Wikidata and Wikibase as complementary research data management services for cultural heritage data

Lozana Rossenova [1], Paul Duchesne [1] and Ina Blümel [1]

[1]TIB – Leibniz Information Centre for Science and Technology, Welfengarten 1B, 30167 Hannover, Germany

Abstract

The NFDI (German National Research Data Infrastructure) consortia are associations of various institutions within a specific research field, which work together to develop common data infrastructures, guidelines, best practices and tools that conform to the principles of FAIR data [1, 2]. Within the NFDI, a common question is: What is the potential of Wikidata to be used as an application for science and research [3]? In this paper, we address this question by tracing current research use-cases and applications for Wikidata, its relation to standalone Wikibase instances [4], and how the two can function as complementary services to meet a range of research needs. This paper builds on lessons learned through the development of open data projects and software services within the Open Science Lab at TIB, Hannover, in the context of NFDI4Culture – the consortium including participants across the broad spectrum of the digital libraries, archives, and museums field, and the digital humanities [5, 6].

Keywords

Wikidata, Wikibase, research data management, cultural heritage, open science, NFDI, NFDI4Culture

1. Introduction

Wikidata was released in 2012 and originally intended to resolve concrete issues pertaining to Wikipedia [7]. It aimed to reduce data redundancy and serve as the language-agnostic data source for infoboxes that are now ubiquitous across Wikipedia pages. Wikidata is distinct from its sister project, Wikipedia, for the fact that it stores structured data and makes that data accessible via a SPARQL endpoint, providing a machine-readable service in contrast to Wikipedia's primarily human-readable interface. As of July 2022, it stores well over a billion statements and 99 million items on subjects across a vast range of knowledge domains [8,9]. It highlights the power of a centralized and distributed approach [10]: a vast amount of information is accessible through a central endpoint in a standardized linked open data format, all items have PIDs (persistent identifiers), and all that information is crowdsourced through collaborative editing efforts. Edits made by both human users and bots are tracked in a reliable version control system with clear provenance and the ability to discuss, debate and (sometimes) revoke each edit [11].

These qualities make Wikidata an attractive environment for data storage, curation and extraction. It is already widely used across many domains of knowledge management,

The 3rd Wikidata Workshop, Workshop for the scientific Wikidata community, @ ISWC 2022, 24 October 2022.
Lozana.Rossenova@tib.eu (L. Rossenova); Paul.Duchesne@tib.eu (P. Duchesne); Ina.Bluemel@tib.eu (I. Blümel)
0000-1111-2222-3333 (L. Rossenova); 0000-0002-3075-7640 (I. Blümel)

including scientometrics (e.g. WikiCite [12] and Scholia [13] initiatives) and academic research data management. Communities of researchers in the life sciences [10], computer science and digital preservation [14], as well as cultural heritage [15], among other fields, have already documented their work and experience using Wikidata as research infrastructure. But they have also documented where the issues born out of dealing with a bottom-up ontology design and curation [16], the burden of vastness of scale on performance [17], the restriction of Wikidata being a secondary database (vs original research repository) [18], and ultimately trustworthiness [19] have become problematic in the context of established scientific practice.

Despite these potential disadvantages, the growing use of Wikidata – both as a repository to upload data to, and a rich resource on the linked open data (LOD) cloud to federate with – can also be presented as an exciting 'proof of concept'. In this paper, we discuss how its underlying software suite (Wikibase) can fulfil many of the core requirements of scientists and researchers dealing with structured data, while at the same time removing some of the issues born out of scale, governance policy, and ontology particularities. Furthermore, we showcase how independent Wikibase instances can be effectively deployed as research infrastructure for cultural data that complements and can further enrich, but also benefit from, Wikidata's already rich network of connections.

2. The drawbacks of an open editing policy for research data

The primary strength of Wikidata is currently its size, both in relation to coverage of subject area and also the sheer scale of data points. The desire to include data pertaining to all areas of human knowledge is ambitious, reflected by the Wikimedia observation that it "only works in practice. It could never work in theory" [20].

This growth has been enabled by providing a relatively low barrier of entry for contributor involvement, with no impediment to anyone with internet access being granted the ability to edit production data.[2] This low barrier of entry has been purposefully designed as a means of encouraging community involvement.[3] An effect of this is that there is no inherent preference given to a user's expertise in a specific area, which means that a domain expert could find their contribution overwritten by a user with only a cursory knowledge of the subject. In practice this appears to be mitigated by obscure specialist data 'hiding in plain sight', and by the fact that vandalism mostly targets highly visible entries (for which circumstances some page protection tools have been developed [21]). The lack of editing limitation is not only applied to human users, as there are many resources for setting up discrete editing bots which amend data based on preconfigured logic, and which can be authenticated to operate autonomously once granted permission by the community [22]. The use of these services does require prior manual editing history to prove the 'good' intent of the individual. Furthermore, the ability to revoke data edits based on contributor information is one of the platform's strongpoints.

There are however no mechanisms to assert authority based on access to physical evidence (which is particularly relevant to cultural institutions, i.e. access to artefacts themselves) or

[2]'Production' is used here in the sense defined by the Agile development model: http://www.agiledata.org/essays/sandboxes.html

[3]Wikidata's vandalism strategy is mostly reparative, not preventative, relying on extensive versioning information to rollback undesirable edits. See: https://www.wikidata.org/wiki/Wikidata:Vandalism

first-hand experiences [23]. The combination of these factors makes an institution's decision to contribute and/or rely on Wikidata a complex one, as they could easily find themselves in the role of 'digital gardener', not just contributing data but having to maintain their statements from alteration [19].

There are also general issues with citation on the platform. There is a well-supported method for applying a reference link to any statement – most often as a web link to a secondary source which supports the claim. However, this is not a mandatory attribute, so many statements are simply presented 'as-is' – without any justification or means of verification. The heavy reliance on web links also means that 'link rot' is a concern. A promising area of investigation is the potential future use of 'signed statements'[4] to allow institutions and authority sources to authenticate and endorse claims which are displayed and attributed to them. This would allow subject-matter experts the ability to assert their domain knowledge and elevate their ability to verify a statement over the claims of a causal user.

