Getting digital preservation data out Wikidata

Katherine Thornton
Yale University Library USA
katherine.thornton@yale.edu
https://orcid.org/0000-0002-44990451

Kenneth Seals-Nutt
Yale University Library USA
kenneth.seals-nutt@yale.edu
https://orcid.org/0000-0002-5926-9245

The Wikidata knowledge base provides a public infrastructure for creating and syndicating machine-readable data about computing resources. We have prepared a set of queries that can be used to gather data sets relevant to digital preservation from Wikidata. We present these data sets in the context of the Wikidata for Digital Preservation portal (Wikidp). Wikidp is a free software portal that allows people to explore data related to digital preservation from the Wikidata knowledge base. Structured data about file formats, the many versions of software titles, and computing environments, are already available in Wikidata. The content of Wikidata is licensed under the Creative Commons Zero license, meaning that anyone can reuse the data for any purpose. The content in Wikidata is available in more than 300 human languages. The data in Wikidata is FAIR data, and it is linked open data. Our portal provides an interface designed for the needs of the digital preservation community.

Wikidata, digital preservation, linked open data Designing and Delivering Sustainable Digital Preservation; Building Capacity, Capability and Community

I. Introduction

Wikidata is the knowledge base that anyone can edit. Twenty thousand people edit Wikidata each month [1]. These editors add structured data in the form of statements of facts, and references for those statements, to the knowledge base. The Wikidata community has created more than five thousand properties for the knowledge base, and editors compose statements using these properties. Many people in the Wikidata community are personally interested in the domain of computing. Collectively, these editors have curated hundreds of thousands of statements related to software and hardware.

II. What is Wikidata?

Wikidata is a cross-domain knowledge base of structured data. Simply put, it is a database of facts that both humans and machines can edit and consume. Wikidata went live in late 2012 [2]. The infrastructure of Wikidata is collaboratively built via commons-based peer production [3]–[5]. Commons-based peer production is the name given to open collaboration systems where users are creating content under the agreement that all content will remain in the public domain. This means that all of the work products of the community are free to be reused by others. The peer-production aspect refers to how users coordinate work themselves. Wikidata is edited by volunteers from all over the world in more than 350 languages [6].

In addition to a free software infrastructure, the Wikidata community also publishes all content in the knowledge base under a Creative Commons Zero License. The Wikidata community makes dumps of previous versions of the content of the knowledge base available. The infrastructure of the Wikidata knowledge base is maintained by an international community of people. For cultural heritage institutions who find structured data in Wikidata relevant for their work flows, this means that there will be much less staff time necessary to design, build and maintain infrastructure for this data. For cultural heritage institutions with limited digital preservation budgets, this means that they can now access descriptive and technical metadata for tens of thousands software titles and more than three thousand file formats without having to create manage or maintain that data locally.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

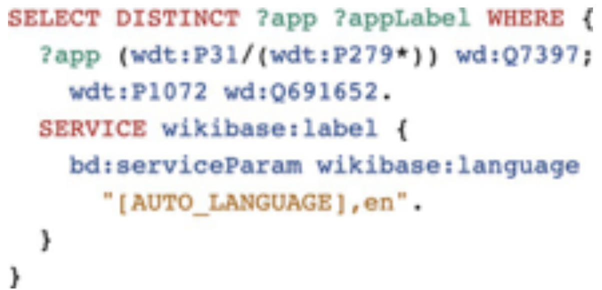
III. What data can I reuse from Wikidata?

The Wikidata community maintains a public SPARQL endpoint. As of October, 2017 the endpoint was consistently handling 8.5 million SPARQL queries per day [7]. Writing SPARQL queries for the Wikidata endpoint allows users to search for data about resources in a flexible way. SPARQL queries enable us to search for file formats by media type, or by file extension, etc. They allow users to search for software titles by their readable file formats, or to search for software titles published within specific windows of time. We can use SPARQL queries to search for software titles by genre, to search for technical specifications that describe a particular file format, or to search for a digitized copy of a user guide for a piece of legacy software. Because Wikidata is a cross-domain knowledge base, the range of data combinations allow users to query data that span technical metadata as well as descriptive metadata aspects of these resources

A. Getting Data from Wikidata

Humans can view data in Wikidata via any of the wiki pages. To access data in bulk, users can access the MediaWiki API[1] or the Wikidata Query Service[2]. Users of the Wikidata Query Service SPARQL endpoint can request subsets of the data contained in Wikidata that match specific patterns. Users can design queries that take advantage of different Wikidata properties such as the examples in the figures below.

