The Wikibase Ecosystem in DH and GLAM

David Lindemann[1] Gustavo Candela[2]
Alessandro Marchetti[3]
Camillo Carlo Pellizzari di San Girolamo[4] Ismael Olea[5]
Christos Varvantakis[6] Tiago Assis[7]

Vera Moitinho de Almeida[7] Ángel Obregón[8]
Alice Santiago Faria[9] Christof Schöch[10] Luis Trigo[7]

[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]Wikimedia Spain 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

Abstract in English

DARIAH-EU Working Group DHwiki presents this White Paper about Wikibase and the Wikibase Ecosystem as a digest of its permanent activity in providing documentation about MediaWiki-based software solutions to researchers. The output of the Working group is accessible on its Wikibase instance at https://dhwiki.wikibase.cloud. As diverse as the Working Group itself is the target group of this report: DH researchers, GLAM institutions, contributors to the software platforms provided by Wikimedia, and software developers. When starting to engage with Wikidata and Wikibase, we have seen ourselves sometimes lost in a diverse landscape of disperse documentation resources with different scope and aimed at differently skilled readers. With this contribution, we hope to provide a comfortable pathway towards adopting Wikibase in FAIR Linked Data publishing and collaborative data curation workflows.

Keywords in English: Linked Data, FAIR data, Wikibase, Wikidata
Received: XX/XX/XXXX
Accepted: XX/XX/XXXX

Acknowledgements

The work leading to the presented results has been supported by DARIAH-EU (2025-26 Theme Call funding). We would like to thank all the members of the DHwiki group who reviewed a preliminary version of the chapters including Meltem Dişli, Shani Evenstein, Ana Salgado, and Tomás Saorín. We would like to thank Wikimedia Germany for hosting our first in-person meeting in Berlin in June 2025.

1 Introduction

This article explores Wikibase, the software behind Wikidata, as a powerful and versatile tool for creating structured knowledge bases following the principles of Linked Open Data (LOD).

In an increasingly data-driven research landscape, individuals and institutions across various sectors face the challenge of managing, connecting, and disseminating large, complex and often interlinked datasets. In a 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 a particular interest. Deploying Semantic Web technologies[1] in Digital Humanities (DH) data collections becomes interesting for a range of reasons:

- Resources are identified with permanent URI, also inside the dataset, on the most fine-grained level;
- Interoperability with other resources, e.g., for data enrichment or dataset integration;
- Straightforward reusability as LOD (instead of closed data siloes);

Wikibase offers a robust and flexible solution for working with such datasets, but Wikibase's most distinctive features compared to other solutions for working with LOD are (1) the fact that every entity described in the database will have its own editable web page listing the data statements associated to it, and (2), the advanced support for collaborative editing and reviewing Wikibase inherits from MediaWiki.

The FAIR data principles (Findable, Accessible, Interoperable, and Reusable) are increasingly recognized as essential for sustainable and transparent research infrastructures. As data volumes grow and interdisciplinary collaboration becomes essential, adopting FAIR-aligned workflows ensures that datasets are not only technically accessible but also understandable and reusable across different contexts, projects, and communities (Candela 2023). In this context, Wikidata and Wikibase are evolving as infrastructure solutions, especially in collaborative workflows (Waagmeester et al. 2019).

Wikibase, by design, is uniquely positioned to support FAIR principles, enabling structured data modelling with rich metadata and persistent identifiers, supporting findability and cross-dataset interoperability, and combining all this with the features necessary for collaborative work. The use of Linked Data standards such as the Resource Description Framework (RDF) and SPARQL endpoints aligns with accessibility and reusability goals, allowing integration into global Linked Open Data networks. Such features resonate with open science policies in Europe and globally, in alignment

1. See the https://5stardata.info/en/ Five Star Open Data scheme for a general introduction.

with concepts like open data, transparency, and ethical data governance, which are key concepts in publicly funded research and projects involving cultural heritage data.

The paper is intended to serve as a practical guide to assist new adopters in implementing their projects on Wikibase effectively. While Wikibase is broadly relevant across disciplines, this paper has specific focus to the DH, and to the Galleries, Libraries, Archives, and Museums (GLAM) sector, hoping to ease the adoption of the discussed software in diverse use cases, and the development of the software itself.

The remainder of this paper is structured as follows: Section 2 introduces the Wikibase platform; section 3 focuses on the concept of data round-tripping in the Wikibase Ecosystem; section 4 explains the role of Wikidata as linking hub between diverse datasets. A closing section highlights some conclusions and an outlook towards feature activities of this Working Group.

