

21. https://wikidata.org/wiki/Wikidata:SPARQL_query_service/Federated_queries

Method	Purpose / Use Case	Key Characteristics
QuickStatements	Batch creation and editing of large volumes of statements	Tabular syntax; ideal for mass imports; supports create/modify/delete operations
Cradle	Guided manual item creation based on schemas	Form-based; uses EntitySchemas; ensures model consistency for human editors
OpenRefine	Reconciliation, enrichment, and upload from spreadsheets or other sources	Connects to any Wikibase; supports matching, editing, and batch upload
MediaWiki Action API	Programmatic read/write access for tools, bots, and integrations	Mature API; supports authentication, page edits, metadata queries, file uploads
Wikibase REST API	Newer structured API for entity-level operations	Simpler than Action API; available in newer distributions and Cloud instances
User scripts	Personalised interface enhancements for editors	Reusable across instances; allows custom gadgets and workflows

Figure 3: Tools for data creation and enrichment in Wikibase

The Wikibase query service offers a range of visualisation types for SPARQL query results.²² By default, variables named in the SELECT expression of the query correspond to the columns of the table containing the result, which can be downloaded in several formats. Additionally, some pre-defined visualisation options are available.²³

Programmatic access to SPARQL endpoints is demonstrated in a collection of Jupyter notebooks, using data from DHwiki Wikibase as a showcase.²⁴ In general, there is a growing range of use cases that involve the reuse of Wikidata content,²⁵ as well as content from other Wikibase instances. For example, the OpenRefine application²⁶ can be connected not only to Wikidata, but also to any Wikibase instance.²⁷ OpenRefine is widely used for entity reconciliation (matching of literal values to Wikibase entities), though it can also be used for data upload.

Data upload

Data creation and enrichment in Wikibase can take several forms, ranging from manual entry to batch upload and programmatic integration. The available tools differ in the level of technical expertise they require and in the editorial scenarios for which they are best suited (see Figure 3).

Having an account in a Wikibase instance allows several forms of interface personalisation, including interface language settings and user scripts. These scripts, which are public and reusable, can often be adapted across Wikibase

22. https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/Wikidata_Query_Help/Result_Views

23. https://dhwiki.wikibase.cloud/wiki/Structure_and_features_of_Wikibase_for_DH

24. <https://github.com/hibernator11/dhwiki-notebooks>

25. The 2025 Data Reuse Days page gives a good overview: https://www.wikidata.org/wiki/Event:Data_Reuse_Days_2025.

26. <https://openrefine.org/>

27. <https://wetneb.github.io/new-openrefine-website/docs/manual/wikibase/configuration>

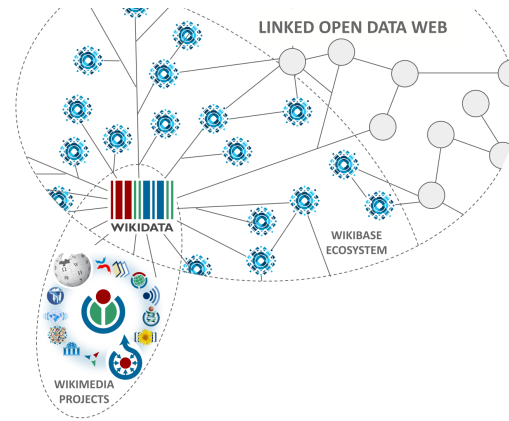


Figure 4: The Wikibase Ecosystem. Credits: CC-BY 4.0 Manuel Merz (WMDE) 2021 Wikimedia Commons

environments. Moreover, user scripts from Wikidata can be reused, sometimes with adaptations, in other Wikibase instances.²⁸

Taken together, these features show that Wikibase is more than a software package: it is an ecosystem of modelling choices, editorial practices, tools and services. Compared to more general-purpose triple store solutions, Wikibase's value lies specifically in making this ecosystem accessible to non-technical contributors while maintaining full Linked Data compliance. This broader perspective is particularly important when considering how data moves across infrastructures, a question addressed in the next section.

