Journal of Open Humanities Data

Documenting Italian Libraries on Wikidata: From Local Projects to a Multilayered National Knowledge Graph

DISCUSSION PAPER

DAVIDE ROLLERI
ALESSANDRO MARCHETTI
CAMILLO CARLO PELLIZZARI DI SAN GIROLAMO
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The representation of Italian libraries in Wikidata has grown through two major data imports. In 2020, under commission from Tuscany region, the Sistema Cultura Toscana dataset was uploaded, raising the number of Italian libraries from fewer than 500 to 1,322 and documenting the methodology of this first large-scale project. A second step followed in 2022 with the addition of 11,239 entries from the ICCU Italian Libraries Database. This process involved merging CC0 datasets, entity alignment, and addressing gaps between the web versions of databases and their dumps. Together, these efforts illustrate both achievements and challenges in enriching Wikidata's coverage of Italian libraries, specifically highlighting the role of iterative, human-curated workflows in large-scale data imports.

CORRESPONDING AUTHOR:
Davide Rolleri
Comune di Bologna, Bologna, Italia
davide.rolleri@comune.bologna.it

KEYWORDS:
Italian Libraries Database; Wikidata import; OpenRefine

TO CITE THIS ARTICLE:
Rolleri, D., Marchetti, A., & Pellizzari di San Girolamo, C. C. (2026). Documenting Italian Libraries on Wikidata: From Local Projects to a Multilayered National Knowledge Graph. *Journal of Open Humanities Data*, 12: 23, pp. 1–15. DOI: https://doi.org/10.5334/johd.478

(1) CONTEXT AND MOTIVATIONS

By virtue of their expertise in knowledge organization, libraries are well positioned to ensure the production of high-quality information within Wikidata, thereby strengthening their institutional authority and contributing to the development of the international Linked Open Data (LOD) ecosystem.

It is therefore particularly relevant to examine how library-related metadata is integrated into Wikidata's collaborative, community-governed framework. Such integration represents a significant case study in how traditional, institutionally-based expertise intersects with a dynamic, participatory model of knowledge production in the new generation of librarians (Lucarelli, 2014; Tharani, 2021; Boccone, 2022). By engaging in this process, libraries reaffirm their more recent role as places where knowledge is actively produced, shared, and negotiated. In doing so, they contribute to the emergence of a central and dynamic node within the broader ecosystem of public data (Pomerantz & Marchionini, 2007).

The import of metadata about Italian libraries into Wikidata is a multi-layered process involving multiple stakeholders, data models, and technical methods. It reflects the long-standing collaboration between the Wikimedia community and the library community, and exemplifies the multifaceted nature of data integration into open platforms such as Wikidata.

The number of Italian libraries described in Wikidata remained low until 2020, when a major regional initiative in Tuscany triggered a significant increase. That import, and subsequent ones, highlighted the challenges and opportunities of integrating metadata of national library registries into the semantic web.

In this study, we examine two major import efforts, primarily associated with the contributions of User:Alexmar983 (Alessandro Marchetti) and User:Divudi85 (Davide Rolleri). Both contributors emerged from a regional community of Wikidata editors based in Tuscany, which also includes User:Giaccai (Susanna Giaccai), User:Manuelarosi (Manuela Musco), and User:Epìdosis (Camillo Pellizzari), all of whom are referenced in this work.

Being part of the same regional community facilitated the editorial process, enabling the co-authors to clarify key similarities and differences between their approaches – insights that may serve as useful reference for other Wikidata contributors. Furthermore, additional data import activities from other Italian users are documented and referenced throughout the study, underscoring the significance of adopting a layered methodological approach and fostering an active community to support large-scale data imports within open collaborative environments.

(1.1) THE *SISTEMA CULTURA* IMPORT

The first structured import, proposed jointly by Tuscan users and the Tuscany region in 2019, involved data from Sistema Cultura – a centralized database launched in 2008 to manage information on museums, later expanded to archives and libraries. Hosted on regional servers and managed with proprietary software for administrative and data purposes, it contained over 2,700 institution records updated via a secure, password-protected web form (Giaccai, 2021). Opening such data aligned with the Region's open access policy, aiming to foster timely and distributed data maintenance by local institutions through the use of an open data repository that is easy to update.[1]

(1.2) OTHER LOCAL IMPORTS

Following the 2020 import of Tuscan records, a distinct disparity remained compared to other Italian regions in the granularity of library data, as no other large-scale import was carried over the next two years.[2] In 2022, two projects of enrichment of library items in Wikidata were underway: one was a manual import of the Emilia Romagna libraries affiliated to the NILDE network by User:Patafisk and the librarians of the CNR Library "Dario Nobili" in Bologna,[3] and

1 See https://www.regione.toscana.it/-/open-data and https://www301.regione.toscana.it/bancadati/atti/DettaglioAttiG.xml?codprat=2013DG00000000048. All the links have been last accessed on 2025-11-14.

2 WikiToscana (2022). *Tweet*. Retrieved from https://x.com/WikiToscana/status/1479816547917082630.

3 https://it.wikipedia.org/wiki/Progetto:GLAM/NILDE.

the other was a manual import of libraries of the Marche region from the *ICCU Italian Libraries Database* by User:Tommasopaiano.

(1.3) *THE* ICCU ITALIAN LIBRARIES DATABASE *IMPORT*

The *Istituto Centrale per il Catalogo Unico, Central Institute for the Unified Catalog* (ICCU) maintains the *Anagrafe delle Biblioteche Italiane* (*Italian Libraries Database*), which, as of October 2025, contains the addresses of 19,477 libraries, 13,714 of which are described in great detail with administrative affiliation, ownership, functional type, and membership in national networks such as *Servizio Bibliotecario Nazionale* (SBN)*, National Library Service*, and *Archivio Collettivo Nazionale dei Periodici* (ACNP)*, National Collective Archive of Periodicals*.

The database is updated yearly through national initiatives of data collection coordinated by *Istituto Nazionale di Statistica* (ISTAT)*, National Institute of Statistics*, with the support of the Ministry of Culture, the Regions, and the Autonomous Provinces, but also with smaller and more frequent updates of specific subsets of libraries stemming from longstanding or newly made agreements with other Italian cultural institutions' networks.[4]

Since 2014, ICCU has made this database available as open data, enabling further research and reuse. In 2015, ICCU and Wikimedia Italia (WMIT) signed a cooperation agreement,[5] making ICCU data available under CC0 license and therefore compatible with Wikidata.[6]

Despite this open availability, the number of Italian libraries represented in Wikidata remained low, as documented in the previous paragraphs. However, after the aforementioned manual import of libraries of the Marche region from the *ICCU Italian Libraries Database* was shared with the community of librarians in Wikidata, the ICCU open data files were analyzed to evaluate the possibility of massive imports.