A full and regularly updated version of the documentation presented here is available at DHwiki Wikibase.[2] In addition to the content summarized in this paper, the full version contains the following dedicated sections:

- Examples of DH-related Wikibase instances;
- Lexical data in Wikibase;
- Requirements for further development emerging in DH and GLAM use cases;
- A resources directory and a bibliography.

2 Structure and features of Wikibase for DH

Wikibase can be understood as a product including a linked data store, an end user editing interface, an end user interface for querying, and a set of APIs based on open standards.

This combination constitutes a versatile product with a relatively low adoption curve. Many Wikibase features are inherited from the underlying MediaWiki framework, like the following:

- Further MediaWiki extensions may be combined in the same instance;
- Multilinguality: support for more than 600 languages and variants;
- Interface customisation with PHP extensions and Javascript accessories;
- a range of APIs;
- Implementation of two constraint systems: QualityConstraints[3] and Shape Expressions.[4]
- The software is in active development by Wikimedia Deutschland and others.
- Free/Libre/Open Source software without intellectual property rights block risks.

Documentation for Wikibase users is already available in several disparate places; topics affecting all users such as hosting, installation, maintenance and backup of Wikibase instances, user management, interface customization through user scripts, etc., are covered to a certain extent. We aim to produce an overview of the available information sources[5] as orientation to new users as well as to assist advanced users with use-case specific requirements.

2. See https://dhwiki.wikibase.cloud/wiki/Documentation.
3. See https://www.mediawiki.org/wiki/Extension:WikibaseQualityConstraints).
4. See https://www.mediawiki.org/wiki/Extension:EntitySchema.
5. See https://dhwiki.wikibase.cloud/wiki/Resource_directory_and_Bibliography.

2.1 Wikibase hosting solutions

In any case, a Wikibase install is a complex choreography of software artifacts. Installing and maintaining a Wikibase instance can be technically demanding, so several solutions have emerged to simplify deployment and hosting:

- Wikibase Cloud[6] is a software as a service, based on WBStack,[7] offered free of charge by Wikimedia Deutschland[8]. It provides cloud-hosted Wikibase instances, providing a user dashboard for the setup. Everybody can have a user account with up to six Wikibase instances. Wikibase Cloud instances include a standard set of extensions.[9]
- Wikibase Suite[10] is an open-source software Docker orchestration provided by Wikimedia Deutschland that bundles all necessary components to deploy a full-featured Wikibase environment on own servers. It includes MediaWiki, Wikibase, Blazegraph (the graph database), QuickStatements, 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.
- Self-hosted solutions: Very advanced users can choose for their own implementation from source code repositories or Docker images. This allows them to configure any detail 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: Some organizations[11] offer Wikibase as a managed service, including setup, maintenance, and support. This can be a good option for institutions lacking in-house technical capacity.

2.2 Wikibase Ontology, the backbone data model

The Wikibase Ontology[12] is the fundamental data model hardcoded in Wikibase, which guarantees interoperability between different Wikibase instances on a basic level. As explained in detail in the Wikibase data model documentation[13], data about entities is structured in the layers *labels*, *aliases*, *descriptions*, and *statements*. In the edit interface, each of the 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 editable by a user, e.g. `/wiki/Item:Q1`, or `/wiki/Property:P1`.

The types of entities modelled in a Wikibase are:

- **Items**: Any concrete or abstract thing (apart from words, i.e. *lexemes*) to be described on the Wikibase, using properties, as subject in a semantic triple,

6. See https://wikibase.cloud.
7. See https://www.wbstack.com/.
8. See https://wikimedia.de.
9. For this instance, DHwiki Wikibase, you can access detailed information at https://dhwiki.wikibase.cloud/wiki/Special:Version#mw-version-ext.
10. See https://wikiba.se.
11. See https://meta.wikimedia.org/wiki/Wikibase/Consultants_and_Support_Providers.
12. The Wikibase Ontology in OWL is available as TTL file at https://wikiba.se/ontology/.
13. See https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer.

or as value of a property with data type item. Items have numeral identifiers preceded by letter Q.

- **Properties**: to be used as predicates in semantic triples. Each property has one datatype,[14] and a numeral identifier preceded by letter P.
- **Lexemes**: Lexeme entities are used to describe words following Ontolex-Lemon. They have a numeral identifier preceded by letter L.
- **EntitySchemes**:[15] Shape expressions used for performing data consistency checks, and as template for the form-based editing tool Cradle, etc. They have a numeral identifier preceded by letter E.