3. The trouble with standard ontologies

The open editing policy also leads to challenges related to ontological coherence. Wikidata was conceived with an inherently flat structure in relation to creative works, due to its primary function as facilitating data exchange between different language Wikipedia resources, which resulted in an initial one-to-one relationship between the two platforms. Increasingly there has been community-driven work to define multi-tiered structures for cultural data (such as WikiProject Books [24]), but there are issues with this approach which can be illustrated with a closer look at the representation of data related to literature.

Contemporary models for cataloguing (for example FRBR [25], BIBFRAME [26] or FIAF [27]) rely on structures generally involving at least three tiers. This begins with a 'work' or 'expression' to represent the artwork as an abstract entity, a 'manifestation' or 'edition' for each interchangeable batch of physical material, and an 'item' to represent an individual physical artefact. For example, *Alice's Adventures in Wonderland* (meaning specifically the book by Lewis Carroll) is an identified singular artistic 'work', individual ISBNs[5] delineate different 'editions', and an 'item' is a physical copy sitting on a shelf. It is unreasonable to expect individual 'items' to be represented in Wikidata unless they are especially notable,[6] although it is worth observing that some disciplines place a great deal of importance on a specific item being identified.[7] The primary tension arises from attempting to enforce the 'work' and 'edition' levels of the schema as distinct ontological elements on a platform not initially intended to facilitate any strict ontological orders of classes and subclasses. The wide reach of Wikidata, coupled with finite infrastructure resources, means that it is not possible to incorporate highly granular data for every represented field. This further contributes to

[4]See the proposal and community discussion for this cryptographic method for endorsing and verifying claims on Wikidata (note, however, that this is still under discussion and not a method already implemented in Wikidata): https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Signed_Statements_(T138708)

[5]ISBNs indeed provide a rare example where a URI is already 'minted', attached to a physical item and ready to be reused as a persistent identifier for that edition.

[6]One of few examples of an 'item' level book entity is the Lincoln Bible. See: https://www.wikidata.org/wiki/Q1816474

[7]There is a great variety amongst extant artefacts of early cinema, in many cases due to distinct coloring or modification processes performed on individual prints.

preventing adoption as a primary research platform, as it is precisely this detail which would make it a more valuable resource for many researchers.

Another area which requires attention is the enforcement of class-specific schemas. Due to a lack of restrictive mechanisms it is currently possible to make data statements which are completely nonsensical. This appears to be mostly avoided by a lack of interest in this form of vandalism, in addition to an active community who pursue what are deemed incorrect edits. Still, a significant step towards being a more trusted resource would be the inclusion of strict data validation which could build upon the existing EntitySchema extension, if it is scaled beyond its current application of serving mostly as a guideline [28].

4. Wikibase as research data management (RDM) service

Thanks to many of the features developed for Wikidata, Wikibase is already more than a single tool. It can instead be considered an umbrella of services [29], including:

- A graphical user interface (GUI) to a database with collaborative and version control features;
- A triplestore, editable down to the triple statement level (thanks to the requirements of Wikidata to serve as a verifiable source and secondary database, every triple statement can be edited and enriched with qualifying statements and references to external sources – all achievable via the GUI);
- A SPARQL endpoint with its own GUI;
- An API for read/write programmatic access;
- A wide range of script libraries (PyWikibot, WikidataIntegrator), as well as additional user-friendly tools (QuickStatements, Cradle, OpenRefine reconciliation service), for data import/export.

Private Wikibase instances rarely have to deal with performance issues born out of Wikidata's scale, at least not until they grow into the magnitude of dealing with many millions of items. Importantly, private Wikibase instances can store original research data, as they don't share Wikidata's policy for being only a secondary database. In that sense, individual Wikibase instances can serve as primary sources for data to be later referenced in Wikidata (this type of decentralization is in fact in alignment with the long-term strategy and vision behind Wikidata itself and the broader Wikimedia movement [30, 31]). Private Wikibase instances can also hold various data licenses, or remain entirely closed off from the open web, depending on the nature of the research and the need for privacy control (note that another restriction of Wikidata is the need for data to be licensed CC0).

When installed 'out-of-the-box', Wikibase provides users with an experience nearly identical to that of Wikidata. There are minimal options for custom branding, but the default interface follows the same templates for creating, editing, or simply viewing item and property pages as Wikidata [32]. The SPARQL endpoint GUI is also identical to Wikidata's query service. Additional tools from the Wikidata ecosystem (as import/export tools) can also be adapted to Wikibase instances [33, 34]. This provides ease of transition from one service to the other. Since there is already extensive documentation and training material for Wikidata users [35], much of the same material and approaches to training can be adopted when introducing Wikibase in research and/or cultural institution environments. Thus it can be argued that Wikibase is well suited to fulfil the need for end-to-end services to 'LOD'-ify research data, while at the same time easing the learning curve to working with LOD

compared to other existing knowledge graph tools [36]. However, Wikibase is not entirely free of the issues born out of the open approach to defining a data model and ontology and the simultaneous lack of validation mechanisms in Wikidata.

Wikibase follows the conventions for data structuring (and PID generation) set out by Wikidata (so all items receive a Q number, whereas properties receive a P number), however there is no requirement to follow Wikidata's upper ontology [16], and Wikibase users can define their own entities (items) and relations (properties) from scratch. Here, the researchers deploying a Wikibase instance for their own datasets will be in control of how data is curated and described, while taking advantage of the familiar graphical user interface for data entry and editing. There will be no need to deal with the 'messiness' of collaborating with a vast and largely anonymous international community, which expands Wikidata's vocabulary primarily in response to the needs of the Wikidata and Wikipedia projects.