(2) DATASET DESCRIPTION

REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.17507984

REPOSITORY NAME

Zenodo

OBJECT NAME

Documenting Italian libraries on Wikidata: From local projects to a multilayered national knowledge graph (dataset)

FORMAT NAMES AND VERSIONS

Import Sistema Cultura (2020) ODS (Note: some files were in XLS but were converted to avoid proprietary formats)

Import ICCU (2022) XLSX, JSON

CREATION DATES

Import Sistema Cultura (2019-12–2020-02)

Import ICCU (2022-06–2022-08)

DATASET CREATORS

Alessandro Marchetti – Wikimedia Switzerland

Davide Rolleri – Comune di Bologna

4 Examples can be found in https://anagrafe.iccu.sbn.it/it/news/. As a result, the imports from ICCU to Wikidata are never to be considered complete per se, more steps are always possible.

5 ICCU and Wikimedia Italia. (2015). Accordo ICCU e Wikimedia Italia. Retrieved from https://www.iccu.sbn.it/it/eventi-novita/novita/Accordo-ICCU-e-Wikimedia-Italia/.

6 https://anagrafe.iccu.sbn.it/it/open-data/.

LANGUAGE

Italian

PUBLICATION DATE

2025-11-14

(3) METHOD

(3.1) SIMILARITIES

(3.1.1) OpenRefine

Manual entry of individual data items is time-intensive, but the process can be greatly accelerated through semi-automated workflows when relevant data already exist digitally in structured formats.

OpenRefine (formerly Google Refine) is an open-source application designed for exploring, cleaning, and transforming heterogeneous or "messy" data.[7] OpenRefine can be used easily even by novices to create facets to identify duplications and blocks of similar information and to structure, reconcile, and convert data between formats, as well as to enrich datasets through the integration of external sources and web-based services.

Both import operators used OpenRefine but, as experienced volunteers, they opposed the idea of large-scale imports without adequate data cleaning; consequently, the use of this tool was complemented by thorough pre- and post-import refinement.

(3.1.2) Import phases and data quality

Both initiatives followed a three-phase workflow.

(1) **Pre-import cleaning:** Given the constraints of Wikidata's data model and operator priorities, only part of the source information could be imported. Initial refinement was conducted in Excel, while OpenRefine—used for the first time by both operators—served primarily for the import rather than for extensive preliminary refinement.
(2) **Import phase:** Data upload was typically preceded by small-scale test batches and subsequently carried out in segmented imports to facilitate close monitoring.
(3) **Post-import refinement:** This phase addressed issues impractical or inefficient to resolve earlier.

Building on this framework, both initiatives underscored the import as the first step of continuous improvement towards a high-quality dataset. Such attention emerged from earlier bulk imports in the Italian community that had been executed with limited analysis of the underlying sources (Giaccai, 2021).[8]

(3.2) *THE* SISTEMA CULTURA *IMPORT*

(3.2.1) Preliminary analysis

The initiative proposed by the Tuscany region was aimed to make three homogeneous datasets about cultural institutions located in its territory available in the semantic web.

The import of the Sistema Cultura database[9] was the first case of an Italian regional cultural database to be imported in Wikidata therefore there were no existing examples to refer to.

7 https://openrefine.org/.

8 In January 2020, 507 Tuscan archives were imported from Archives Portal Europe (Q15427386), creating items with P17, P131, P6375 (without street numbers), P856 (SAN website), and P7764 (Archives Portal Europe ID) and labels that were often misaligned with *Sistema Cultura* names, requiring later merges. This experience reminded Tuscan users the importance of carefully designing imports for continuous community-driven improvement (Giaccai, 2021).

9 See the project page: https://www.wikidata.org/wiki/Wikidata:Sistema_Cultura.

An in-depth analysis was conducted on the records due to the complexity of handling three distinct institution types ("library networks",[10] archives, libraries) managed by separate offices. Such records were sent as attached spreadsheet files via email on November 14, 2019 while awaiting the final update of the content by the members involved in the regional system, which happened before the end of the year (the updated files were sent again in January). The data about "library networks" was presented in a single-sheet format, while in the case of libraries and archives, data were distributed in three tabs with partially overlapping columns that were therefore recombined (see Table 1).[11]

ORIGINAL FILES (DATE OF FINAL VERSION)	COMBINED NUMBER OF ROWS [EXCLUDING INDEX] (VALUES IN TABS)	COMBINED NUMBER OF COLUMNS (VALUES IN TABS)	NUMBER OF COLUMNS (AND ROWS) IN END-FILE
Library networks (2019-11-14)	16	23	15 (16 rows)
Libraries (2020-01-10)	1,116 (1,116 / 1,081 / 1,116)	57 (51 / 14 / 12)	11 (1,116 rows)
Library networks (2019-11-14)	16	23	15 (16 rows)
Archives (2020-01-10)	333 (332 / 219 / 333)	58 (53 / 15 / 11)	YEAR –11 (41 rows) NO YEAR –10 (292 rows)

Table 1 The effects of data refinement on the final files originally sent by Tuscany region via mail on November 14, 2019 (library networks, no update necessary) and on January 10, 2020 (libraries and archives, after final update from the network users). Each file corresponds to a separate import. For libraries and archives, the number of rows and columns is cumulative since the data were distributed in three different tabs (values in round brackets), with duplicated columns.

A meticulous effort was required to verify data quality and to decide which data to import in Wikidata in the corresponding properties. Some cells with descriptive paragraphs that would have been more suitable as Wikipedia content, were omitted. Sensitive (privacy-related) data and frequently-changing data (e.g. websites and contact emails) were initially excluded, as well as data that probably had already become obsolete (e.g. fax numbers). Sometimes data granularity was reduced in order to prepare for the import (e.g. structured street addresses were merged into a single string). Italian descriptions were created directly in the spreadsheet from the composition of different cells. Coordinates in the databases were also checked via random sampling and proved consistent.[12]

(3.2.2) Test and actual imports.

All imports were concentrated between December 31, 2019 and February 28, 2020.[13] The import was tested with the small file of "Library networks", while "Archives" were imported on January 28 and February 13, and "Libraries" in different batches in the last week of February.

The imports were split along different values that were missing in the original file, for example in the case of libraries a subset of items with no ISIL (P791) was separately created, including 204 items originally inserted in *Sistema Cultura* with a placeholder ISIL code (IT-XX-[X]XXX, with an additional hyphenation after the province code).

Some imports were carried out in separate steps, spaced a few hours or days apart, as reflected in the item history—particularly in the initial phase. Shorter descriptions in languages other than Italian were added during the import.

10 The correct term is "reti documentarie" (see https://www.regione.toscana.it/-/reti-documentarie-locali) which could be translated as "documentation network" and indicates a specific structure coordinating both libraries and archives. Between various options (including "GLAM network") we decided to refer to it as "library networks" here and in Zenodo.

11 The "library" spreadsheet originally included a fourth tab titled "collection", which was subsequently separated and prepared as an independent file, but never uploaded (see Section 5.2). Please notice that in May 2020 another spreadsheet about museums was imported, but this is not addressed in this article.