Wikibase structured data about entities is stored as MediaWiki page data in JSON format; each entity has the assertions that are made about it (*labels*, *descriptions*, and *statements* involving other entities or data values) stored on its own entity data page, as described in detail in the Wikibase documentation.[16] This JSON representation is the format required for upload or modification of entity data using the MediaWiki API.

In parallel to the JSON data, which is stored in a relational database, changes to entity data are propagated to a separate graph database, where the RDF representation of the entity data is stored. The RDF data can be accessed as entity data (i.e., the equivalent to a JSON entity page) in Turtle format,[17] and as RDF graph through the graphical SPARQL query service, or through the SPARQL endpoint.

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

Table 1 lists those properties that are part of the Wikibase Ontology and used by default in the RDF representation of entities as equivalent to the keys in the JSON representation. For a detailed description of Wikibase lexemes, which are modelled according to OntoLex-Lemon (McCrae et al. 2017), see a dedicated overview paper (Lindemann 2025).

Wikibase *statements* are the part of the Wikibase entity data that involves properties defined on the same instance. Statements include by default a mechanism for further describing the main *claim* of a statement using *qualifiers*, *ranks*, and *references*. A graphical model of a Wikibase statement's RDF representation is given in the figure, where *entity* represents an *item*, *lexeme*, *sense*, *form*, or *property* node, each of which has its own URI in the main entity namespace of the Wikibase instance, and where the blue-colored *value* nodes, depending on the property datatype, represent entities of the same Wikibase, or data values, including strings, external identifiers, date-time objects, globe coordinates, et cetera. Each statement node may be linked to several *qualifier* values, and to several *reference* nodes (in the RDF representation attached to the statement blank node using `prov:wasDerivedFrom`), which make up blocks of

14. See the list of datatypes on the Wikidata help pages at https://www.wikidata.org/wiki/Help:Data_type.

15. See https://www.wikidata.org/wiki/Wikidata:WikiProject_Schemas and https://www.weso.es/WShEx/.

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

17. See an example for an item at https://wikidata.org/wiki/Special:EntityData/42.ttl (Douglas Adams), and for a lexeme at https://wikidata.org/wiki/Special:EntityData/L1347.ttl

Table 1: RDF properties used in Wikibase for basic entity description (Q=item, P=property, E=EntityScheme)

JSON key	RDF property	Notes
labels	`rdfs:label`	Q,P,E preferred label, one per language code
descriptions	`schema:description`	Q,P,E short description, one per language code
aliases	`skos:altLabel`	Q,P,E alternative labels, multiple per language code
sitelinks	`schema:about`	(Wikidata only: Concept links from other Wikimedia projects)
lemmas	`wikibase:lemma`	Lexeme lemma, one per language code
lexicalCategory	`wikibase:lexicalCategory`	Lexeme lexical category
language	`dct:language`	Lexeme language, one per lexeme
senses	`ontolex:sense`	Lexeme sense
glosses	`skos:definition`	Lexeme sense definition
forms	`ontolex:lexicalForm`	Lexeme form
representations	`ontolex:representation`	Lexeme form representation, one per language code
grammaticalFeatures	`wikibase:grammaticalFeatures`	Lexeme form grammatical feature

references. *Ranks* are used for annotating multiple statements for the same property with one of the three values *normal*, *preferred*, or *deprecated*.

Wikibase implements standard property **datatypes**, some of them are available default in the Wikibase extension, and others trough additional extensions.[18] Each Wikibase property is associated to one datatype; when uploading data using the MediaWiki Action API, the value is validated according to the datatype (for example, only a valid URL will be accepted as value for a property with datatype URL).

2.3 SPARQL Queries

The above described RDF representation is stored in the Wikibase instance's graph database and can be queried using the SPARQL query language, either through a graphical interface, the Query Service,[19] or, programmatically, through the SPARQL endpoint. How to use the graphical Query Service is explained on the Wikidata Help pages.[20] These explanations hold through for any Wikibase, with the exception of pre-configured namespace prefixes, which may differ.

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

18. For a complete description of datatypes, see https://dhwiki.wikibase.cloud/wiki/Structure_and_features_of_Wikibase_for_DH#Property_data_types.

19. For example, the query service at DHwiki Wikibase is available at https://dhwiki.wikibase.cloud/query/. Wikidata's query service is available at https://query.wikidata.org/; for Wikidata, the University of Freiburg provides, as an alternative, https://qlever.dev/wikidata.