4.1. Addressing the 'open world' scenario of modelling data in Wikibase

The possibility to define new LOD models with the help of a GUI can be a liberating prospect for disciplines that defy conventions and do not fit neatly in previously established metadata standards (particularly good case in point being contemporary art data [36]). But this can also be an issue for researchers who want (and in some cases may be required) to work within an established domain standard and to reuse ontologies maintained elsewhere on the semantic web. Although there are ways to map local Wikibase properties and classes to external ontologies, the native RDF structure remains relatively 'flat' and lacks the semantics needed for more sophisticated reasoning operations over its triplestore [3].[8] Furthermore there are no formal constraints or validation rules that can be applied to Wikibase's flat data structure as is.[9]

An established tradition in free and open source software (FOSS) culture is that if a tool doesn't perform to user expectations, users have the choice to either find workarounds or they can write a patch, or an extension of the tool, and submit a pull request for the issue at hand. As a FOSS tool, Wikibase benefits from this ethos and its community has been collaborating in finding workarounds (e.g. see Figure 1) or developing entirely new features.

Mapping local properties and items to Wikidata, for example, has been piloted and shown to deliver successful results in query federation [37] and data syncing [38]. Developing data modelling principles and import/export pipelines for Wikibase that explicitly match to established schemas or ontologies, such as the standard ontology for cultural heritage data CIDOC-CRM [39], has also shown promising results in moving away from the complete 'open world' paradigm [40, 41].

---

[8]Still, Wikidata and Wikibase offer some capabilities out-of-the-box that are actually less 'flat' than other RDF resources, for example the possibility to attach references or qualifiers to individual triple statements. In addition, it is worth noting that Wikidata contains properties like wdt:P279 and wdt:P31 which are direct translations of rdf:type and rdfs:subclass of. Waagmeester [10] has demonstrated the potential to use CONSTRUCT queries to get the semantics needed for reasoning also in Wikidata, e.g. see this example query: https://w.wiki/4x49

[9]This is somewhat addressed by the EntitySchema extension [28] which can bring machine-readable boundaries to any Wikibase, though the values in statements defined by an entity schema are still not validated apart from data type, e.g. a statement with a property calling for a country can still be (wrongly) populated by a city item, because the only constraint is to add a Wikidata item in the value field associated with the country property.

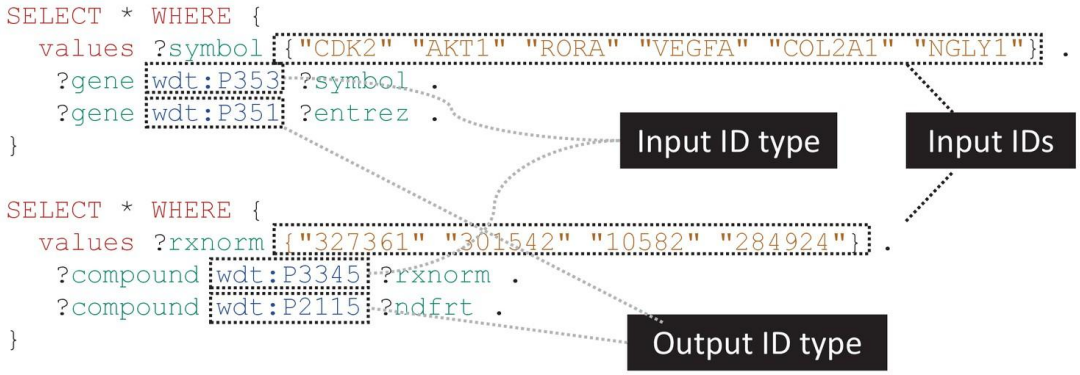

Figure 1: Generalizable SPARQL template for identifier translation. These simple SPARQL examples show how identifiers of any biological type can easily be translated using SPARQL queries. Image by Waagmeester, et al (2020) [10].

While the above can be described as workarounds or mitigation tactics, there are also efforts among user communities to redevelop the explicit logic of how RDF is written to the Wikibase triplestore. Within NFDI4Culture, the team at FIZ Karlsruhe have experimented with methods that write semantic RDF directly into the triplestore, thus circumventing the limitations of the MySQL database that is connected to the Wikibase user frontend and serves as default source for RDF statements written to the triplestore [3]. The issue with the latter methods is that these statements are not bi-directional and cannot be used when continuous edits need to be performed via the frontend user interface. An alternative plan for an extension that creates explicit mappings between Wikibase items and properties and standard OWL or SKOS classes and properties is being discussed and prepared among a Wikibase community group dedicated to new (re)developments supplementing the official product roadmap set by Wikimedia Germany (the lead maintainers of Wikibase) [42]. While this extension alone does not solve all the issues regarding the use of formal semantics and standard ontologies in Wikibase, it is an illustration of the capacities within FOSS communities to respond to user needs – an important factor to consider when researchers select tools and infrastructures for long-term projects.

4.2. Wikibase RDF extension

The Wikibase RDF extension [43] currently in development will work via implementing a graphical configuration window into each Wikibase entity page, which will enable users to specify correlations between Wikibase entities and externally defined RDF resources, using commonly-used OWL and SKOS methods of declaring two resources to be ‘same’ or ‘similar’ (see Figure 2).

This extension will also transfer these correlation statements into the triplestore. However, the triplestore’s default configuration currently precludes leveraging any of the inferencing or knowledge generation features which are present in modern knowledge graphs. The next step is to reconfigure the Blazegraph triplestore to incorporate and appropriately process these newly generated mapping triples, and also to explore routing data generated by this extension directly into other triplestores. At the Open Science Lab, we are actively supporting

work on this extension and triplestore reconfiguration, with the aim to allow independent Wikibase instances to be used both as practical tools facilitating introduction to the semantic web, and also as fully-featured knowledge graphs meeting the needs of the NFDI research consortia.