12 See https://www.wikidata.org/wiki/Wikidata:Sistema_Cultura/Raffinamento#Coordinate and https://www.wikidata.org/wiki/Wikidata:Sistema_Cultura/Raffinamento#Indirizzi Verification checks were primarily performed to resolve ambiguities related to coordinates of municipal libraries and archives (their locations might overlap, or indicated erroneously on the town hall) and to ensure cross-checking against the items of Tuscan archives imported in January 2020, as detailed in Note 8.

13 See https://www.wikidata.org/w/index.php?title=Special:Contributions&end=2020-03-01&namespace=0&newOnly=1&start=2019-12-30&tagfilter=&target=Alexmar983&offset=20200227154257&limit=500 and https://www.wikidata.org/w/index.php?title=Special%3AContributions&target=Alexmar983&namespace=0&tagfilter=&newOnly=1&start=2019-12-30&end=2020-03-01&limit=500.

All statements were imported with proper references including P248 (stated in). P813 (retrieved) was filled with the date when summary files were extracted and sent to the operators (see Table 1). The original files contained however the date of the first and sometimes also of the last update in the system, which were encoded as P577 (publication date) and P5017 (last update) respectively.[14]

(3.2.3) Refinement phase and trainings

The main refinement actions—for instance, manually verifying phone contacts or checking official websites when available—were catalogued.[15] These tasks were not carried out solely by wiki volunteers; rather, in cooperation with regional authorities, a sustainable model and clear framework for ongoing data refinement and curation was established.

Responsibility for maintaining up-to-date content on Wikidata – continuing the practice of the former *Sistema Cultura* databases – was assigned to network centers or to individual archives, libraries, and museums. This subsequent phase included a training program for network representatives, organized in collaboration with AIB Toscana (*Associazione Italiana Biblioteche Toscana, Italian Library Association – Tuscany chapter*) in spring 2020. The program comprised two groups of 24 and 23 participants,[16] each attending five lessons followed by two joint sessions. Each session lasted one hour and 30 minutes, totaling 18 hours of instruction. A further training session was held for three operators from the Province of Siena in October 2020.[17]

In the case of Tuscany the long-term goal was also to increase the coverage of the local "beni culturali" (cultural heritage assets) in open data ecosystems, thereby enhancing their visibility for educational and touristic purposes.[18]

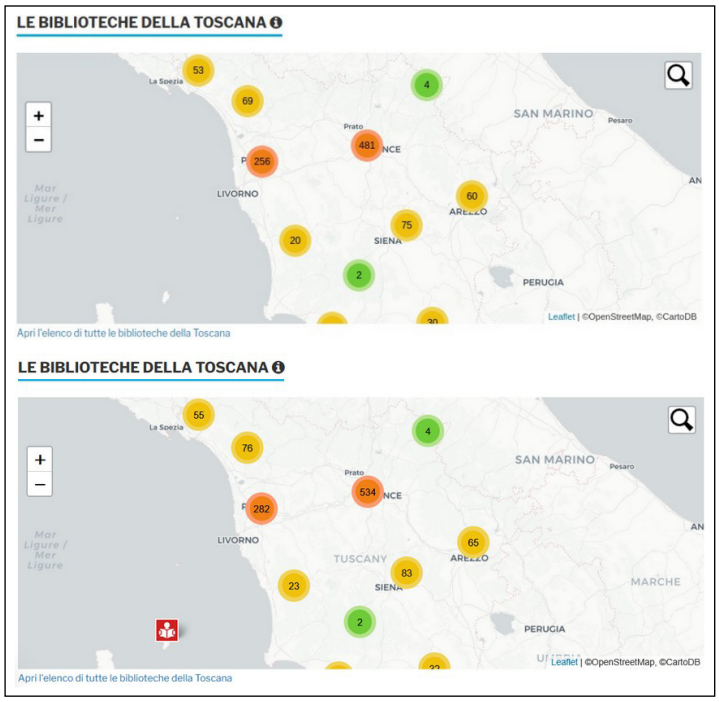

Figure 1 The results of the import on the website BiblioToscana, snapshot from https://biblio.toscana.it/. Situation as of January 2021 (above), and October 2025 (below), respectively.

The refinement phase continued thanks to the active reuse of the data and the constant monitoring by the employees of the GLAM institutions in the following years. Such reuse of data is exemplified through the integration of OpenStreetMap on the front-page map of the BiblioToscana website (Figure 1) that became public in December 2020 (Baldi et al., 2021).

14 These dates might be the same but some members updated their entries regularly, and other ones did so upon request before the import and the dismissal of the internal database.

15 https://www.wikidata.org/wiki/Wikidata:Sistema_Cultura/Raffinamento.

16 https://www.wikidata.org/wiki/Wikidata:Sistema_Cultura/Corsi.

17 https://www.wikidata.org/wiki/Wikidata:Sistema_Cultura/manutenzione_dei_dati/esercizi_REDOS.

18 Wikimedian volunteers in Tuscany are active in the topic of cultural heritage, see e.g. the metrics of the Wiki Loves Monuments competition (Giaccai, 2019) https://commons.wikimedia.org/wiki/Commons:Wiki_Loves_Monuments_in_Italy/Tuscany#Metriche and other imports, such as https://www.wikidata.org/wiki/Wikidata:Progetto_Partigiani_Toscani.

This system demonstrates a high degree of flexibility and maintains real-time connectivity with Wikidata, enabling continuous updates to be visualized over time. A comparative analysis of the map after nearly five years reveals numerous additions, many of which are directly attributable to the import of records from the *ICCU Italian Libraries Database,* as discussed in the following section.

(3.3) *THE* ICCU ITALIAN LIBRARIES DATABASE *IMPORT*

(3.3.1) Preliminary analysis

At the start of the import,[19] since the database contained at the time 18,883 records, part of them with minimal information, the first of many steps was to explore the data and create subsets to identify possible problems.

First and foremost, it is important to underline that only 67% (12,660) of the libraries in the database had the field "stato di registrazione" (registration status) blank; the others had a note in this field, most frequently "Biblioteca non censita" (non-registered library) or "Biblioteca non più esistente" (no longer existing library), implying that the data could be not up-to-date and thus discouraging their import.

Secondly, only 45.65% (8,621) of the records were updated to 2022 and another 12.65% (2,388) of the records were updated to 2021. 8.85% (1,671) had data older than five years, and 19.5% (3,685) lacked an update date – mostly the aforementioned non-registered or no longer existing libraries (Figure 2).