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

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

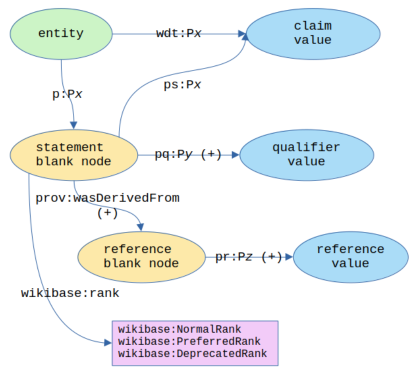

Figure 1: Wikibase Statement RDF model

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, a federation of an own Wikibase with Wikidata will make sense, for example for person metadata: Details about persons change over time, and synchronizing updates between databases can be avoided if federating with a central database, which for persons would be Wikidata.

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 (JSON, CSV, HTML). In addition, some pre-defined visualisation options are available.[23] Each visualization type requires specific variable names or data structures in the query results to render properly (for example, maps typically need coordinates, timelines need date/time values, etc.).

Programmatical access to SPARQL endpoints is exemplified in a Jupyter notebook collection, using data from DHwiki Wikibase as showcase.[24] In general, there is an ever-growing landscape of use cases that involve the reuse of Wikidata content,[25] but also involving other Wikibase instances. For the latter, we highlight two examples:

- The INCEpTION text annotation tool[26] can be connected to Wikidata[27] or

22. See https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/Wikidata_Query_Help/Result_Views.
23. See examples at https://dhwiki.wikibase.cloud/wiki/Structure_and_features_of_Wikibase_for_DH#Query_result_visualisations.
24. See 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. See homepage at https://inception-project.github.io/.
27. See https://inception-project.github.io/releases/39.2/docs/user-guide.html#knowledge_bases_in_getting_started.

any other Wikibase for the retrieval of items for manually supervised entity matching, or of lexemes, senses and/or forms for dictionary linking.[28]

- The OpenRefine application[29] can be connected not only to Wikidata but to any Wikibase.[30] OpenRefine is widely used for entity reconciliation (matching of literal values to Wikibase entities), though it can also be used for data enrichment, i.e., for upload.[31]

2.4 Data upload

QuickStatements[32] is a widely used tool for batch editing Wikibase instances. It allows users to create, modify, or delete large numbers of statements, labels, or descriptions through a simple tabular syntax.

Cradle[33] is a tool for creating Wikibase items using forms, which are created according to Entity Schemes. It is particularly useful for Wikibase administrators who want to facilitate the reliable manual creation of large numbers of items while ensuring they adhere to predefined data models and remain consistent across the database.

OpenRefine can read source data from spreadsheets and other formats, and perform upload, edit and delete operations using the MediaWiki Action API.

The **MediaWiki Action API** is a underlying technical web service interface that enables programmatic access to any MediaWiki-powered site. It allows developers and tools to read, edit, and manage wiki content without using the standard web interface. The MediaWiki API provides a number of methods to interact with Wikibase.

Through dedicated endpoints, users can send HTTP requests and retrieve page content, query metadata, upload files, modify pages, and perform administrative actions. The API supports multiple output formats (JSON, XML, PHP) and includes robust authentication mechanisms to ensure secure access.

This makes it essential for bots, automated maintenance tools, third-party applications, and integration services with external systems that need to interact with Wikibase at scale. Software development packages like the python WikibaseIntegrator[34] or Wikibase-sdk[35] for Javascript use this API. The API documentation[36] and versioning system help maintain backward compatibility while supporting the diverse ecosystem of tools built around MediaWiki platforms.

The **Wikibase REST API**[37][38] is modern web service interface included in newer distributions of Wikibase, and is available in Wikibase Cloud instances. It is designed

28. For linking text tokens to lexical entries, see https://github.com/inception-project/inception/discussions/4928.
29. See homepage at https://openrefine.org/.
30. See https://wetneb.github.io/new-openrefine-website/docs/manual/wikibase/configuration/.
31. See complete guidelines at https://en.wikiversity.org/wiki/OpenRefine_to_Wikibase:_Data_Upload_Pipeline.
32. See https://wikidata.org/wiki/Help:QuickStatements; the tool for Wikidata is available at https://quickstatements.toolforge.org/; a rewritten version of QuickStatements is available at https://qs-dev.toolforge.org/ (the manual is https://meta.wikimedia.org/wiki/QuickStatements_3.0/Documentation).
33. This page explains Cradle for Wikidata: https://wikidata.org/wiki/Wikidata:Cradle.
34. See https://toolhub.wikimedia.org/tools/wikibaseintegrator.
35. See https://github.com/maxlath/wikibase-sdk.
36. See https://www.mediawiki.org/wiki/API:Action_API.
37. See https://tech-news.wikimedia.de/2023/09/07/starting-fresh-the-wikibase-rest-api/.
38. See https://doc.wikimedia.org/Wikibase/master/js/rest-api/.