Data round-tripping in the Wikibase ecosystem

Wikidata²⁹ is a free and openly editable knowledge graph maintained by the Wikimedia Foundation. Although it is the most visible and widely used public Wikibase instance, it is only one part of a broader ecosystem that also includes many independent and domain-specific Wikibase installations.

The expanding ecosystem around Wikidata and Wikibase has created new possibilities for structuring, managing and exchanging knowledge across diverse domains. While Wikidata remains the central community-driven hub for LOD, volunteers and institutions are adopting their own Wikibase instances to manage specialised, sensitive or domain-specific information. Such a shift has made it essential to understand the practical, technical and governance-related differences between working directly within Wikidata and operating an independent Wikibase environment.

Data round-tripping describes the process of moving data across formats, sources or infrastructures while preserving its quality, integrity and consistency

28. https://dhwiki.wikibase.cloud/wiki/Structure_and_features_of_Wikibase_for_DH#User_scripts

29. <https://www.wikidata.org>

throughout its lifecycle. The term refers not only to technical aspects (e.g., format conversion) but also to content-related ones (e.g., preserve semantic equivalence).

Within Wikidata, the term has traditionally been used mainly for the synchronisation of data between Wikidata and external databases to which it links. In that sense, it has often referred to round-tripping with third-party databases outside the Wikimedia ecosystem proper. As of December 2025, 9,875 properties from a total of 13,088 have the datatype external identifier. Such limited integration suggests that, for many Wikidata contributors, data round-tripping is still understood mainly in relation to external databases rather than within the broader Wikibase ecosystem.

For projects operating in this ecosystem, one of the main challenges of operating a custom Wikibase instance is synchronisation with Wikidata and other data sources. Good practices can be established, but the process of data round-tripping for Wikibase users remains inherently complex and fragmented due to multiple methodological factors, including: i) avoiding entity duplication, ii) respecting the specific data models of all involved Wikibase instances and iii) preserving provenance information together with the data throughout the editorial workflow.

Tools developed to facilitate data import and alignment can support the round-tripping cycle, though in-depth knowledge of the involved identifiers, properties and community conventions is still required.

Round-tripping therefore highlights one of the central tensions of the Wikibase ecosystem: the need to balance local control, interoperability and participation in wider linked data networks. This tension becomes especially visible in the domain of identifiers, authority files and controlled vocabularies, to which we now turn.

Wikidata and Wikibase as a hub for identifiers

A central challenge in digital cultural infrastructures is that knowledge is distributed across many separate systems with limited interoperability. Identifiers play a crucial role because they enable entities to be disambiguated, linked and reused across catalogues, authority files and digital collections. Wikidata, and increasingly other Wikibase instances, have become important hubs in this landscape.

Wikidata and authority files

Authority control offers one of the clearest examples of how Wikidata functions as a linking hub. In GLAM contexts, authority data is essential for disambiguation, consistency and cross-catalogue interoperability, and Wikidata provides a shared environment in which these connections can be maintained and enriched.

GLAM institutions, most notably libraries, maintain catalogues of their collections in order to make them accessible to the public. In such environments, each resource can be found through different keys, including the title, the author(s) name, etc. However, the names of these entities are often ambiguous: an author can have multiple names (e.g., due to different transliterations) and/or homonym authors can exist. In order to ensure an effective indexing of the catalogued resources, the keywords used most frequently by users them are treated as controlled access points. The process of creating controlled access points, known as authority control, allows users to find all of and only the resources they are looking for, if properly applied. Each controlled access point (e.g., each author) has an entry, known as an authority record, containing basic data and linked to all of and only the pertinent resources. Authority records play a crucial role in disambiguating homonym entities and in collecting the multiple names of each entity, in order to ensure consistency and accuracy in indexing resources. The entire collection of the authority records curated by an institution, or group of institutions, is known as an authority file.

The use of Wikidata in authority control has been described in several presentations and case studies, most recently by (Pellizzari di San Girolamo 2025, 2026). These workflows can be grouped into two broad areas: reconciliation and cooperation. Together, they show that Wikidata is not only a passive hub for identifiers, but also an active environment for maintaining and improving authority data across systems.