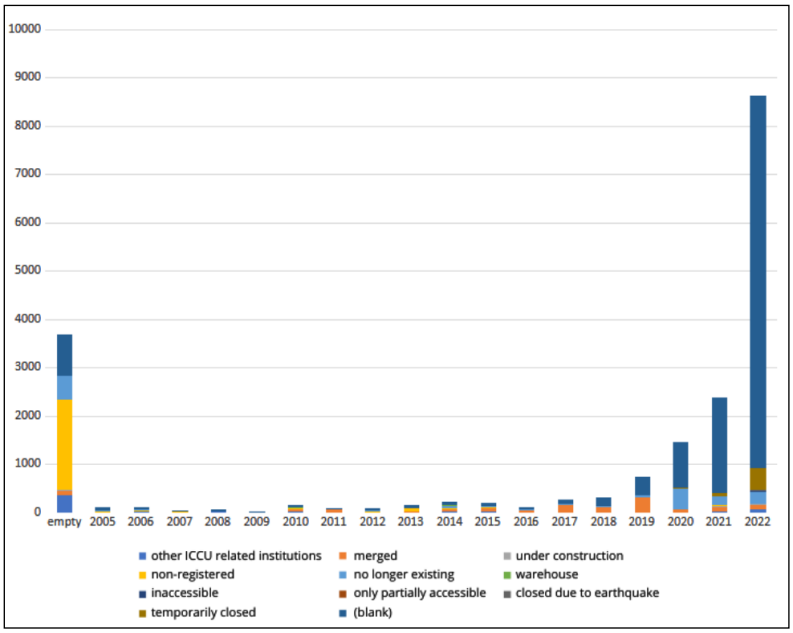

Figure 2 Distribution of libraries of the ICCU database by last update and registration status.

An in-depth analysis was conducted on the files, identifying potential problems within the JSON database. The choice was to use a less rich file, *indirizzi.csv*, containing "codice ISIL" (ISIL ID), "denominazione" (name of the library), "indirizzo" (address), "cap" (postal code), "città" (city), "provincia" (province), "regione" (region), "latitudine" (latitude) and "longitudine" (longitude) and adding data from the columns of the other files. The columns "latitudine" (latitude) and "longitudine" (longitude) were merged to be imported as P625 (coordinate location). Using the ISIL ID as the primary key and the *add column(s) from other projects* OpenRefine function, the columns "telefono" (telephone number), "email" and "url" (website url) were imported from the file *territorio.csv* while the columns "stato registrazione" (registration state), "anno censimento" (first census date), "data aggiornamento" (last update), "tipologia-amministrativa" (administrative type), "tipologia-funzionale" (functional type), "ente controllante" (parent organization), and "identificativo ACNP di una biblioteca" (ACNP library ID) were imported from the file *biblioteche.json* (See Table 2).

19 See the project page: https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche/Anagrafe_ICCU.

ORIGINAL FILES	NUMBER OF ROWS [EXCLUDING INDEX]	NUMBER OF COLUMNS	NUMBER OF COLUMNS USED IN END-FILE
indirizzi.csv	18,883	12	8 (latitude and longitude merged)
territorio.csv	12,362	17	3
biblioteche.json	18,883	34	7
complete db.csv	18,883	-	18

Table 2 Difference in number of rows and columns between the original files and the merged one used for the OpenRefine import. From the starting three files with 63 total columns, the output file had only 18. Of the 18,883 rows only 11,239 were imported, some libraries already existed in Wikidata and others didn't have enough information or the information was outdated.

(3.3.2) Test and actual imports

To better comprehend other potential problems with the file thus created, a first test import was necessary. A subset of 391 rows was created filtering "tipologia-amministrativa" (administrative type): "Comune" (municipality), "tipologia-funzionale" (functional type): "Pubblica" (public), "regione" (region): Emilia Romagna and "stato registrazione" (registration state): *blank*.

After a time-consuming effort to clean the telephone numbers, e-mail addresses and website URLs for the subset, this information was excluded in the subsequent imports due to the overall poor quality of this data. Telephone number formatting, personal emails and Facebook page URLs used instead of institutional emails and web pages were amongst the main problems with this type of data.

The same test file of 391 public libraries of the Emilia Romagna region made apparent the problem of duplicated names of libraries: 24.04% (94) of the libraries necessitated disambiguation in the denomination. On the whole database the same problem affected 22.25% (4201) of the records. As mentioned previously, the data of the ICCU database is mostly collected through annual initiatives of data collection coordinated by ISTAT; these initiatives take the form of questionnaires sent to every library, compiled by a representative of the library (librarian, functionary, delegate or volunteer depending on the library) and collected at the regional level. Since it is common in smaller cities to have a single library, most responders simply used a generic denomination like "Biblioteca comunale", "Biblioteca civica", "Biblioteca diocesana". This problem is not corrected at any stage of the data collection. Using GREL in OpenRefine the corresponding city was added to the library denomination, e.g. Biblioteca comunale → Biblioteca comunale di [Città] (Public library → Public library of [City]). The same solution was applied to libraries dedicated to the same public figure; the most frequent dedications were to Aldo Moro, Alessandro Manzoni and Cesare Pavese. In 13 cases there were actual duplicate libraries with different ISIL IDs, a report was made to ICCU that corrected the problem.

The first actual import consisted of 366 new items created and 25 existing items edited.[20]

The import process continued preparing and uploading a subset for specialized libraries in Emilia Romagna region[21] and then subsets for public libraries region by region. It was initially helpful to import no more than a couple of hundred items at a time to check the process for errors and inconsistencies. After the imports of Piemonte and Lombardia regions, respectively of 667[22] and 1,170[23] items – the biggest of these first imports concerning public libraries – it was necessary a first major correction since every item created had labels and descriptions only in Italian. The issue was solved with two edit groups adding respectively 4,779[24] labels and 4,815[25] descriptions in English. From that moment, the imports creating new items had labels and descriptions in both languages.

20 https://editgroups.toolforge.org/b/OR/abc314a0588/.

21 https://editgroups.toolforge.org/b/OR/d32e8f44734/.

22 https://editgroups.toolforge.org/b/OR/c0ffc828271/.

23 https://editgroups.toolforge.org/b/OR/ae8c7724423/.

24 https://editgroups.toolforge.org/b/OR/133184f70e0/.

25 https://editgroups.toolforge.org/b/OR/1d66a661797/.

Next in line were 1,130[26] university libraries (22 already existing and only edited) and 801[27] school libraries, then private libraries (850 created, 139 edited)[28] and finally libraries depending directly from Regions or the Italian state (811 created, 207 edited) (Figure 3).[29]

A total of 11,239 library items were created.

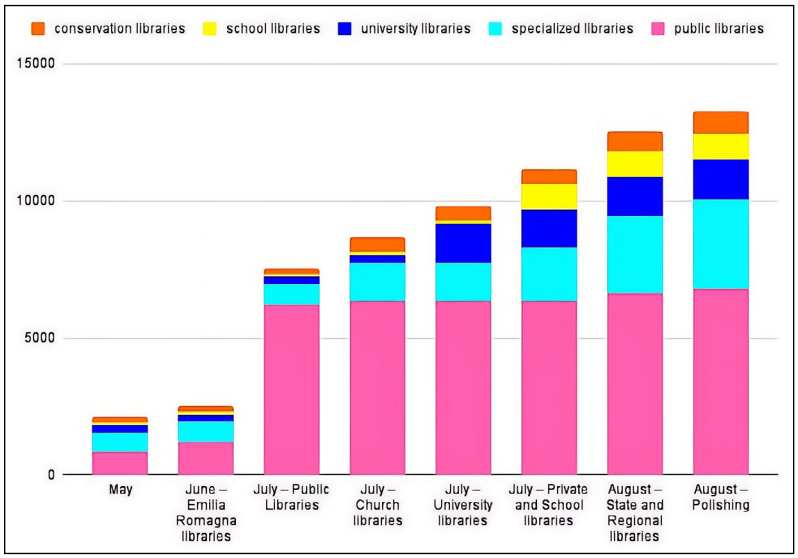

Figure 3 Increments of library items in Wikidata from May (initial situation) through August 2022 (after the refinement phase). The extended data is available in the Zenodo repository.