1. Return all software titles known to read .dxf files, see Figure 1.

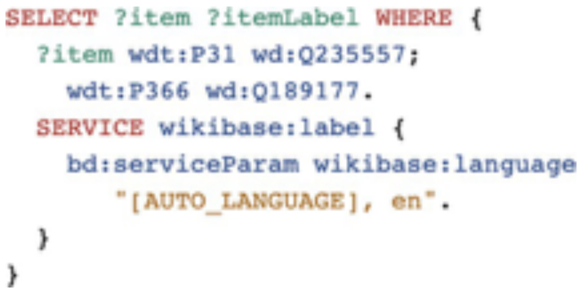


```
SELECT DISTINCT ?app ?appLabel WHERE {
  ?app (wdt:P31/(wdt:P279*)) wd:Q7397;
    wdt:P1072 wd:Q691652.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language
      "[AUTO_LANGUAGE],en".
  }
}
```

Figure 1: A screenshot of a SPARQL query to request software titles that can read .dxf files.

Try this query! Code for this query.

2. File formats used for 3D graphics, see Figure 2.

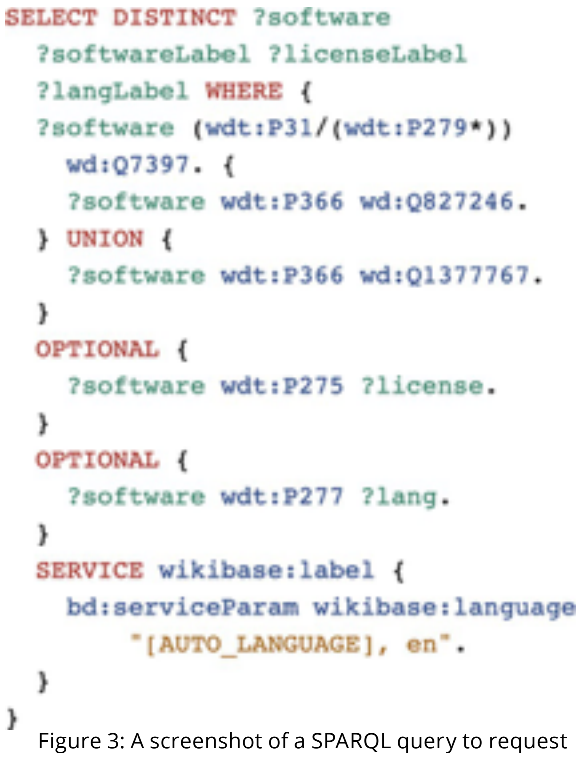


```
SELECT ?item ?itemLabel WHERE {
  ?item wdt:P31 wd:Q235557;
    wdt:P366 wd:Q189177.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language
      "[AUTO_LANGUAGE], en".
  }
}
```

Figure 2: A screenshot of a SPARQL query to request file formats used for 3D data.

Try this query! Code for this query.

3. Sequence alignment software with date of publication, programming language and license, see Figure 3.



```
SELECT DISTINCT ?software
  ?softwareLabel ?licenseLabel
  ?langLabel WHERE {
  ?software (wdt:P31/(wdt:P279*))
    wd:Q7397. {
    ?software wdt:P366 wd:Q827246.
  } UNION {
    ?software wdt:P366 wd:Q1377767.
  }
  OPTIONAL {
    ?software wdt:P275 ?license.
  }
  OPTIONAL {
    ?software wdt:P277 ?lang.
  }
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language
      "[AUTO_LANGUAGE], en".
  }
}
```

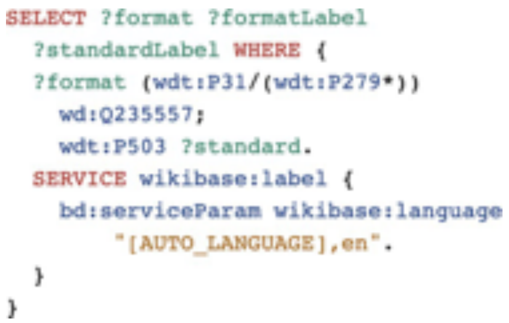
Figure 3: A screenshot of a SPARQL query to request sequence alignment software with date of publication, programming language and license.