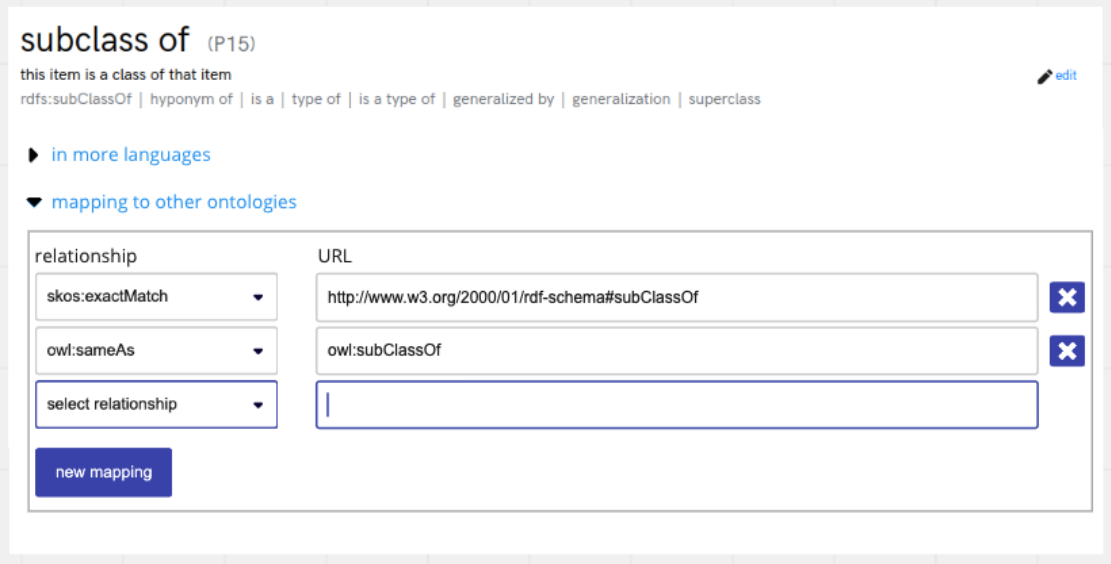

Figure 2: Design mockup of the new graphical interface for property mapping via the Wikibase RDF extension. Courtesy of Dragan Espenschied (Wikibase Stakeholder Group).

5. Linked research data: increasing granularity and specificity

So far we have examined some of the respective strengths and weaknesses of Wikidata and Wikibase, but comparing one against the other as a service is not the aim. Rather, the approach we see as most productive for a comprehensive research data management infrastructure is one where the two services are utilized side-by-side to produce richly linked research data at various levels of granularity and specificity, while retaining licensing and a degree of 'openness' appropriate to each use-case's context. In this approach, Wikidata remains a 'hub' service, linking together a family of independently maintained Wikibases, each containing a vast expansion of detail for a given subject area. If we take the previously used example – *Alice's Adventures in Wonderland* – the book remains as a single entity on Wikidata (the 'work'), connected via external identifiers with other Wikibases which need to link their resources (most obviously in this instance, Wikibases run by libraries).[10]

The decision to either use Wikidata as primary data repository, or deploy an independent Wikibase and then interlink with relevant Wikidata entities, ultimately depends on the granular nature of the data itself. At the Open Science Lab we have two ongoing cultural data projects which provide good illustration of the decisions involved in choosing either path. We

[10]There are already a number of library initiatives working in this direction See: https://www.wikidata.org/wiki/Wikidata:WikiProject_LD4_Wikidata_Affinity_Group and https://www.wikimedia.de/the-wikilibrary-manifesto/

discuss these use cases below to illustrate how both projects support the above-mentioned approach to data management.

5.1. Case study A: DigAMus Award

To visualize the advantages of structured, networked information that can be edited by anyone, we first discuss the 'DigAMus goes Wikidata' project. The DigAMus Award, which honors successful digital offerings from museums in German-speaking countries, grew out of a grassroots movement, and continues to be organized on a voluntary basis [44]. The first call for submissions in 2020 received 129 entries – far more than expected. Building on our experience with similar projects with potential to facilitate work through communities [45], we suggested curating structured data about DigAMus in Wikidata, instead of a simple spreadsheet. We were able to demonstrate the benefits to museums who can themselves expand Wikidata entries on an ongoing basis with further information on their projects. We also explained the benefits for museums in maintaining information about them and their activities available as LOD, and thus accessible for further applications [46]. Crucially, the decision to choose Wikidata was informed by the fact that there was no need to create any new items or properties in order to represent the knowledge as desired.[11] We created and mapped an appropriate data model,[12] imported data via the easily accessible, open source tool OpenRefine [47], and provided pre-built queries[13] to demonstrate the possibilities of the networked data. Visualizations of the query results were a particularly valuable outcome of the project – presenting a comprehensive and visually-appealing overview of all submitted projects.

For legal reasons [48], we were not allowed to simply create thumbnail-images from the project web pages, store them in Wikimedia Commons and link to Wikidata so that they can be displayed in the visualizations. This was a learning experience for the Award organizers and the submission of a CC-BY-licensed thumbnail image for the respective project was requested for the second call in 2021. The projects from the second year's DigAMus Award were added to Wikidata right at the point of submission so that they could already be searched via queries, along with former projects.

Despite the successful track record for adoption of Wikidata in GLAMs,[14] Wikidata and especially its SPARQL endpoint at first sight appear challenging to many museum staff members who can be potential contributors. Therefore, we emphasized community building throughout this project and distributed information on contributions via a dedicated WikiProject page [49]. Together with DigAMus, we offered an online hands-on workshop explaining how to contribute to the project and how to make one's own institution more

[11] See a list of the items and properties we utilised at: https://www.wikidata.org/wiki/Wikidata:WikiProject_Digital_projects_of_museums/DigAMus_Award#DigAMus_Award_-_Wikidata_items

[12] Graphical draw.io data model for mapping the DigAMus data in Wikidata is available at: https://www.wikidata.org/wiki/Wikidata:WikiProject_Digital_projects_of_museums/DigAMus_Award#Datamodel_visualization

[13] Selected queries are listed at: https://www.wikidata.org/wiki/Wikidata:WikiProject_Digital_projects_of_museums/DigAMus_Award#Querys

[14] See numerous GLAM-institutional pages related to their work in Wikidata, which are linked from: https://www.wikidata.org/wiki/Wikidata:GLAM

visible via Wikidata and Wikimedia Commons.[15] Unexpected feedback from the workshop was that due to the deliberately-chosen low entry level, the participants "dared to ask very basic questions for the first time" – an important step towards much needed increase in adoption of LOD and FAIR data principles in the cultural heritage and humanities fields [50, 51].