The statements created by the imports were referenced with P248 (stated in), P791 (ISIL), P5017 (last update) with values from the column "data aggiornamento" (last update) and P813 (retrieved) for the date of the import.

(3.3.3) Refinement phase

After a long manual refinement it was possible to import a limited set of information about library dedicatees not included in the database as a separate field but extrapolated from the denomination. 616[30] values of P138 (named after) were added to libraries, only for the most identifiable historical figures. For the libraries dedicated to minor figures, it is necessary the effort of local librarians and historians to create their Wikidata items and disambiguate them appropriately from possible homonyms.

After an equally long refinement and reconciliation of pre-existing Wikidata items, it was possible to add 905[31] values of P749 (parent organization).

In some cases the postal code was missing from the database as a standalone field: 1,088[32] values of P281 were added after an extrapolation from the corresponding complete street address.

A more specific set of values of P31 (instance of) was added in 985[33] pre-existing library items with a generic instance of Q7075 (library).

To increase interoperability and facilitate subsequent imports and data comparison, 1,182[34] values of P10667 (ACNP library ID), where available in the ICCU database, were added.

---

26 https://editgroups.toolforge.org/b/OR/12b0819e18e/, https://editgroups.toolforge.org/b/OR/6456f4f42c0/ and https://editgroups.toolforge.org/b/OR/ce2acded92a/.

27 https://editgroups.toolforge.org/b/OR/ce24b0e5cea/.

28 https://editgroups.toolforge.org/b/OR/15ae16f16b8/ and https://editgroups.toolforge.org/b/OR/242053ef30e/.

29 https://editgroups.toolforge.org/b/OR/c2bef7e1ce2/, https://editgroups.toolforge.org/b/OR/b9c933b155d/, https://editgroups.toolforge.org/b/OR/214b6ac84e9/ and https://editgroups.toolforge.org/b/OR/d4d25e1053f/.

30 https://editgroups.toolforge.org/b/OR/ec62ba534f5/.

31 https://editgroups.toolforge.org/b/OR/c87e9a527d9/.

32 https://editgroups.toolforge.org/b/OR/3e693389717/ and https://editgroups.toolforge.org/b/OR/7c602c5be54/.

33 https://editgroups.toolforge.org/b/OR/6b273e27db1/.

34 https://sigma.toolforge.org/summary.y?name=Divudi85&search=P10667&max=500&server=wikidatawiki&ns=%2C%2C&enddate=&startdate=.

The item Q113223474 (SBN hub, i.e. a subgroup of libraries participating in SBN) was created to be used as value of P31 for the 102 SBN hubs and 5,201[35] values of P463 (member of) added to the libraries which were members of a hub, using the list of libraries contained in the web page of each hub. The increasing number of statements can be seen in Figure 4.[36]

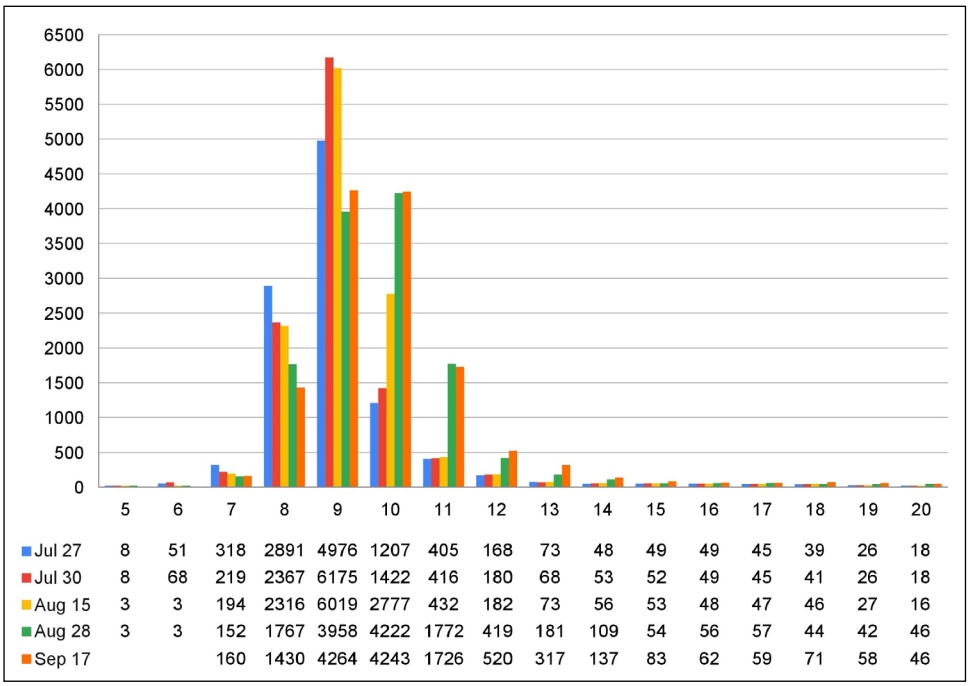

Figure 4 Number of statements increase in library items through the refinement phase. Number of statements on the x axis, number of libraries on the y axis.

(4) RESULTS AND DISCUSSION

(4.1) COPYRIGHT ASPECTS

Copyright remains a critical concern when importing data massively into Wikidata, as strict compliance with the CC0 license is essential.[37] It is not uncommon, however, for publicly available open data to be distributed under a CC BY license instead. Some of the imports discussed in this article involved datasets that were formally released under CC BY,[38] yet used with the explicit authorization of the rights holder within a project initiated by them[39] and whose content was also in part under CC0 in the ICCU database.[40] This practice, however, is not advisable and has been characterized in various public discussions as a necessary but problematic compromise.

Even in cases where the data owner explicitly authorizes or requests the import of CC BY data into Wikidata, it must be emphasized that removal requests may nonetheless arise at a later stage as it occurred with other imports.[41] In 2020, volunteers conducted a manual review to verify that the imported material represented factual information. The data were examined prior to upload; the inclusion of explicit references for each statement was accepted by the rights holder as compliant with their CC BY license; and, subsequently, the information was manually revised and, in some cases, re-entered during dedicated editing sessions. Most importantly, the multi-layered process of review and subsequent editing has further mitigated the issue, as the data have since been elaborated and validated through numerous workflows and appropriately CC0-compliant sources. Consequently, it has become *de facto* impossible to isolate and remove portions of the original import, as most statements are now supported by multiple references.