Try this query! Code for this query.

[1] https://www.mediawiki.org/wiki/API:Main_page

[2] https://query.wikidata.org/

4. File formats to which the defining ISO standard been linked, see Figure 4.

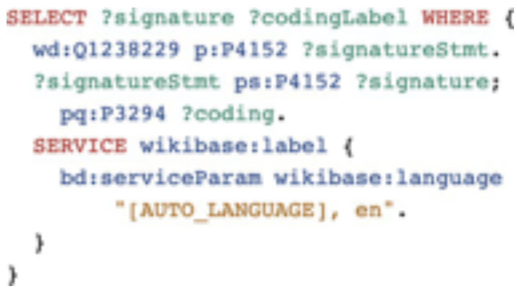


```
SELECT ?format ?formatLabel
  ?standardLabel WHERE {
  ?format (wdt:P31/(wdt:P279*))
    wd:Q235557;
    wdt:P503 ?standard.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language
        "[AUTO_LANGUAGE],en".
  }
}
```

Figure 4: A screenshot of a SPARQL query to request the a list of file formats described by ISO standards.

Try this query! Code for this query.

5. What is the signature of the SteroLithography file format, see Figure 5.



```
SELECT ?signature ?codingLabel WHERE {
  wd:Q1238229 p:P4152 ?signatureStmt.
  ?signatureStmt ps:P4152 ?signature;
    pq:P3294 ?coding.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language
      "[AUTO_LANGUAGE], en".
  }
}
```

Figure 5: A screenshot of a SPARQL query to request the file format signature of the .stl file format.

Try this query! Code for this query.

IV. WikiDP Reports

The Wikidata for Digital Preservation portal (WikiDP) is a specialized interface that provides a view of the data in Wikidata tailored to the domain of digital preservation[1]. The Wikidata for Digital Preservation portal can be used to access and download multiple datasets derived from Wikidata. Users will find reports on a dedicated portal page[2]. Each time a person runs a query, the result set is computed live, thus results for many queries will change over time. The purpose of this page is to gather queries that return datasets of potential interest to the digital preservation community.

Reusing these queries over time allows us to gain a deeper understanding of how the data changes over time.

The reports featured on the portal website are a subset of the total reports we track[3]. As we write additional queries this inventory of useful datasets will also grow.

V. Fair data

Long-term preservation and governance of metadata for the domain of computing is an important issue for the digital preservation community [8], [9]. Centralization of technical metadata for the domain of computing benefits all creators and users of this metadata. Wikidata is a multilingual knowledge base, leveraging the mappings created through years of conceptual alignment among the different language versions of Wikipedia and Wikidata items [10]. This means that more users will have access to metadata related to the domain of computing in their language, an important step in reducing the dominance of the English language which disadvantages other linguistic communities.

The data contributed to Wikidata is compliant with the FAIR data principles [11]. By creating data that aligns with the FAIR data principles, we ensure that this metadata is easy to find and easy to reuse. This technicalital preservation professionals must be able to identify and refer to, will be more complete if we distribute our effort. Redundant, fragmented descriptions in siloed repositories are frustratingly incomplete. Many governmental bodies and international consortia have endorsed the FAIR data principles as a key aspect of their open science or open data initiatives [12]. The data contributed to Wikidata is linked open data[4]. Experts from libraries, archives, museums and technologists of the World Wide Web Consortium (W3C) recommend linked data for library metadata published on the web[13].

[1] A description of the WikiDP system.