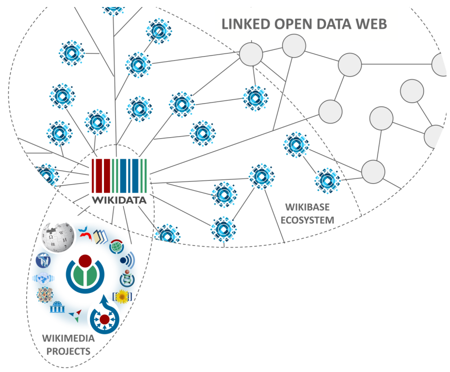

Figure 2: The Wikibase Ecosystem.

to offer a simpler, more structured way to interact programmatically with Wikibase entity data compared to the traditional, more complex MediaWiki API, which is intended to eventually supersede. Today, most of the tooling relies on the MediaWiki Action API.

Having an account in a Wikibase instance allows many **personalisations** to the user interface, involving the interface language, and own **user scripts**, the code of which entirely public, so that other users adopt may adopt or adapt them. Moreover, user scripts from Wikidata can be, sometimes with adaptations, reused in other Wikibase instances.[39]

3 Data round-tripping in the Wikibase ecosystem

Wikidata[40] is a free open knowledge graph maintained by the Wikimedia Foundation. While Wikidata is the most well-known and widely used public Wikibase instance, it is not the only possible one. In fact, running a custom Wikibase instance is technically feasible and often advisable. Any potential users should however be mindful of how interactions operate within this ecosystem to ensure effective integration and use.

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 Linked Open Data, volunteers and institutions are adopting their own Wikibase instances to manage specialized, sensitive, or domain-specific information. Such 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.[41]

39. See a more comprehensive summary of personalisation options at https://dhwiki.wikibase.cloud/wiki/Structure_and_features_of_Wikibase_for_DH#User_scripts.

40. Accessible at https://www.wikidata.org.

41. See a complete comparison of features in Wikidata vs. Wikibase at https://dhwiki.wikibase.cloud/wiki/Wikibase_and_Wikidata,_differences_and_round-tripping#Differences_between_working_with_Wikidata_and_an_own_Wikibase_instance.

Data round-tripping describes the process of iterating data across formats, sources, or infrastructures while preserving its quality, integrity, and consistency throughout its lifecycle. The term can be related to strictly technical aspects but also content-related ones more related to the need of preserving semantic equivalence.

On Wikidata, the term was primarily used to refer to the synchronization of data between Wikidata and the external databases to which it links, therefore implicitly referring to third-party databases and archives outside of the proper Wikimedia ecosystem (including Wikibase instances). As of December 2025, out of 13,088 properties,[42] there are 9,875 properties with the datatype *external identifier*;[43] of those, only 17 link to a Wikibase instance.[44] Due to such limited integration, the perception of data round-tripping for Wikidata volunteers largely still does not address the Wikibase ecosystem.

Considering the role of custom Wikibase instances is going to increase in this ecosystem at large,[45] data round-tripping will therefore also refer to their synchronization with Wikidata and other data sources, that is, to the complete cycle of exporting, modifying, and reimporting data between these platforms. Under this specific scenario, it is possible to outline certain good practices, which are especially relevant for projects that begin in a private Wikibase instance and plan to publish part of their data to Wikidata, projects that plan to maintain a local authoritative version while gradually contributing to the federated Wikibase ecosystem.

One of the main challenges of operating a custom Wikibase instance is synchronization with Wikidata and other data sources. The process of data round-tripping for Wikibase users remains inherently complex and fragmented due to multiple methodological factors, including:

- the imperative to avoid entity duplication;
- the essential requirement to respect the specific data models of all involved Wikibase instances;
- the necessity to preserve 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. Tools do not eliminate the conceptual complexity of round-tripping, but they significantly reduce the operational burden for editors working across multiple Wikibase environments.

In addition to the upload tools listed in Section 2.4, in this context, the *Beam Me Up* tool[46] is remarkable. That service automates the complex process of transferring data from Wikidata to FactGrid,[47] Traditional methods, like manual creation or importing via QuickStatements, can be time-consuming and prone to errors. This tool streamlines the entire workflow by automatically mapping properties and items between the two Wikibase environments.