35 https://editgroups.toolforge.org/b/OR/1d35de09d8e/.

36 https://www.iccu.sbn.it/it/SBN/poli-e-biblioteche/.

37 See https://www.wikidata.org/wiki/Wikidata:Licensing.

38 The Tuscany region does not include CC0 in its opens access policy see https://esperienze.formez.it/content/open-data-regione-toscana.html and https://www.regione.toscana.it/documents/10180/339418/Open+Data+-+Strategie+RT/a26b13e9-94eb-40d8-a4ba-80b9e1ef7c58.

39 https://www.wikidata.org/wiki/Wikidata:Sistema_Cultura#Licenza.

40 https://web.archive.org/web/20190327155731/https://anagrafe.iccu.sbn.it/it/open-data/ See the description of the ICCU Database as of 2019, when the licenses were debated before the import.

41 See this comment https://it.wikipedia.org/w/index.php?title=Wikipedia%3ABar%2FDiscussioni%2FTemplatizzare_dati_statici&diff=88464090&oldid=88462158.

Nevertheless, this approach should not be regarded as a recommended practice for future imports—particularly when no subsequent development or curation of the affected items is foreseen.[42]

(4.2) DATA MODELLING AND ONTOLOGICAL ASPECTS

A major challenge concerned the classification of libraries. In 2020, Wikidata was missing many items for library types to be used as values of P31.[43]

The volunteers decided to import a generic Q7075 (library) as value of P31, but they stressed in the classes for the personnel that a possible improvement was to add more specific instances. They also modelled some descriptive items that were missing (e.g. Q86065117 "bibliocoop") and posed the issue at the related WikiProject.[44]

The effect can be seen in Figure 5, which shows how the 2022 import revisited the issue and made further progress.

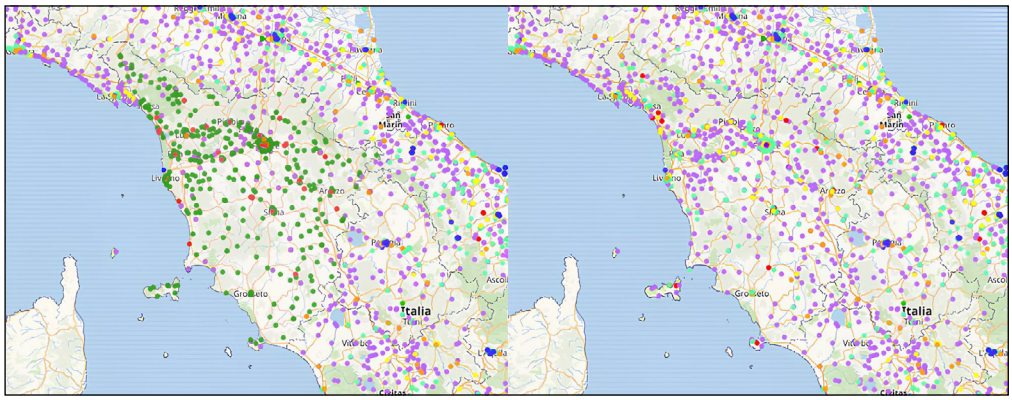

Figure 5 On the left, colored in green, the distribution of the generic value Q7075 (library) used for P31 (instance of) in most of the items related to Tuscany as of July 2022. In August 2022 the ICCU import further refined the descriptions and aligned the region with the rest of Italy using more specific values of P31, i.e. public library (in mauve), specialized library (in light blue), conservation library (in orange), university library (in blue) and school library (in yellow).

In 2022, the synchronization of the ICCU database proved challenging too, since it uses both "administrative type" and "functional type" for libraries, whilst Wikidata's ontology relies on a data model that does not fully correspond to Italian categories. For example, there is no precise equivalent for Italian "provincial" or "regional" libraries.

To improve consistency, comparisons were made with the controlled vocabulary curated by the BNCF *Biblioteca Nazionale Centrale di Firenze* (BNCF), National Central Library of Florence), the *Nuovo soggettario*,[45] but full alignment proved difficult. The discussion highlighted the need for greater participation of librarians and cultural data experts in refining Wikidata's ontology.

This means that the limited efforts made in 2020 and the analysis conducted in 2022 have left the discussion far from concluded and it will likely be further examined by the project ontology.[46]

In Table 3 the properties addressed by both imports are summarized: while core identifiers and administrative details are robust in general, other attributes (such as contact details, website URLs, and historical data on foundation or closure) are often missing or out of date. It can be complex for top-down imports to properly handle such aspects, providing structured training of dedicated users to do so (as in the case of the *Sistema Cultura* import) is a good practice but not established.

42 See https://www.wikidata.org/wiki/User_talk:Alexmar983/archive/2022#NILDE_ID. The NILDE import (also affected by a CC BY license) was in the end performed manually as well.

43 https://www.wikidata.org/wiki/Wikidata:Sistema_Cultura/Raffinamento.

44 https://www.wikidata.org/wiki/Wikidata_talk:WikiProject_Libraries#Types_of_libraries.

45 https://thes.bncf.firenze.sbn.it/.

46 A dedicated coordination subpage on Wikidata was created https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology/Controlled_vocabularies_Task_Force/Lib raries, where the issue is addressed in order to move towards a robust and informed community consensus on the ontology model including queries and authoritative external identifiers. A preliminary contact with Nuovo Soggettario (Italy's official subject indexing system) was also established.

PROPERTY	LABEL (EN)	SISTEMA CULTURA	ICCU DATABASE
P31	instance of	✔	✔
P138	named after		[✔]
P17	country	✔	✔
P131	located in the administrative territorial entity	✔	✔
P625	coordinate location	✔	✔
P463	member of	(✔)	(✔)
P749	parent organization		(✔)
P6375	street address	✔	✔
P281	postal code	✔	✔
P1329	telephone number	(✔)*	(✔)*
P968	e-mail address	(✔)*	(✔)*
P856	website url	(✔)*	(✔)*
Identifier property	Label (en)		
P791	ISIL	✔	✔
P10667	ACNP library ID		(✔)

Table 3 Summary of the properties addressed in the imports. The properties on the leftmost column are clustered in two sections, i.e. statements and external identifiers, and are ordered according to how they appear in Wikidata items as of October 2025.[47]

✓ – data in the start file; (✓) – data not present for all entries, [✓] – data inferred, * – data proven in part untrustworthy.

(4.3) OVERALL METRICS

In 2020, the upload of the *Sistema Cultura* Toscana dataset raised the number of Italian libraries from fewer than 500 to 1,322. In 2022, before the ICCU database import, the coverage was still fragmented, but with the addition of 11,239 entries the national coverage became homogenous. The effect of such a multi-layered approach is illustrated in Figure 6.