[2] www.wikidp.org/reports

[3] For a more complete list see https://github.com/emulatingkat/SPARQL

[4] https://www.w3.org/DesignIssues/LinkedData.html

FAIR is an acronym for findable, accessible, interoperable and reusable. Metadata for the domain of computing that we contribute to Wikidata are **findable** in that Wikidata items are indexed by all large search engines. The Qids assigned to Wikidata items are their unique, persistent identifiers.

These metadata are **accessible** because the entity data associated with their unique ids (all statements and references asserted about an item) are dereferencable via the HTTP protocol. They are **interoperable** in that they link to many other databases and systems through the collection of external ids as seen in Figure 7.

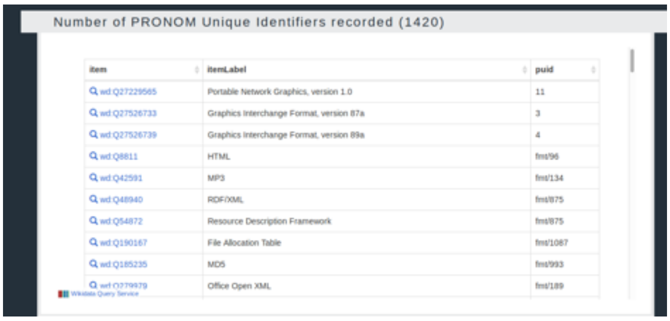
These metadata are **reusable** due to the use of the CCO license for the content of Wikidata. Anyone can reuse Wikidata data for any purpose. Publishing data in the Wikidata knowledge base fulfills the most complete degree of FAIRness, level F, "FAIR data, Open Access, Functionally Linked", as described in [12].

The Wikidata for Digital Preservation Portal provides direct links to items in Wikidata. If a user would like to consult Wikidata to view additional information related to vocabularies that have been stored, they may consult the item of interest by following the links provided in the Portal.

VI. Discussion: a centralized repository of FAIR, linked open data

Wikidata is growing. We have been participating in Wikidata by structuring data in the domain of computing since August, 2016. In the years of our participation we have seen growth in Wikidata as a whole, and improved data coverage for computing topics.

Wikidata is a project of Wikimedia Deutschland[1] and has been supported by the chapter budget, grant awards and donations. In 2016 the Wikimedia Foundation announced that it would begin funding the software engineering activity for Wikidata[2]. This is a strong signal that the infrastructure of Wikidata will continue to be supported in the future.



Number of PRONOM Unique Identifiers recorded (1420)

item	itemLabel	puid
wd:Q27229565	Portable Network Graphics, version 1.0	11
wd:Q27526733	Graphics Interchange Format, version 87a	3
wd:Q27526739	Graphics Interchange Format, version 89a	4
wd:Q8811	HTML	fmt/96
wd:Q42591	MP3	fmt/134
wd:Q48940	RDF/XML	fmt/875
wd:Q54872	Resource Description Framework	fmt/875
wd:Q190167	File Allocation Table	fmt/1087
wd:Q185235	MD5	fmt/993
wd:Q279979	Office Open XML	fmt/189

Figure 6: This is a screenshot of the Wikidp Reports page. When a user selects a report, a query is performed on WQS and the most recent result set is returned.

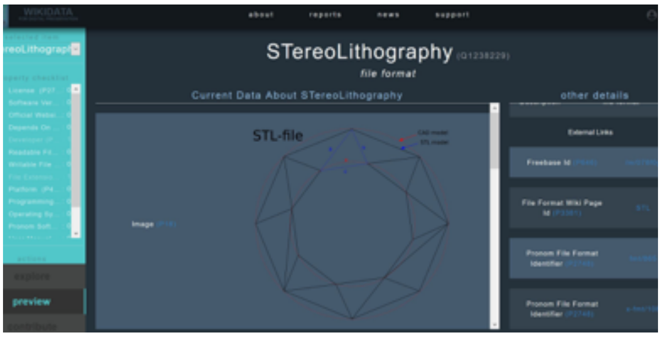


Figure 7: A screenshot of the Wikidata Item for the Sterolithography file format showing the collection of links to external resources that also describe the format on the right-hand side of the interface.