42. See https://qlever.cs.uni-freiburg.de/wikidata/SVzmaE.
43. See https://qlever.cs.uni-freiburg.de/wikidata/ws7qR6.
44. https://w.wiki/EeLf lists Wikidata properties linking to another Wikibase.
45. See https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy2021/Wikibase.
46. See https://factgrid.wikidata.dbis.rwth-aachen.de/. The tool is itself based on the WikibaseMigrator (see https://github.com/tholzheim/WikibaseMigrator).
47. See https://database.factgrid.de/ a Wikibase instance. Factgrid supports projects with a specific interest in historical data and is part of NFDI4Memory (https://4memory.de/).

4 Wikidata and Wikibase as a hub for identifiers

The world's cultural and academic knowledge is stored in a multitude of separate systems that have shown limited interconnection for years, until the Semantic Web started to slowly shape form.

In such complex ecosystem, Wikidata has become an indispensable component of the global Semantic Web infrastructure, thanks to the amount of entities it describes (as of October 2025, more than 119 M)[48] and the number of external resources to which it is connected (as of October 2025, its properties with datatype *external-id* are more than 9 thousand[49]).

4.1 Wikidata and authority files

GLAM institutions, most notably libraries, maintain catalogues of their collections in order to make them accessible by the public. In such legacy 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 the users to find them are treated as controlled access points. The process of creating controlled access points, known as authority control, allows users to find all and only the resources they are looking for. Each controlled access point (e.g., each author) has an entry, known as authority record, containing basic data and linked to all 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 authority file.

Since 2013, Wikidata items link, among various types of databases, to authority files; the second and the third *external-id* properties created in Wikidata were the ones for ISNI (`P213`) and VIAF (`P214`). As of October 2025, Wikidata has 225 *external-id* properties for library authority files, used as value in 23.2 M statements contained in 5.8 M unique items, the majority of them (4.2 M) being personal items (i.e. having `P31` = `Q5`, *instance of human*). Among these 225 properties, the 5 most used by number of unique items are VIAF (3.8 M items), GND (2.4 M items), ISNI (2.2 M items), LC (1.7 M items) and IDREF (1.0 M items).[50] Wikidata URI t have become de facto a central pillar in the landscape of authority files (Van Veen 2019).

The workflows regarding the use of Wikidata in authority control, in the framework of a cooperation between cataloguers and Wikidata editors, have been illustrated in many presentations (most recently, (Pellizzari di San Girolamo 2025)). These workflows can be divided into the two parts described below.

4.1.1 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

48. See https://www.wikidata.org/wiki/Special:Statistics.
49. See https://www.wikidata.org/wiki/Help:Data_type
50. See https://qlever.dev/wikidata/XxlFHJ.

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.

The reconciliation between authority records and Wikidata items can be performed in three ways: manually, importing a previous list of matches (e.g. into Wikidata through QuickStatements (see section 2.4), or with the help of a reconciliation tool. The two main reconciliation tools used in Wikidata are Mix'n'match[51] and OpenRefine (see section 2.4). 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 to create new items putting together entries from different catalogues; conversely, OpenRefine is a software running on an individual computer, so each reconciliation project is developed by one individual user, and it contains functions allowing to clean messy data. Mix'n'match, in general, is more suitable for projects which are meant to be completed in a long span of time and require collaboration, and when most 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.

4.1.2 Cooperation

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 massively authority records and Wikidata items.

Regarding point 1: everyone can edit Wikidata items, but only cataloguers can usually edit authority records, so if a Wikidata editor finds in Wikidata a mistake imported from an authority record, there should be some mistake-reporting procedure, so that, once the mistakes are corrected in Wikidata, the same also happens in the authority record - the so-called data round-tripping (see Section 3). The mistake-reporting email or webform can be indicated directly in the Wikidata property of the authority file through `P10923`. On the other side, cataloguers can (as Wikidata editors) improve Wikidata items 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 items to 2 or more authority records) are clues of mistakes. These mistakes can be duplications or conflations, 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

51. See https://mix-n-match.toolforge.org/.

conjunct effort can produce an effective workflow, allowing to improve both Wikidata and the authority file. If the authority file has a SPARQL endpoint, making federated queries with Wikidata allows to easily spot inconsistencies (e.g., in gender, dates) and work on them (Kerboul 2025).[52]

Regarding point 3: Wikidata can be used as a source to massively enrich authority records; this can be done in two ways, retrieving the data on-the-fly through the API and showing it in an infobox, or extracting the data periodically and copying it into the authority records (especially the empty or nearly-empty ones). Both these ways have pros and cons: the second gives more control on the data showed, whilst the first has the advantage of being always up-to-date. Examples of the first include the AuthorityBox in the ILS Koha (Bargioni 2020) and the use of Wikidata inside the ILS Primo,[53] whilst the main example of the second is the enrichment of SBN authority records with data from Wikidata (Ravelli 2024). On the other side, 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 data import from an authority file, currently continuing on a regular basis through a bot, is the one managed by the National Library of the Czech Republic (Dostál 2021; Jansová, Maixnerová, and Šťastná 2024; Dostál 2025).