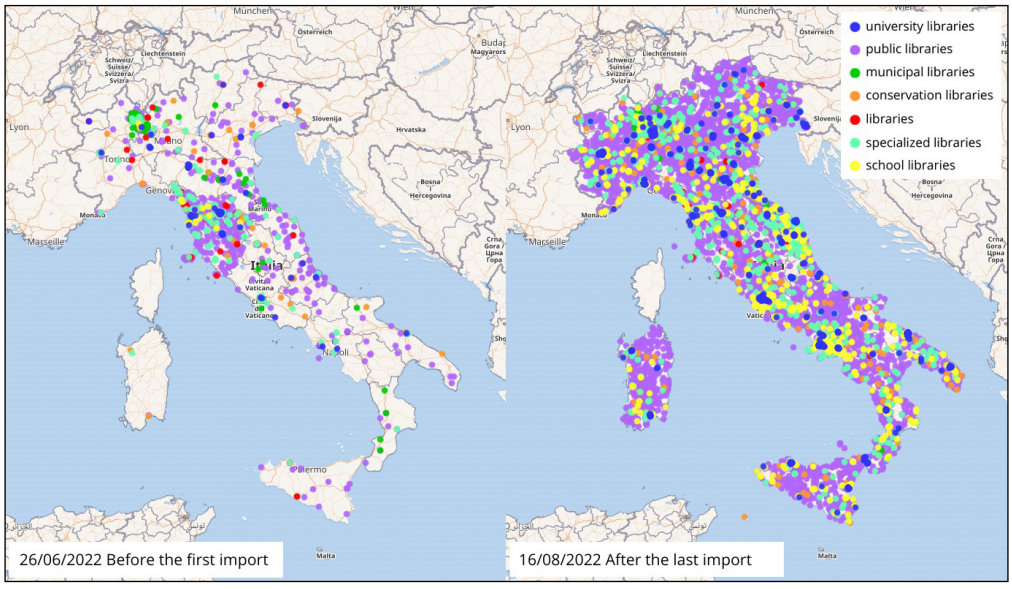

Figure 6 The effects of different imports of Wikidata items related to Italian libraries on the Italian map as of 2022, adapted from (Rolleri, 2022).

(5) IMPLICATIONS/APPLICATIONS

National-scale integrations of library metadata into Wikidata have been explored in other contexts as well (Obregón Sierra, 2022), but the integration of Italian libraries data into Wikidata offers valuable insights for digital humanities, library science, and Linked Open Data, highlighting key considerations for future contributors undertaking similar initiatives.

(5.1) ESTABLISHED IMPLICATIONS

Heterogeneous regional, national, and institutional registries can be integrated into a global open knowledge graph: even when imported in distinct phases—including updated versions of the same database—such datasets interconnect effectively thanks to Wikidata's flexible data

47 https://www.wikidata.org/wiki/MediaWiki:Wikibase-SortedProperties.

model, user-friendly interface, and active community. The comparison of the two initiatives produced useful guidelines for future data imports.

- For library management: libraries can take ownership of their digital identity by maintaining and updating their own items in Wikidata, thereby improving visibility and integrating their presence into the wider LOD environment.
- For import organization: as with other comparable large-scale data imports, both initiatives converged toward a structure composed of three phases (see Paragraph 3.1.2)- (1) data analysis, including modelling and identification of gaps; (2) data import, often preceded by testing and divided into smaller batches to allow monitoring; and (3) data refinement, based on the principle that an import is the beginning of continuous improvement, not its conclusion.
- For import modelling: successful imports require careful design of the data model and selection of metadata. The final versions of the spreadsheets uploaded on Zenodo show the work needed to prepare data and the need of reducing the granularity of information. Accurate selection of specific instances remains a key prerequisite. Gaps in the ontology should be explicitly identified and addressed (see next sections).
- For data quality: bottom-up approaches and peer-to-peer collaboration are important in refining data from centralized institutions, where different portions or aspects may originally adhere to inconsistent quality thresholds.
- For process governance: the use of standardized edit summaries, and updated project pages to share documentation practices and monitoring tools provides a framework for sustainable and accountable data import activities.

(5.2) SUBSEQUENT INITIATIVES AND DEVELOPMENTS

The Tuscany project was designed as a pilot model for integrating regional library datasets. It also outlined possible future development steps (Baldi et al., 2021), as well as the 2022 import did (Rolleri, 2024).

- The projects exposed limitations in the data model currently used in Wikidata for representing libraries and highlights the need for expert involvement to improve semantic precision. The topic of controlled vocabularies is gaining attention across different areas of research and, within the Wikidata community, a more focused effort on the quality of the ontology has been emerging (Zhang et al., 2015; Pastor-Sánchez, 2021; Bagov et al., 2022) and led to a more focused effort on the quality of the ontology on Wikidata (see the improvement of the dedicated project since 2023).[48]
- The import projects excluded data about the accessibility and opening hours of libraries, to be uploaded to OpenStreetMap in a second phase. Later, the 2024 import from WMIT[49] imported such metadata with P2848 (Wi-Fi access), P3025 (opening days) and P8026 (opening times); the main challenge is now keeping these data up-to-date in Wikidata.
- Structured, open data on thousands of libraries make it possible to study the historical distribution of libraries, their dedications (e.g., to notable figures by gender or profession), and the evolution of library networks across regions (Rolleri, 2022).
- The import of cultural institutions can be linked to the greater visibility of cultural heritage as open data on the web, thus contributing to the valorisation of lesser-known cultural resources. Wikidata helps libraries and archives achieve both sustainability and local engagement, while also offering additional possibilities such as contributing entries about their collections (Okuonghae, 2024). There was a debate between the volunteers and the Region to also import metadata about book collections described in *Sistema Cultura*, but it was rejected because at the time these items were not fully modelled. A project was started in 2025 to create a data model for book collections.[50] Also the WMIT 2024 import started to insert P1436 (“collection or exhibition size”), showing interest in a more complete modelling of such aspects.

48 See https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology and, since 2023, https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology/Cleaning_Task_Force.

49 https://www.wikidata.org/wiki/Wikidata:WikiProject_Italy/Empowering_Italian_GLAM.

50 https://www.wikidata.org/wiki/Wikidata:WikiProject_Personal_Collections.

(5.3) CONCLUSIONS

This study documents a multi-layered effort to integrate Italian library metadata into Wikidata, demonstrating a pathway to transform heterogeneous institutional data into a coherent knowledge graph through careful modeling and community coordination, deliberately relying on manual quality control. This choice was methodological rather than incidental: human judgment is necessary to handle data heterogeneity — a topic that has been examined across different domains (Kumar et al., 2021; Hansen & Quinon, 2023).

The three-phase workflow (pre-import cleaning, import, post-import refinement) is applicable to any large-scale and medium-scale Wikidata import project (Bianchini et al., 2023; De Monaco, 2025), the core element is the human factor. While large, top-down imports play a significant role, this work shows that coordinated bottom-up, small-scale local contributions and active cooperation among established contributors can collectively enhance the global knowledge graph, making Wikidata a living, collaboratively maintained catalogue of cultural institutions. This vision fits within the emerging philosophy of "Slow Data" approach that was described in 2024,[51] but was progressively built up locally in Tuscany, where various training initiatives were established over the years[52] creating a community of users with different backgrounds.