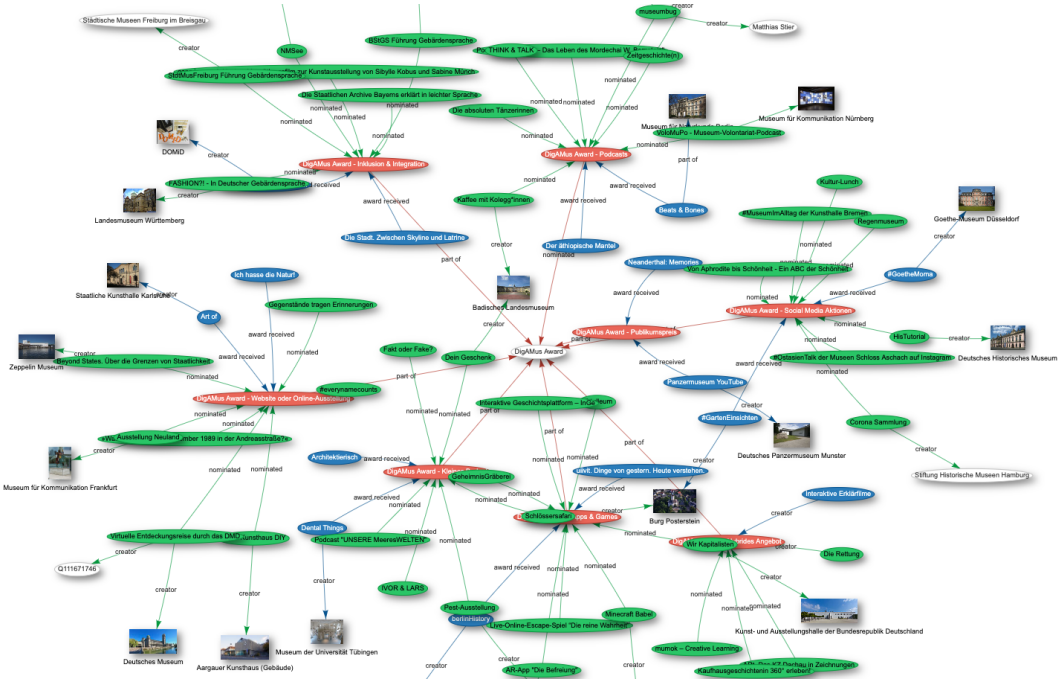

Figure 3: Image showing museum projects in context. Query available from the WikiProject page at: https://bit.ly/3Ju2O3E

Following the proposed model in this paper – of Wikidata acting as a hub service – individual museums could easily run dedicated Wikibase instances for specific projects that may be included in DigAMus in the future. Such Wikibases would contain far more granular data – needed for a specific project execution, for example, and then link out to the more general, overview data about all the Award submissions and related institutions in Wikidata.

The DigAMus Award project was well-suited to highlight the importance of LOD for cultural data, given that the Award itself promotes the exploration of varied applications of digital technology in museums.

5.2. Case study B: Semantic annotation of 3D models, a minimum viable product (MVP)

In the context of our work within NFDI4Culture's 'Task Area 1: Data Capture and Enrichment', we have identified the need for open and collaborative digital infrastructure that

[15] See workshop documentation: https://docs.google.com/document/d/12KmJqfEIRTj3DCQTygDMc5A67SnRmN6RvxivaO82YVY/edit

allows for the storage, access and annotation of various digitized cultural heritage objects, including complex 3D media files. Such an infrastructure needs to address the challenges specific to 3D media, e.g. copyright licenses that preclude 3D files which utilize special textures from being uploaded to open, public resources like Wikimedia Commons. Furthermore, the need for scholarly annotations presents a good illustration of the granularity argument. Unlike the more general data needed to describe digital museum projects presented in the previous case, individual annotations created by different scholars to describe a single 3D model and to make a scholarly argument are far too granular to store in a data repository such as Wikidata. Hence, the need to store data in a dedicated Wikibase instance, which can be linked to relevant, less specific items already stored in Wikidata such as physical buildings, geographic locations, historical figures, artistic styles, and more. To address these concrete infrastructural challenges, we developed an integrated toolchain that consists of three main open source software components (see Figure 4): 1) OpenRefine – for data reconciliation and batch upload; 2) Wikibase – for linked open data storage; and 3) Kompakkt – for rendering and annotating 3D models, and other 2D and AV media files [52]. All components of the toolchain feature graphical user interfaces aiming to lower the barrier of participation for a wide range of cultural practitioners and researchers.

In Phase 1 of this project, we developed an MVP which works with a specific art and architectural data research project – reconstruction work and 3D modelling of the Weikersheim castle and its baroque ceiling paintings [53]. This very specific dataset allowed us to work with real world data and develop all aspects of the toolchain with concrete user requirements in mind. At the same time, the combination of art historical data, architectural data, 2D images, 3D models and attendant metadata, as well as annotations, provides a very rich sample of highly heterogeneous data in order to serve as proof of concept that this toolchain can be deployed in multiple instances and applied beyond the field of architecture to a wider range of related cultural disciplines.