A. *Infrastructure and Maintenance*

The Wikidata for Digital Preservation portal can support the work of many users, and only one local developer [14] with knowledge of Wikidata is required. The local developer inspects the data models and the infrastructure of Wikidata in order to make recommendations about the user interface and the interaction design of the portal, effectively articulating work for many users.

We conclude that reusing the infrastructure of the Wikidata knowledge base, which has been assigned to the public domain, is a compelling model for cultural heritage institutions looking for a centralized repository for metadata. Commons-based peer production of infrastructure allows for distributed stakeholders to collaboratively maintain the infrastructure [4]. The wiki software is developed by engineers who have agreed to make their work available to all by releasing it under free software licenses

[1] https://wikimedia.de/wiki/Hauptseite

[2] https://blog.wikimedia.org/2016/10/04/supporting-the-future-of-wikidata/

The Wikidata community maintains a public SPARQL endpoint for the knowledge base. This SPARQL endpoint allows users to write flexible, powerful queries to retrieve subsets of the data in the knowledge base. In contrast, maintaining a public SPARQL endpoint is not often feasible for a software or format registry developed within an institution or project context. The Wikidata community has enhanced the SPARQL endpoint by providing multiple visualization options for the data returned in queries, from bubble charts to graphs of many varieties. The developers who work on the SPARQL endpoint have also created an interface that supports users who do not yet know SPARQL in writing or modifying SPARQL queries[1]

B. Collaboration

The structure of Wikidata allows the crowd to collaborate. A boundary object is a tool for thinking that allows people from different communities of practice to use a shared form to bridge the differences in their experiences and effectively collaborate [15]. When multiple boundary objects are used in conjunction they can become parts of systems of boundary objects [16]. Star and Bowker introduced the concept of "boundary infrastructure" to theorize about systems of boundary objects. Boundary infrastructure allows for collaboration without consensus [16]. Wikidata, the knowledge base of structured data that anyone can edit, is an example of boundary infrastructure that allows people from many communities of practice, from many walks of life, specialists and non-specialists across many domains, to effectively collaborate to structure data, and make it available for reuse.

C. Sustainability

Digital Preservation is a expensive activity for many institutions. Institutions with limited budgets for digital preservation can reuse this data at no cost. The boundary infrastructure of Wikidata provides a means for digital preservation professionals from different parts of the world, working in different languages, to collaborate by creating structured data in the knowledge base. This reduces the risk of redundant effort to describe the same file format in numerous local format registries. The boundary infrastructure of the knowledge base also supports contributions from the crowd, people who have interest in, and information about, the domain of computing. This allows for collaborations that otherwise might not happen without the boundary infrastructure that facilitates communication in a community of practice.

Members of the general public will also have access to this information. Having this information in an accessible, structured repository will allow more people to consult it, which could lead to people making different computing choices in their lives, for example choosing an open format, which could impact the work of future generations of digital preservation professionals. Wikidata's CC0 license ensures that this data will have an equalizing force, as it will not be controlled by any single institution, or even any consortium of institutions. Anyone with access to the internet will be able to inspect and reuse this data for their own projects or systems. Institutions that do not yet have budgets for digital preservation will have access to this metadata and will not have to recreate it in their local systems.

VII. Conclusion

The Wikidata for Digital Preservation Portal facilitates increased communication between members of the Wikidata community and the international digital preservation community.

Centralizing the metadata for the domain of computing, eliminates redundant labor of individual institutions creating structured data within their local systems. When we collaboratively create metadata and publish it in Wikidata, anyone can reuse it. This allows metadata professionals to focus on the administrative, preservation, and use metadata pertinent to their local settings.

Making use of infrastructure supported by the Wikimedia foundation, built and maintained by an active community of tens of thousands of contributors is a new option for cultural heritage institutions. The fact that this infrastructure is built in conformance to open standards and is comprised of free software means that we can audit this system and to

[1] https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service

see if we can continue to trust it to store and access our data.