Stable cooperation between the Wikidata community and authority files involve, besides the National Library of the Czech Republic, the Italian National Library Service (SBN)[54] and the Union of Roman Ecclesiastical Libraries (URBE).[55]

4.2 Wikibase and authority files

Wikibase is currently used to maintain two authority files, the Semantic Name Authority Repository Cymru (SNARC) by the National Library of Wales,[56] started in 2022, and the National Library of Nigeria Semantic Name Authority Repository (NLN SNAR),[57] started in 2025 (Osuigwe, Boakye-Achampong, and Appiah 2025). Wikibase is also being experimented as a software for the authority file of the National Library of Greece; a proof of concept was deemed successful in 2024 (Zapounidou et al. 2024) and, as of 2025, the project is progressing (Gerolimos et al. 2025).

4.3 Controlled vocabularies

Wikidata can be used as a hub to interlink controlled vocabularies, both general-purpose ones and thematic ones. The main ongoing project regards the Nuovo Soggettario thesaurus, developed by the National Central Library of Florence, Italy:[58] the two focuses of the project have been fixing the previously accumulated non-biunivocal links, also through federated SPARQL queries and improving the interconnection

52. See https://www.wikidata.org/wiki/Property_talk:P269/Gender_mismatches.
53. See https://knowledge.exlibrisgroup.com/Primo/Product_Documentation/020Primo_VE/Primo_VE_(English)/150End_User_Help/Searching_Linked_Open_Data_-_Person_Entity.
54. See https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche/SBN.
55. See https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche/Parsifal.
56. See https://snarc-llgc.wikibase.cloud/.
57. See https://nlnsnar.wikibase.cloud/.
58. See https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche/Nuovo_soggettario.

on 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 Library and Information Science), into a Wikibase instance,[59] with the aim of discarding the previous platform and continuing the development of the thesaurus only on Wikibase.

4.4 GLAM data

GLAM institutions have been publishing digital collections during the last 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,[60] or networks such as Research Data Alliance[61] and the International GLAM Labs Community[62] promote the publication of digital collections able to support computational use. Relevant examples of institutions making their digital collections available in the form of reusable data are the National Library of Scotland[63] and the National Library of Luxembourg.[64] A project consortium at IFLA has identified Wikibase as a powerful and valuable 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 dedicated properties to link resources such as artists, art works, places, topics, etc.[65]

Rhizome ArtBase[66] is an outstanding example of deploying Wikibase for a GLAM collection, a collection of art works, in this case.

5 Final considerations and future prospects

Wikibase is emerging as a robust, collaborative, open, and sustainable infrastructure that can offer a structured ecosystem, particularly for cultural heritage and digital humanities projects, enabling data interoperability, discoverability, and long-term preservation.

Through its commitment to openness, technical interoperability, and community-centered governance, Wikibase can become a foundational component within global Linked Open Data ecosystems. Emerging research data infrastructures such as data spaces enable new ways of publishing and accessing research data according to transparency, trustworthiness and security. Other platforms such as Galaxy Community Hub[67] and the European Collaborative Cloud for Cultural Heritage[68] will promote

59. See https://tdkiv.wikibase.cloud/.
60. See https://collectionsasdata.github.io/.
61. See https://www.rd-alliance.org/groups/collections-as-data-ig.
62. See https://www.glamlabs.io/.
63. See https://data.nls.uk/.
64. See https://data.bnl.lu/.
65. https://qlever.dev/wikidata/IT9Tyc lists Wikidata properties related to museums, and the number of uses for each of them.
66. Accessible at https://artbase.rhizome.org.
67. https://galaxyproject.org
68. https://www.echoes-eccch.eu/

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.

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 Digital Humanities include:

- Provision of training materials in known platforms such as the Social Sciences and Humanities Open Marketplace[69] to facilitate the adoption of Wikibase as a tool for data publication and reuse.
- Promote the adoption of Wikibase in platforms such as DARIAH Campus[70] to plan, design, and implement educational programs based on Wikibase and DH.