During Phase 2 of the project development, we are working to develop a common data model around media files, their annotations and connections to objects in the physical world, that can fit different cultural digitization use cases [54]. The first version of this data model intentionally features significant overlap with Wikidata properties, to facilitate federated querying (see Figure 5). These property mappings only cover relations concerning physical objects and historical figures. For annotations of digital media, we have developed a custom mapping between Kompakkt's data model, which follows the W3C Web Annotation standard [55], and Wikibase. Once it is possible to deploy the new Wikibase RDF extension, we will be able to map this data model in RDF-compliant ways to the W3C Annotation standard, as well as to more established standard schemas and ontologies commonly used in the cultural field, such as CIDOC-CRM, and better integrate widely used vocabularies and thesauri, such as the Getty's Art and Architecture Thesaurus [56] and Iconclass [57], among others.

Thanks to the capacity to store data in RDF format and the public SPARQL endpoint of any Wikibase instance, deploying multiple instances of our toolchain for multiple project partners does not mean that their resources will be siloed. Data from multiple collections representing different degrees of granularity and specificity within a given cultural domain, will remain interoperable by following our common, generalized data model with links to Wikidata – the latter retaining its role as the hub service – and with mappings to common ontologies and thesauri.

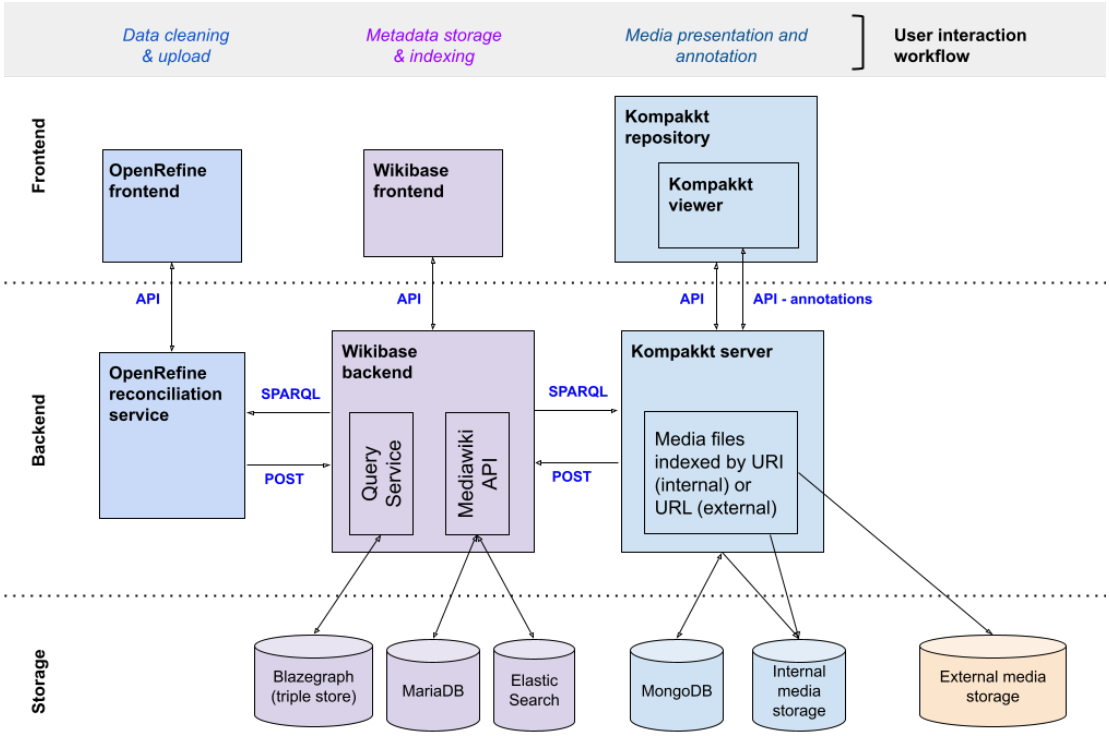

Figure 4: Diagrammatic representation of the MVP toolchain architecture

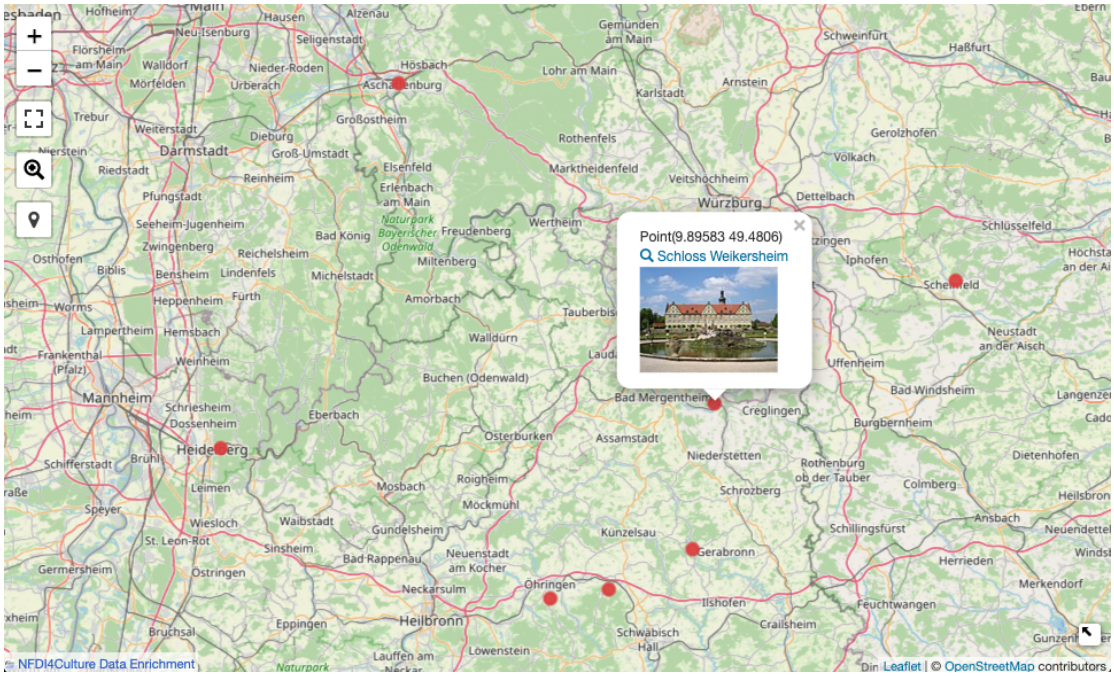

Figure 5: Map visualization of a sample federated query showing castles built in Renaissance architectural style within 100km of Weikersheim: https://tinyurl.com/26q6bl5j

6. Outlook and long-term sustainability

As the many projects we have cited in this paper and the two case studies from the Open Science Lab attest, both Wikidata – as a public, centralized *and* distributed vast repository of knowledge – and Wikibase – with its possibility to be deployed as stand-alone software product – can serve as working, albeit imperfect, solutions for the needs of scientists and researchers managing heterogeneous datasets.