Reconciliation

Authority records can always link to Wikidata, in the 017 field of UNIMARC and 024 field of MARC 21. Conversely, Wikidata can link to authority records only if they have a URI. If an authority file does not provide URIs for its authority records, it cannot be linked from Wikidata, but it can link to Wikidata.

Reconciliation between authority records and Wikidata items can be performed in three ways: manually, by importing a previous list of matches (e.g. into Wikidata through QuickStatements), or with the help of a reconciliation tool. The two main reconciliation tools used in Wikidata are Mix'n'match³⁰ and OpenRefine (see section). There are some relevant differences between them: Mix'n'match is a web tool, so it can be used by multiple users at the same time, and it allows users to create new items by putting together entries from different catalogues; conversely, OpenRefine is a software application running on an individual computer, so each reconciliation project is developed by one individual user, and it contains functions that allow users to clean messy data. Mix'n'match, in general, is more suitable for projects which are meant to be completed over a long span of time and require collaboration, and when most

30. <https://mix-n-match.toolforge.org/>

of the entries in the dataset to be reconciled with Wikidata are already present in Wikidata; conversely, OpenRefine is more suitable for projects managed by one person, and when most of the entries need to be imported as new items into Wikidata. Only datasets licensed under [CC0](#) can be massively imported into Wikidata.

Cooperation

Beyond reconciliation, sustainable cooperation between authority files and Wikidata depends on workflows for correction, comparison, enrichment and synchronisation. These workflows are especially important when authority control is understood as an ongoing collaborative process rather than a one-time alignment task.

The main areas of cooperation include:

1. editing individual authority records and Wikidata items;
2. comparing and editing both authority records and Wikidata items;
3. editing authoritative records and Wikidata items at scale.

These three areas are closely connected, and together they illustrate that authority work in LOD environments involves both local editorial control and distributed collaboration.

Regarding point 1, while Wikidata items can be edited by anyone, authority records are typically edited by cataloguers. If a Wikidata editor finds in Wikidata an error imported from an authority record, there should be some mistake-reporting procedure. Once the mistakes are corrected in Wikidata, the same should also be reflected in the authority record – the so-called data round-tripping (see section). On the other side, cataloguers can (as Wikidata editors) improve Wikidata items by linking to their authority records; this ensures that the authors they care about are well described in Wikidata, a source that can be used by other librarians to find data for their authority records and by the library itself to display more data for the catalogue’s users (e.g., information about an author, or links to related resources). Finding Wikidata items with missing and/or unreferenced data can be easily done through queries.

Regarding point 2, the correspondence between Wikidata items and authority records is usually biunivocal (1:1), although exceptions are possible, particularly in the case of pseudonyms, whose treatment differs between Wikidata and the various authority files (Chen and Yuxuan 2024; Chou 2025). Consequently, most of the cases of non-biunivocal correspondence, i.e. unique-value constraint violations (2 or more Wikidata items to 1 authority record) and single-value constraint violations (1 Wikidata item to 2 or more authority

records) suggest the presence of mistakes. These mistakes can be duplications or confluents, in Wikidata (Pellizzari di San Girolamo 2024) or in the authority file. In order to empty the lists of constraint violations, fixing the issues on one side is insufficient, because the other issues remain, cluttering the list. Only a concerted effort can produce an effective workflow, allowing improvements to both Wikidata and the authority file. If the authority file has a SPARQL endpoint, making federated queries with Wikidata allows us to easily spot inconsistencies (e.g., in gender, dates) and work on them (Kerboul 2025).

Regarding point 3, Wikidata can be used as a source to massively enrich authority records. This can be done in two ways: by retrieving the data on-the-fly through the API and showing it in an infobox, or by extracting the data periodically and copying it into the authority records (especially into empty or nearly-empty ones). Both these methods have pros and cons: the first has the advantage of being always up-to-date, whilst the second gives more control over the data shown. Examples of the first method include the AuthorityBox in the Koha ILS (Bargioni 2020) and the use of Wikidata inside the Primo ILS,³¹ whilst the main example of the second method is the enrichment of SBN authority records with data from Wikidata (Ravelli 2024, 2026). Moreover, authority records can be used to improve Wikidata through massive imports (if licensed under CC0), adding new high quality items and raising the quality of the existing ones. The main project of importing data from an authority file, currently done on a regular basis through a bot, is the one managed by the National Library of the Czech Republic (Jansová, Maixnerová, and Šťastná 2024; Dostál 2025).