Acknowledgment

This portal was inspired by the Su Lab and their Wikigenomes portal. We thank Wikimedia Deutschland[1] for supporting the Wikidata project. We thank the Wikidata community for sharing thoughts, feedback, data curation and data modeling work. Specifically, would like to thank all participants of WikiProject Informatics[2] for engaging with us as we work to describe the domain of digital preservation in Wikidata. To review the source code for this project please visit our Github repository[3]. We appreciate the informative conversations we had with participants of Wikicite 2016, Wikicite 2017. This work is supported by the Council on Library and Information Resources, Andrew W. Mellon Foundation, Alfred P. Sloan Foundation, and the Open Preservation Foundation.

References

[1] Wikidata, Statistics, 2019. [Online]. Available: https://www.wikidata.org/wiki/Special : Statistics.

[2] D. Vrandečić, “Wikidata: A new platform for collaborative data collection,” in Proceedings of the 21st International Conference Companion on World Wide Web, ACM, 2012, pp. 1063–1064.

[3] Y. Benkler, “Coase’s penguin, or, linux and the nature of the firm,” Yale Law Journal, pp. 369–446, 2002.

[4] Y. Benkler, A. Shaw, and B. M. Hill, “Peer production: A modality of collective intelligence,” Collective Intelligence, 2013.

[5] C. Müller-Birn, B. Karran, J. Lehmann, and M. Luczak-Rösch, “Peer-production system or collaborative ontology engineering effort: What is wikidata?” In Proceedings of the 11th International Symposium on Open Collaboration, ACM, 2015, p. 20.

[6] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić, “Introducing wikidata to the linked data web,” in The Semantic Web–ISWC 2014, Springer, 2014, pp. 50–65.

[7] L. Pintscher, “State of the project,” WikidataCon 2017, https://upload.wikimedia.org/wikipedia/commons/b/b0/WikidataCon_2017-_State_of_the_Project.pdf, year=2017.

[8] P. McKinney, S. Knight, J. Gattuso, D. Pearson, L. Coufal, D. Anderson, J. Delve, K. De Vorsey, R. Spencer, and J. Hutař, “Reimagining the format model: Introducing the work of the nsla digital preservation technical registry,” New Review of Information Networking, vol. 19, no. 2, pp. 96–123, 2014.

[9] P. McKinney, D. Pearson, D. Anderson, J. Hutař, S. Knight, L. Coufal, J. Delve, J. Gattuso, K. DeVorsey, and R. Spencer, “A next generation technical registry: Moving practice forward,” iPRES 2014: 11th International Conference on Digital Preservation, 2014.

[10] S. Burgstaller-Muehlbacher, A. Waagmeester, E. Mitraka, J. Turner, T. Putman, J. Leong, C. Naik, P. Pavlidis, L. Schriml, B. M. Good, et al., “Wikidata as a semantic framework for the gene wiki initiative,” Database, vol. 2016, baw015, 2016. https://www.wikidata.org/wiki/Wikidata:WikiProject_Informatics12 https://github.com/WikiDP

[11] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg,G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., “The fair guiding principles for scientific data management and stewardship,” Scientiic data, vol. 3, p. 160 018, 2016.

[12] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson, “Cloudy, increasingly fair; revisiting the fair data guiding principles for the european open science cloud,” Information Services & Use, no. Preprint, pp. 1–8, 2017.

[13] T. Baker, E. Bermès, K. Coyle, G. Dunsire, A. Isaac, P. Murray, and M Zeng, “Library linked data incubator group final report,” report, W3C Incubator Group, October, vol. 25, 2011.

[14] M. Gantt and B. A. Nardi, “Gardeners and gurus: Patterns of cooperation among cad users,” in Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, 1992, pp. 107–117.

[15] S. L. Star and J. R. Griesemer, “Institutional ecology,translations’ and boundary objects: Amateurs and professionals in berkeley’s museum of vertebrate zoology, 1907-39,” Social studies of science, vol. 19, no. 3, pp. 387–420, 1989.

[16] S. L. Star, “This is not a boundary object: Reflections on the origin of a concept,” Science, Technology and Human Values, vol. 35, no. 5, 601–617, 2010.

[1] https://www.wikimedia.de/wiki/Hauptseite

[2] https://www.wikidata.org/wiki/Wikidata:WikiProject_Informatics

[3] https://github.com/WikiDP