Furthermore, the ability to deploy Wikibase is becoming increasingly easier given a focus on providing containerized packages and installation templates [58], with community engagement and growth being a goal explicitly stated in the Wikimedia Linked Open Data Strategy [31]. As with all ecosystems, the likelihood of this model being fully realized is dependent on adoption, and the willingness of relevant organizations to engage with this vision. It is also worth noting that developing any infrastructure (whether a Wikibase instance, or another triplestore) in the context of a research project is tied to (and limited by) grant funding cycles. At the end of a research project, infrastructure often decays. Using Wikibase offers seamless alignment with Wikidata, the latter being a stable infrastructure independent of such funding cycles. Data originally stored in a Wikibase instance can easily be exported and made permanently available on Wikidata. It can also live on in its RDF form on many other RDF platforms. In short, Wikibase can work as a 'proxy' in the research data landscape that allows long-term sustainability of the acquired knowledge [59].

Acknowledgements

NFDI4Culture is funded by the Deutsche Forschungsgemeinschaft (DFG) under grant no. 441958017.

References

[1] NFDI, Homepage, 2022. URL: https://www.nfdi.de/.
[2] Mark Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al, The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[3] Ina Blümel, Paul Duchesne, Lozana Rossenova, Harald Sack, NFDI InfraTalk: Wikibase - knowledge graphs for RDM in NFDI4Culture (7 March 2022). URL: https://www.youtube.com/watch?v=RPMkuDxHJtI&ab_channel=NFDIDirektorat.
[4] Wikibase, Homepage, 2022. URL: https://wikiba.se/.
[5] Open Science Lab, Homepage, 2022. URL: https://www.tib.eu/en/research-development/research-groups-and-labs/open-science.
[6] NFDI4Culture, Homepage, 2022. URL: https://nfdi4culture.de/.
[7] Denny Vrandečić, Markus Krötzsch, Wikidata: A Free Collaborative Knowledgebase. Communications of the ACM 57 (10) 78–85 (2014). https://doi.org/10.1145/2629489.
[8] Wikidata, Stats Homepage, 2022. URL: https://wikidata-todo.toolforge.org/stats.php.
[9] Wikidata, Homepage, 2022. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page.

[10] Andra Waagmeester, et al, Science Forum: Wikidata as a knowledge graph for the life sciences. *eLife* 2020; 9:e52614 (2020). DOI: 10.7554/eLife.52614.
[11] Andrea Piscopo, Structuring the World's Knowledge: Socio-Technical Processes and Data Quality in Wikidata. PhD thesis, University of Southampton, UK, 2019.
[12] WikiCite, Homepage, 2022, URL: https://meta.wikimedia.org/wiki/WikiCite.
[13] Scholia, Homepage, 2022. URL: https://scholia.toolforge.org/.
[14] Katherine Thornton, Kenneth Seals-Nutt, Euan Cochrane, Carl Wilson, Wikidata for Digital Preservation, in: Proceedings of iPRES'18, Cambridge, MA, USA, September 24–27, 2018.
[15] Effie Kapsalis, Wikidata: Recruiting the Crowd to Power Access to Digital Archives. Journal of Radio & Audio Media 26 (2019) 134–142.
[16] Lydia Pintscher, Silvan Heintze, Ontology issues in Wikidata. in: Data Quality Days, online 2021. URL: https://commons.wikimedia.org/w/index.php?title=File%3ADataQualityDaysontologyissues.pdf
[17] Mike Pham, et al, Scaling Wikidata Query Service – unlimited access to all the world's knowledge for everyone is hard, in: WikidataCon 2021, online, 2021. URL: https://www.youtube.com/watch?v=oV4qelj9fxM&ab_channel=wikimediaDE.
[18] Daniel Mietchen, Gregor Hagedorn, Egon Willighagen, et al, Enabling Open Science: Wikidata for Research (Wiki4R). Research Ideas and Outcomes 1: e7573 (2015). doi: https://doi.org/10.3897/rio.1.e7573
[19] Martin Zeinstra, Returning Commons Community Metadata Additions and Corrections to Source, Swedish National Heritage Board, 2019. URL: https://meta.wikimedia.org/wiki/File:Research_Report_–_Returning_commons_community_metadata_additions_and_corrections_to_source.pdf
[20] Foundation-l Mailing List, The problem with Wikipedia, 2010. URL: https://lists.wikimedia.org/pipermail/foundation-l/2010-June/059273.html.
[21] Wikidata, Protection Policy, 2022. URL: https://www.wikidata.org/wiki/Wikidata:Protection_policy.
[22] Wikidata, Requests for permissions, 2022. URL: https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot.
[23] Philip Roth, An Open Letter to Wikipedia. New Yorker (September 6, 2012). URL: https://www.newyorker.com/books/page-turner/an-open-letter-to-wikipedia.
[24] WikiProject Books, Homepage, 2022. URL: https://www.wikidata.org/wiki/Wikidata:WikiProject_Books.
[25] IFLA, Functional Requirements for Bibliographic Records, 2009. URL: https://www.ifla.org/publications/functional-requirements-for-bibliographic-records.
[26] Library of Congress, Bibliographic Framework Initiative, 2022. URL: https://www.loc.gov/bibframe/.
[27] Natasha Fairbairn, Maria Assunta Pimpinelli, Thelma Ross, The FIAF Moving Image Cataloguing Manual, International Federation of Film Archives (2016). URL: https://www.fiafnet.org/pages/E-Resources/Cataloguing-Manual.html.
[28] Mediawiki, EntitySchema Extension, 2021. URL: https://www.mediawiki.org/wiki/Extension:EntitySchema.
[29] Renat Shigapov, RaiseWikibase: Towards fast data import into Wikibase. in: 2nd Workshop on Wikibase in Knowledge Graph based Research Data Management (NFDI)