Existing forms of stable cooperation between the Wikidata community and authority file initiatives, including those of the National Library of the Czech Republic, the Italian National Library Service (SBN) and the Union of Roman Ecclesiastical Libraries (URBE), suggest that this model can be sustained when technical workflows are supported by clear institutional collaboration.³²

Wikibase and authority files

Wikibase is currently being used not only as a platform for LOD publication, but also as software for the maintenance of authority files. Examples include the Semantic Name Authority Repository Cymru (SNARC) of the National Library of Wales³³ and the National Library of Nigeria Semantic Name Au-

31. [https://knowledge.exlibrisgroup.com/Primo/Product_Documentation/020Primo_VE/Primo_VE_\(English\)/150End_User_Help/Searching_Linked_Open_Data_-_Person_Entity](https://knowledge.exlibrisgroup.com/Primo/Product_Documentation/020Primo_VE/Primo_VE_(English)/150End_User_Help/Searching_Linked_Open_Data_-_Person_Entity)

32. https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche/SBN

33. <https://snarc-llgc.wikibase.cloud/>

thority Repository (Osuigwe, Boakye-Achampong, and Appiah 2025).³⁴ The National Library of Greece is testing Wikibase as the software for its authority file, following a successful proof of concept in 2024 (Gerolimos et al. 2025).

Controlled vocabularies

Wikidata can be used as a hub to interlink controlled vocabularies, both general-purpose ones and thematic ones. The main ongoing project is the Nuovo Soggettario thesaurus, developed by the National Central Library of Florence, Italy:³⁵ the two focuses of the project have been fixing the previously accumulated non-biunivocal links (through federated SPARQL queries) and improving the interconnection in a specific topic area, photography, which proved successful under many aspects (Cencetti, Pellizzari di San Girolamo, and Viti 2025; Viti and Pellizzari di San Girolamo 2025).

In 2025, the National Library of the Czech Republic exported one of its controlled vocabularies, TDKIV (the Czech Explanatory Terminology Database of the Library and Information Science), into a Wikibase instance,³⁶ with the aim of discarding the previous platform and continuing the development of the thesaurus only on Wikibase.

GLAM data

GLAM institutions have been publishing digital collections in recent decades, including a wide diversity of content such as images, paintings, postcards, metadata, text and videos. Cultural Heritage institutions have been exploring new ways to make the content available in the form of reusable data (Tasovac, Chambers, and Tóth-Czifra 2020). Recent trends such as Collections as Data and networks such as the Research Data Alliance and the International GLAM Labs Community³⁷ promote the publication of digital collections supporting computational use. Relevant examples of institutions making their digital collections available in the form of reusable data are the National Library of Scotland³⁸ and the National Library of Luxembourg.³⁹ A research project has identified Wikibase as a powerful tool for librarians and LOD (Godby et al. 2019).

Wikidata has played a leading role as a collaborative publication platform, connecting and enriching GLAM resources. Wikidata enables the creation of

34. <https://nlnsnar.wikibase.cloud/>

35. https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche/Nuovo_soggettario

36. <https://tdkiv.wikibase.cloud/>

37. <https://www.glamlabs.io/>

38. <https://data.nls.uk/>

39. <https://data.bnl.lu/>

dedicated properties to link resources such as artists, art works, places, topics, etc.⁴⁰

Rhizome ArtBase⁴¹ provides a particularly clear example of the use of Wikibase for a GLAM collection. It demonstrates how Wikibase can support the publication and management of art-related data in ways that combine semantic structure, curatorial control and public accessibility.

These examples show that Wikidata and Wikibase increasingly function as infrastructural hubs through which identifiers, authority data, vocabularies and collections can be connected. Their relevance, however, depends not only on technical capacity but also on long-term sustainability, governance and responsible reuse.

