Modeling the Domain of Digital Preservation in Wikidata

Katherine Thornton[*]
Euan Cochrane
katherine.thornton@yale.edu
euan.cochrane@yale.edu
Yale University

Thomas Ledoux
Bertrand Caron
Bibliothèque nationale de France
Paris, France

Carl Wilson
Open Preservation Foundation
UK

ABSTRACT

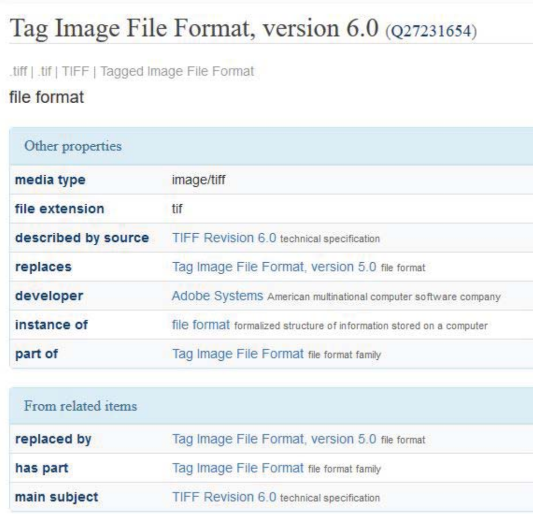
Members of the digital preservation community collate and capture metadata to describe file formats, software, operating systems and hardware, and use it to inform and drive digital preservation processes. In this work we describe how the infrastructure of Wikidata meets the requirements for a technical registry of metadata related to computer software and computing environments. Collaboratively creating this metadata, and making it available as linked open data, will reduce the amount of redundant work digital preservation professionals do in order to describe resources. Having machine-readable, linked open data that describes the digital preservation domain will also allow us to reuse this data in our software applications and information systems, reducing the overhead when building new tools. Furthermore the Wikidata social and technical infrastructure will enable the long term continued access to the data digital preservation practitioners collate and capture.

Wikidata is a project of the Wikimedia Foundation (WMF), and is created through commons-based peer production [3]. Simply put, Wikidata is a knowledge base of structured data that anyone can edit [40]. The infrastructure of Wikidata is created using free software, and is designated to the public domain. All content in Wikidata is licensed so that other may freely reuse the data. Volunteer editors, coordinating their own work, add data to Wikidata. Through this analysis we demonstrate how the infrastructure of Wikidata provides distinct advantages to the cultural heritage domain that proprietary knowledge bases do not provide.

ACM Reference format:
Katherine Thornton, Euan Cochrane, Thomas Ledoux, Bertrand Caron, and Carl Wilson. 2017. Modeling the Domain of Digital Preservation in Wikidata. In *Proceedings of ACM International Conference on Digital Preservation, Kyoto, Japan, September 2017 (iPres'17),* 10 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Metadata about software, file formats and computing environments is necessary for the identification and management of these entities. Machine-readable metadata about software, file formats and computing environments allows digital preservation practitioners to then automate programmatic interactions with these entities.



Tag Image File Format, version 6.0 (Q27231654)

.tiff | .tif | TIFF | Tagged Image File Format

file format

Other properties	
media type	image/tiff
file extension	tif
described by source	TIFF Revision 6.0 technical specification
replaces	Tag Image File Format, version 5.0 file format
developer	Adobe Systems American multinational computer software company
instance of	file format formalized structure of information stored on a computer
part of	Tag Image File Format file format family

From related items	
replaced by	Tag Image File Format, version 5.0 file format
has part	Tag Image File Format file format family
main subject	TIFF Revision 6.0 technical specification

Figure 1: A screenshot of the Wikidata item for TIFF version 6.0 and related properties.

For example, if we ingest a file into the a digital preservation system, we could configure the system to use a tool such as DROID [2] or Sigfreid [20] to automatically identify the file formats of resources at the time of ingest into the digital preservation system. Based on that data, we could then present the user with a list of available software in a collection that could be used to interact with files of that type. This is an example of machine-actionable interaction that is possible with a digital preservation system using DROID in combination with Wikidata data to determine the software options.

A technical registry [6] is a data store of descriptions of file formats, software used to create or interact with files, configured hardware environments, operating systems and sustainability factors. In Figure 1, there is a screenshot of a Wikidata item and the properties used to describe the item. The metadata for file formats in Wikidata includes information about who developed the format, how it is related to other formats, what the file extensions are, and the media type(s).