Projects, online, 2021. URL: https://madoc.bib.uni-mannheim.de/60059/1/29.07.2021-RaiseWikibase-Shigapov.pdf.
[30] Lydia Pintscher, et al, Strategy for the Wikibase Ecosystem (2019). URL: https://upload.wikimedia.org/wikipedia/commons/c/cc/Strategy_for_Wikibase_Ecosystem.pdf.
[31] Wikimedia. Strategy 2021: Wikibase ecosystem, 2021. URL: https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy2021/Wikibase.
[32] Lozana Rossenova, ArtBase Archive—Context and History: Discovery Phase and User Research 2017–2019. 2020. URL: https://lozanaross.github.io/phd-portfolio/docs/1_Report_ARTBASE-HISTORY_2020.pdf.
[33] Lozana Rossenova and Lucia Sohmen, Using OpenRefine with arbitrary Wikibase instances, in: WikidataCon 2021, online, 2021. URL: https://pretalx.com/wdcon21/talk/XDNW9A.
[34] Alexander Derveaux, Demo of upload process for a Wikibase instance, in: WikidataCon 2021, online, 2021. URL: https://pretalx.com/wdcon21/talk/LERPAG.
[35] Wikidata, Training, 2022. URL: https://www.wikidata.org/wiki/Wikidata:Training.
[36] Sandra Fauconnier, Dragan Espenschied, Lyndsey Moulds, Lozana Rossenova, Many Faces of Wikibase: Rhizome's Archive of Born-Digital Art and Digital Preservation, Wikimedia Blog (2018). URL: https://wikimediafoundation.org/news/2018/09/06/rhizome-wikibase/.
[37] Rhizome, Welcome to the ArtBase Query Service: Federation and Advanced Queries (2021). URL: https://artbase.rhizome.org/wiki/Query#Federation_and_Advanced_Queries.
[38] Dennis Diefenbach, Max de Wilde, Samantha Alipio, Wikibase as an Infrastructure for Knowledge Graphs: the EU Knowledge Graph. in: ISWC 2021, France, online, 2021. URL: https://hal.archives-ouvertes.fr/hal-03353225/document.
[39] CIDOC-CRM, Homepage, 2022. URL: https://www.cidoc-crm.org/.
[40] David Fichtmueller, Using Wikibase as a Platform to Develop a Semantic Biodiversity Standard. in: 1st NFDI Wikibase Workshop, online, 2021. URL: https://docs.google.com/presentation/d/1i91OB9xPZVVovd8c7Cm2sOglQLM8CEeZed8grdVaFwU/edit.
[41] Jose Emilio Labra Gayo, et al, Representing the Luxembourg Shared Authority File based on CIDOC-CRM in Wikibase. in: SWIB 2021, online, 2021. URL: https://swib.org/swib21/slides/05-03-gayo.pdf.
[42] Wikibase Stakeholders Group, Homepage, 2022. URL: https://wbstakeholder.group/.
[43] ProfessionalWiki Github, Wikibase RDF Extension, 2022. URL: https://github.com/ProfessionalWiki/WikibaseRDF.
[44] DigAMus Award, Homepage, 2022. URL: https://digamus-award.de
[45] Ina Blümel, Lucia Sohmen, Nils Casties, Integration of Wikidata 4OpenGLAM into data and information science curricula, in: WikidataCon 2021, online, 2021. URL: https://pretalx.com/wdcon21/talk/MK3ZBH/.
[46] DigAMus Award. DigADigAMus goes Wikidata, 2021. URL: https://digamus-award.de/2021/07/29/digamus-goes-wikidata/.
[47] OpenRefine, Homepage, 2022 https://openrefine.org/.
[48] Wikimedia Commons, Commons: Copyright rules, 2021. URL: https://commons.wikimedia.org/wiki/Commons:Copyright_rules.

[49] DigAMus Award, WikiProject Digital projects of museums, 2022. URL: https://www.wikidata.org/wiki/Wikidata:WikiProject_Digital_projects_of_museums.
[50] Ulrike Wuttke, Here be dragons: Open Access to Research Data in the Humanities (2019). URL: https://ulrikewuttke.wordpress.com/2019/04/09/open-data-humanities/.
[51] Erzsebet Tóth-Czifra, Ulrike Wuttke, Loners, Pathfinders, or Explorers? How are the Humanities Progressing in Open Science? (2019). doi: https://doi.org/10.25815/x516-wf23.
[52] Kompakkt, Homepage, 2022. URL: https://kompakkt.de/home.
[53] Bayerische Akademie der Wissenschaften, Corpus der barocken Deckenmalerei in Deutschland, 2021. URL: https://deckenmalerei.badw.de/.
[54] NFDI4Culture 3D Data Enrichment MVP. Data Model, 2022. URL: https://wikibase.semantic-kompakkt.de/wiki/Data_Model.
[55] Rob Sanderson, Paolo Ciccarese, Benjamin Young (Eds.), Web Annotation Data Model: W3C Recommendation 23 February 2017, 2017. URL: https://www.w3.org/TR/annotation-model/.
[56] The Getty Research Institute, Getty Vocabularies, URL: https://www.getty.edu/research/tools/vocabularies/.
[57] Iconclass, Homepage, 2022. URL: https://iconclass.org/.
[58] WMDE Github, Wikibase Release Pipeline, 2022. URL: https://github.com/wmde/wikibase-release-pipeline.
[59] Andra Waagmeester, in: Email conversation with the author Lozana Rossenova, March 14, 2022.