Final considerations and future prospects

This article has argued that Wikibase should be understood not only as the software behind Wikidata, but as a broader and increasingly relevant infrastructure for collaborative, interoperable and reusable knowledge organisation in DH and GLAM contexts. Its significance lies in the combination of structured semantic data, persistent identifiers, open standards and editorial workflows that remain accessible to a wide range of users and institutions.

The contribution of this paper lies in bringing together technical, infrastructural and workflow-oriented perspectives into a single synthesis oriented toward DH and GLAM adoption. Rather than introducing a new tool, it has sought to clarify how the Wikibase ecosystem can be approached in practice, what kinds of decisions it requires and what broader implications it has for interoperability, data stewardship and institutional sustainability.

This work has some limitations in terms of scope. A selection of topics, such as hosting services, data modelling and tools, is covered in order to facilitate the adoption of Wikibase by the community. The DHwiki Wikibase instance has been curated exclusively by members of the community. In addition, the Jupyter Notebooks provided are preliminary examples of how the Wikibase instance can be accessed and edited. The use of Wikibase requires extensive knowledge of several technologies, tools and services. Additional work is required in order to foster Wikibase as a platform for the DH and GLAM sectors.

Through its commitment to openness, technical interoperability and community-centred governance, Wikibase can become a foundational component within global LOD ecosystems. Emerging research data infrastructures, such as data spaces, enable new ways of publishing and accessing research data according

40. <https://qlever.dev/wikidata/IT9Tyc>

41. <https://artbase.rhizome.org>

to transparency, trustworthiness and security. Other platforms such as Galaxy Community Hub⁴² and the ECCCH will promote inclusion and democratise access to knowledge and digital assets for citizens and researchers. Wikibase can play a relevant role in these new data infrastructures not only as a data source but also as a collaborative environment to create knowledge bases.

By adopting Wikibase in their infrastructures, institutions can unlock new possibilities for research, public engagement and collaborative knowledge building, while retaining control and sovereignty over their data.

There are some aspects that need to be addressed to establish a more sustainable, ethical and impactful approach to data stewardship with Wikibase for the field of DH. For example, the provision of training materials on known platforms such as the Social Sciences and Humanities Open Marketplace⁴³ or the [DARIAH-EU Campus](#)⁴⁴ to facilitate the adoption of Wikibase as a tool for LOD editing, publication and reuse.

We conclude with more general considerations regarding sustainability and ethical issues. As for the long-term sustainability of Wikibase implementations, institutions must develop robust community engagement strategies that foster active participation and collaboration among diverse stakeholders, with accessible training programs. To assess the success of Wikibase as a knowledge infrastructure, institutions should define measurable key performance indicators aligned with their strategic goals, enabling clear impact measurement for funders and financing institutions. This aspect, which is critical for demonstrating the value of open infrastructures, remains underexplored in existing literature on Wikibase implementations. While Wikibase aligns well with the technical dimensions of FAIR principles, these should be complemented by CARE principles (collective benefit, authority to control, responsibility, and ethics) (Carroll et al. 2020). This is particularly important if Wikibase is to contribute to forms of data stewardship that are not only interoperable and reusable but also socially responsible, inclusive and attentive to the needs of Indigenous and marginalised communities.

Glossary

Application Programming Interface A set of protocols that allows different software applications to exchange data. 5

CC0 Creative Commons Zero. Public domain dedication allowing the unrestricted reuse of data. 13

DARIAH-EU Digital Research Infrastructure for the Arts and Humanities. 17

42. <https://galaxyproject.org>

43. <https://marketplace.sshopencloud.eu/>

44. <https://campus.dariah.eu/>

DH Digital Humanities. Scholarly field at the intersection of humanities research and digital technologies. 3

DHwiki DARIAH-EU Working Group DHwiki. 3

GLAM Galleries, libraries, archives and museums. Sector dealing with cultural heritage institutions. 3

ILS Integrated library system. Software used by libraries to manage catalogues and operations. 14