ACKNOWLEDGEMENTS

For the 2020 import we acknowledge the Directorate of Culture and Research, Tuscany region (Francesca Navarria and Giancarla Brusoni), Associazione Italiana Biblioteche Tuscany and Wikidata User:Giaccai and User:Manuelarosi.

For the 2022 imports we acknowledge Wikidata User:Patafisik and the librarians of the CNR Library "Dario Nobili" for the work on the NILDE Network, and Wikidata User:Tommasopaiano for the manual import of the Marche region libraries.

For continuing the work of enrichment of libraries in Wikidata we acknowledge Wikidata User:Adert (WMIT).

FUNDING STATEMENT

The Tuscany imports were paid by AIB Toscana (2020); the ICCU imports were not funded.

COMPETING INTERESTS

One of the authors of the paper is an editor of the special issue Wikidata Across the Humanities: Datasets, Methodologies, Reuse, to which the paper was submitted.

AUTHOR CONTRIBUTIONS

Davide Rolleri – Data curation, Conceptualization, Writing – original draft, Writing – review & editing

Alessandro Marchetti – Data curation, Conceptualization, Writing – original draft, Writing – review & editing

Camillo Carlo Pellizzari di San Girolamo – Writing – review & editing

AUTHOR AFFILIATIONS

Davide Rolleri orcid.org/0009-0001-9791-1729
Comune di Bologna, Bologna, Italia

Alessandro Marchetti orcid.org/0000-0002-6125-3920
Wikimedia Switzerland, 8008 Zürich, Switzerland; Dipartimento di Biologia, Università di Pisa, Pisa, Italy

Camillo Carlo Pellizzari di San Girolamo orcid.org/0000-0003-2699-1693
Classe di Lettere e Filosofia, Scuola Normale Superiore, Pisa, Italy

51 See https://commonists.wordpress.com/2024/10/09/small-data-slow-data-a-snail-approach-to-wikidata/.

52 See https://meta.wikimedia.org/wiki/Gruppi_locali_di_lingua_italiana/Toscana.

REFERENCES

Bagov I., Greiner C., & **Garabedian N.** (2022). Collaborative Metadata Definition using Controlled Vocabularies, and Ontologies. *Research Ideas and Outcomes, 8*, e94931. https://doi.org/10.3897/rio.8.e94931

Baldi, P., Navarria, F., Becchi, G., Ferracani, A., & **Marchetti, A.** (2021, October 7) Presentazione del Portale di ricerca bibliografica BiblioToscana della Regione Toscana [Presentation]. Internet Festival 2021, Pisa, Italy. https://2021.internetfestival.it/programma/presentazione-del-portale-di-ricerca-bibliografica-bibliotoscana-della-regione-toscana/.

Bianchini, C., Marchitelli, A., & **Moi, A.** (2023). Metodi e strumenti di un progetto di valorizzazione delle riviste italiane di biblioteconomia in Wikidata, *AIB Studi, 63*(2), 313–335. https://doi.org/10.2426/aibstudi-13893

Boccone, A. (2022). Il ruolo del Wikidata librarian in un rinnovato universo bibliografico: "next generation metadata", next generation librarians. *JLIS.it, 13*(2), 45–57. https://doi.org/10.36253/jlis.it-460

De Monaco, S. (2025). Il "Dizionario degli scrittori italiani contemporanei pseudonimi" in Wikidata. Metodologia e risultati [MA thesis]. Università di Pisa, Pisa, Italy. URL: https://etd.adm.unipi.it/theses/available/etd-05132025-160428/

Giaccai, S. (2019). Wiki loves monuments: il concorso fotografico del mondo wiki. *Bibelot: notizie dalle biblioteche toscane, 25*(3), 1–7. URL: https://riviste.aib.it/bibelot/article/view/12011

Giaccai, S. (2021). Archivi, biblioteche e musei toscani in Wikidata. *Bibelot: notizie dalle biblioteche toscane, 27*(1), 1–7. URL: https://riviste.aib.it/bibelot/article/view/13166

Hansen, J. U., & **Quinon, P.** (2023). The importance of expert knowledge in big data and machine learning. *Synthese, 201*. https://doi.org/10.1007/s11229-023-04041-5

Kumar G. M., Basri, S., Imam, A. A., Khowaja, S. A., Capretz, L. F., & **Balogun, A.** (2021). Data Harmonization for Heterogeneous Datasets: A Systematic Literature Review. *Applied Sciences, 11*(17), 8275. https://doi.org/10.3390/app11178275

Lucarelli, A. (2014). "Wikipedia loves libraries": in Italia è un amore corrisposto…. *AIB Studi, 54*(2/3), 241–259. https://doi.org/10.2426/aibstudi-10108

Obregón Sierra, Á. (2022). Inserción de metadatos de las bibliotecas españolas en Wikidata: un modelo de datos abiertos enlazados. *Revista Española De Documentación Científica, 45*(3). https://doi.org/10.3989/redc.2022.3.1870

Okuonghae, O. (2024). Beyond the Library Catalogue: Connecting Library Metadata to Wikidata. *Folia Toruniensia*, *24*, 183–192. https://doi.org/10.12775/ft.2024.009

Pastor-Sánchez, J. (2021). Wikidata como herramienta para elaborar ontologías y vocabularios controlados. *Anuario ThinkEPI, 15*. https://doi.org/10.3145/thinkepi.2021.e15f01

Pomerantz, J., & **Marchionini, G.** (2007). The digital library as place. *Journal of Documentation, 63*(4), 505–533. https://doi.org/10.1108/00220410710758995

Rolleri, D. (2022). L'Anagrafe delle Biblioteche italiane in Wikidata. *Bibelot: notizie dalle biblioteche toscane, 28*(3), 1–7. URL: https://riviste.aib.it/bibelot/article/view/13809

Rolleri, D. (2024, November 9). Importazione dell'Anagrafe delle Biblioteche Italiane in Wikidata [Presentation]. Wikidata Days, Bologna, Italy. Retrieved from https://commons.wikimedia.org/entity/M154805958

Tharani, K. (2021). Much more than a mere technology: A systematic review of Wikidata in libraries. *The Journal of Academic Librarianship, 47*(2). https://doi.org/10.1016/j.acalib.2021.102326

Zhang, Y., Ogletree, A., Greenberg, J., & **Rowell, C.** (2015). Controlled vocabularies for scientific data: Users and desired functionalities. *Proceedings of the Association for Information Science and Technology, 52*(1). https://doi.org/10.1002/pra2.2015.145052010054

TO CITE THIS ARTICLE:
Rolleri, D., Marchetti, A., & Pellizzari di San Girolamo, C. C. (2026). Documenting Italian Libraries on Wikidata: From Local Projects to a Multilayered National Knowledge Graph. *Journal of Open Humanities Data*, 12: 23, pp. 1–15. DOI: https://doi.org/10.5334/johd.478

Submitted: 14 November 2025
Accepted: 18 December 2025
Published: 04 February 2026

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.