We conclude with more general considerations, regarding sustainability and ethical issues:

- Regarding 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.
- For assessing the success of Wikibase as a knowledge infrastructure, institutions should define measurable Key Performance Indicators aligned with their strategic goals, enabling clear impact measurement toward funders and financing institutions. This aspect, critical for demonstrating the value of open infrastructures, remains underexplored in existing literature on Wikibase implementations.
- While the technical dimensions of FAIR principles are well aligned with Wikibase capabilities, these should be complemented by CARE principles (Collective Benefit, Authority to Control, Responsibility, and Ethics) (Carroll et al. 2020) to ensure that data stewardship is human-centered, ethically grounded, and attentive to Indigenous and marginalized community needs. Current literature on Wikibase has yet to fully emphasize its potential contributions to the human and ethical dimensions of data stewardship, highlighting an opportunity for further exploration and alignment with global ethical frameworks.

References

Bargioni, Stefano. 2020. "From Authority Enrichment to AuthorityBox: Applying RDA in a Koha environment." *JLIS.it* 11 (1): 175–189. ISSN: 2038-1026. https://doi.org/10.4403/jlis.it-12595.

Candela, Gustavo. 2023. *A workflow to publish Collections as Data: the case of Cultural Heritage data spaces.*

69. Accessible at https://marketplace.sshopencloud.eu/.
70. Accessible at https://campus.dariah.eu/.

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. ISSN: 1683-1470, accessed November 27, 2025. 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. ISSN: -.

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,* 63–74. Baden-Baden: Ergon – ein Verlag in der Nomos Verlagsgesellschaft. ISBN: 978-3-9874004-7-6. 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. ISSN: 0163-9374. https://doi.org/10.1080/01639374.2025.2502329.

Dostál, Vojtěch. 2021. *Bringing Czech authority files into 21st century: Integration with Wikidata.* Presentation. online, September.

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. Accessed April 22, 2024. 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. ISSN: -. https://doi.org/10.5282/O-BIB/6081.

Kerboul, Thomas. 2025. *Enhancing authority files through SPARQL federated queries.* online, November.

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. Accessed July 7, 2025. https://doi.org/10.5281/zenodo.15471514.

McCrae, John, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. "The OntoLex-Lemon Model: Development and Applications." In *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017,* edited by Iztok Kosem, Carole Tiberius, Miloš Jakubíček, Jelena Kallas, Simon Krek, and Vít Baisa, 587–597. Brno: Lexical Computing CZ s.r.o.

Osuigwe, Nkem, Stanley Boakye-Achampong, and Doreen Appiah. 2025. *How to Wikibase: The Case of the NLN SNAR.* Presentation. Nairobi, Kenya, August.

Pellizzari di San Girolamo, Camillo Carlo. 2024. "Conflations and Duplications in Wikidata Items: Causes, Detection, Solutions, and Issues." In *Proceedings of the Wikidata Workshop 2023 co-located with 22nd International Semantic Web Conference (ISWC 2023).* Athens: CEUR.

Pellizzari di San Girolamo, Camillo Carlo. 2025. *Links between Wikidata, the Thesaurus Nuovo soggettario and other thesauri,* September. https://doi.org/10.5281/ZENODO.17235305.

Ravelli, Elena. 2024. *Authority file di SBN e Wikidata. Un progetto per l'arricchimento dei dati del Catalogo nazionale.* Presentation. Bologna, Italy, November.

Tasovac, Toma, Sally Chambers, and Erzsébet Tóth-Czifra. 2020. *Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper.*

Van Veen, Theo. 2019. "Wikidata: From "an" Identifier to "the" Identifier." *Information Technology and Libraries* 38, no. 2 (June): 72–81. ISSN: 2163-5226. https://doi.org/10.6017/ital.v38i2.10886.

Viti, Elisabetta, and Camillo Carlo Pellizzari di San Girolamo. 2025. *Nuovo Soggettario and Wikidata: comparison between two KOSs.* Presentation. Bern, Switzerland, August.

Waagmeester, Andra, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M. Good, Malachi Griffith, Obi Griffith, Kristina Hanspers, Henning Hermjakob, Toby S. Hudson, and Kevin Hybiske. 2019. "Wikidata as a FAIR knowledge graph for the life sciences." *bioRxiv,* 799684. Accessed May 20, 2024.

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, no. 2 (April): 111–131. ISSN: 1938-6389. https://doi.org/10.1080/19386389.2024.2307208.