JSON JavaScript Object Notation. Lightweight data-interchange format used for structured data. 7

LOD Linked Open Data. Structured data published and interlinked according to web standards. 3

MARC 21 Machine-Readable Cataloguing (21st century format). Standard format for bibliographic and authority data. 12

RDF Resource Description Framework. A standard model for data interchange on the web. 4

SBN Servizio Bibliotecario Nazionale. National Library Service of Italy. 14

SPARQL SPARQL Protocol and RDF Query Language. A semantic query language for databases, used to retrieve data in RDF format. 4

URI Uniform Resource Identifier. Unique identifier used to name resources on the web. 3

VIAF Virtual International Authority File. International authority file aggregating multiple national authority records. 4

WBStack Wikibase hosting infrastructure stack. Platform supporting Wikibase Cloud deployments. 5, 6

References

- Bargioni, Stefano. 2020. "From Authority Enrichment to AuthorityBox: Applying RDA in a Koha environment." *JLIS.it* 11 (1): 175–189. <https://doi.org/10.4403/jlis.it-12595>.
- Berners-Lee, Tim, James Hendler, and Olli Lassila. 2001. "The Semantic Web in Scientific American." *Scientific American Magazine* 284 (May).
- Candela, Gustavo. 2026. "Towards a semantic approach in GLAM Labs: The case of the Data Foundry at the National Library of Scotland." *J. Inf. Sci.* 52 (1): 3–21. <https://doi.org/10.1177/01655515231174386>.
- Candela, Gustavo, Sally Chambers, Alba Irollo, Nuno Freire, Vicky Dritsou, Antoine Isaac, Agiatidis Benardou, Vicky

- Garnett, and Toma Tasovac. 2025. "A Workflow to publish Collections as Data: looking back at Europeana.eu and forward to the common European data space for cultural heritage." *Transformations: A DARIAH Journal* 1:1–18. <https://doi.org/10.46298/transformations.14774>.
- Candela, Gustavo, Mirjam Cuper, Olga Holownia, Nele Gabriëls, Milena Dobrev, 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, 15178:34–50. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72440-4_4.
- Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, et al. 2020. "The CARE Principles for Indigenous Data Governance." *Data Science Journal* 19 (November): 43. <https://doi.org/10.5334/dsj-2020-043>.
- 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.
- Chen, Chen, and Zhong Yuxuan. 2024. "A Comparative Study on the Approaches of Name Authority Control and Wikidata Identity Management." In *Knowledge Organization for Resilience in Times of Crisis: Challenges and Opportunities*, edited by Wei Lu, Thiago Henrique Bragato Barros, Lu An, and Daniel Martínez-Ávila, 63–74. Baden-Baden: Ergon. <https://doi.org/10.5771/9783987400476-63>.
- Chou, Charlene. 2025. "Linking Challenges for Pseudonyms: The Studies of VIAF, ISNI, Wikidata, Meridian, and Beyond." *Cataloging & Classification Quarterly* (July): 1–45. <https://doi.org/10.1080/01639374.2025.2502329>.
- Dijkshoorn, Chris, Lizzy Jongma, Lora Aroyo, Jacco van Ossenbruggen, Guus Schreiber, Wesley ter Weele, and Jan Wielemaker. 2017. "The Rijksmuseum collection as Linked Data." *Semantic Web* 8 (5): 707–715. <https://doi.org/10.3233/SW-170257>.
- Dostál, Vojtěch. 2025. "Czech Authority Files Meet Wikidata." Presentation. Bern, Switzerland, August.
- Gerolimos, Michalis, Effie Koufakou, Lazaros Ioannidis, Charalampos Bratsas, and Sofia Zapounidou. 2025. "MARC data on Wikibase using RDA/RDF." Presentation. Bern, Switzerland, August.
- Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Kalan Knudson Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, et al. 2019. *Creating library linked data with Wikibase: Lessons learned from Project Passage*. Technical report. August. <https://doi.org/10.25333/faq3-ax08>.
- Jansová, Linda, Lenka Maixnerová, and Petra Šťastná. 2024. "Library Data in Wikimedia Projects: Case Study from the Czech Republic." *O-bib. Das offene Bibliotheksjournal* 11, no. 4 (December): 1–19. <https://doi.org/10.5282/O-BIB/6081>.
- Kerboul, Thomas. 2025. "Enhancing authority files through SPARQL federated queries." Online presentation, November. <https://doi.org/10.5281/zenodo.17650716>.
- Koho, Mikko, L. P. Coladangelo, Lynn Ransom, and Doug Emery. 2023. "Wikibase Model for Premodern Manuscript Metadata Harmonization, Linked Data Integration, and Discovery." *ACM Journal on Computing and Cultural Heritage* 16 (3): 56:1–56:25. <https://doi.org/10.1145/3594723>.

- Lindemann, David. 2025. "Ontolex-Lemon in Wikidata and other Wikibase instances." In *Proceedings of the 5th Conference on Language, Data and Knowledge: Workshops*, 287–297. UniorPress, September. <https://doi.org/10.5281/zenodo.15471514>.
- Osuigwe, Nkem, Stanley Boakye-Achampong, and Doreen Appiah. 2025. "How to Wikibase: The Case of the NLN SNAR." Presentation. Nairobi, Kenya, August.
- Padilla, Thomas, Hannah Scates Kettler, Stewart Varner, and Yasmeen Shorish. 2023. "Vancouver Statement on Collections as Data." <https://doi.org/10.5281/zenodo.8342171>.
- 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.
- Pellizzari di San Girolamo, Camillo Carlo. 2025. "Links between Wikidata, the Thesaurus Nuovo soggettario and other thesauri." Database. <https://doi.org/10.5281/ZENODO.17235305>.
- Pellizzari di San Girolamo, Camillo Carlo. 2026. "Wikidata and authority files: reconciliation and cooperation procedures." *JLIS.it* 17, no. 2 (May): 33–50. <https://doi.org/10.36253/jlis.it-698>.
- Ravelli, Elena. 2024. "Authority file di SBN e Wikidata. Un progetto per l'arricchimento dei dati del Catalogo nazionale." Presentation. Bologna, Italy, November.
- Ravelli, Elena. 2026. "Il progetto di arricchimento dei nomi di persona di SBN con i dati di Wikidata." *JLIS.it* 17 (1): 97–105. <https://doi.org/10.36253/jlis.it-677>.
- Shieh, Jackie. 2025. "Smithsonian Libraries and Archives: BIBFRAME linked data experiment with the Share Family technology." *Art Libraries Journal* 50 (1): 12–18. <https://doi.org/10.1017/alj.2025.5>.
- Tasovac, Toma, Sally Chambers, and Erzsébet Tóth-Czifra. 2020. "Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper."
- Viti, Elisabetta, and Camillo Carlo Pellizzari di San Girolamo. 2025. "Nuovo Soggettario and Wikidata: comparison between two KOSs." Presentation. Bern, Switzerland, August.
- W3C. 2013. "SPARQL 1.1 Overview." <https://www.w3.org/TR/sparql11-overview/>.
- W3C. 2014. "RDF 1.1 Concepts and Abstract Syntax." Edited by Richard Cyganiak, David Wood, and Markus Lanthaler.
- Waagmeester, Andra, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M. Good, Malachi Griffith, Obi Griffith, Kristina Hanspers, et al. 2019. "Wikidata as a FAIR knowledge graph for the life sciences." *bioRxiv*, 799684. <https://doi.org/10.1101/799684>.
- Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3. <https://doi.org/10.1038/sdata.2016.18>.
- Zapounidou, Sofia, Lazaros Ioannidis, Michalis Gerolimos, Eftychia Koufakou, and Charalampos Bratsas. 2024. "Entity Management Using RDA and Wikibase: A Case Study at the National Library of Greece." *Journal of Library Metadata* 24 (2): 111–131. <https://doi.org/10.1080/19386389.2024.2307208>.
- Zhao, Fudie. 2023. "A systematic review of Wikidata in Digital Humanities projects." *Digital Scholarship in the Humanities* 38 (2): 852–874. <https://doi.org/10.1093/dlhc/fqac083>.