2 FRAGMENTED AND INCOMPLETE

Many researchers and practitioners in the field of digital preservation have identified the fragmented nature of technical registries [21]. When registry data is stored in multiple systems, communication also takes place in distributed settings. In response to the fragmented landscape, several groups have presented plans for how

[*] The corresponding author

iPres'17, Kyoto, Japan
© 2017 Copyright held by the owner/author(s). 123-4567-24-567/08/06…$15.00
DOI: 10.475/123_4

to centralize and unify technical registry information [22, 37, 38]. Both GDFR and UDFR have concluded work. The sustainability of a technical registry is important set of factors to consider. The creation of a sustainability plan for the infrastructure of the registry is a crucial component to consider. The solution proposed by the New Zealand team is forthcoming, and infrastructural development described as the subsequent phase of work [22]. In this paper, we explore an alternative to creating infrastructure for a repository of technical registry data through use of existing, independently supported, infrastructure provided by the knowledge base of structured data, Wikidata.

Wikidata went live in late 2012 [39]. The infrastructure of Wikidata is collaboratively built via commons-based peer production [3, 4, 25]. Commons-based peer production is the name given to open collaboration systems where users are creating content that will become part of 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 amongst themselves. Wikidata is edited by volunteers from all over the world in more than 350 languages [14].

The MediaWiki software [1] and WikiBase software [41] are the primary technical components of the knowledge base itself. In the domain of computational systems, the concept of infrastructure is used to describe technologies that support information systems. Theorists of infrastructure, Star and Ruhleder, note that infrastructure is often invisible, and because of this, many people take it for granted [35]. By referring to infrastructure as 'invisible' these authors highlight the fact that infrastructure is often purposely designed to be available only to those who are building or repairing it. For example, the infrastructure of the search algorithms used by Google are not made visible to users of the search engine. The database structure of Amazon.com is not made visible to visitors of the website. In contrast, Wikidata's infrastructure is open for inspection because the technical components of the system are free software [33] and the source code for the software is shared publicly. Using Wikidata as the technical registry of metadata for the domain of digital preservation equates to using the infrastructure of the Wikidata system to store digital preservation metadata.

3 WIKIDATA: A KNOWLEDGE BASE OF STRUCTURED DATA

The knowledge base of structured data, Wikidata, combines a data model with structured data. Editors add content and provide source information for structured data [17]. Wikidata contains data about entities structured in a way that is machine-readable as well as human-readable [14]. As the data management platform of all Wikimedia Foundation projects, the data is free and open for reuse within all WMF projects, and also is freely available for reuse outside of WMF projects [14].

In Figure 2, we see a screenshot of the Wikidata page for entity Q42332 *Portable Document Format.* Each item is allotted a page in Wikidata and has a unique identifier, with prefix *Q* plus a string of numbers, ex. *Q42332*, which is assigned to the item *Portable Document Format.* Wikidata consists of two entity types: **items** and **properties**. In Figure 2 the four properties *Commons category*, *BNCF Thesaurus*, *image*, and *topic's main category* are used to assert statements about the item *Q114678.* In Figure 2 we see four statements about *Portable Document Format.* Each of these statements expressed a property of the item Q114678[1]. Each item contains a list of claims in the form of triples. The subject of the triple is the Wikidata item to which the claim refers, the predicate is a Wikidata property, and the object is a date, string, quantity, URL, an external identifier or another Wikidata item. Claims can be made more precise through the use of qualifiers. These qualifiers indicate the contexts in which the claim is valid. Claims can be annotated through the inclusion of references. A claim and its references are considered to be a *statement.* A description of the Wikidata data model can be found in [42].

The knowledge base, Wikidata, contains many millions of such items, many of these items are already described by a set of statements. The Wikidata pages for items allow users to view and enter data [14]. These pages are open for editing by anonymous IP, or by registering for an account as an editor. Volunteer editors contribute to the project in coordination with one another and all content in the system is designated as belonging within the public domain. The content in Wikidata spans general and specialized domains, and is relevant for many application areas [14].

3.1 Collaboratively creating data models

One way that the structure of Wikidata is extended is through the creation of properties to represent specific relationships between items. We participated in efforts to extend Wikidata to more precisely represent information from the domain of digital preservation. We identified pages on the Wikidata wiki where the community was actively describing computing. We found a active group engaged in Wikidata's WikiProject Informatics [2]. On the pages of this wiki project we communicated with other Wikidatans to discuss the models for representation of file formats, software, computing environments and hardware environments. We created items and proposed properties to describe different aspects of computing. We gave feedback on ideas proposed by other editors, and we gathered feedback about our ideas from other editors. Much of the discussion within WikiProject Informatics is based around the creation of new items and new properties for Wikidata. Items can be created by any editor of Wikidata. Editors use the *create new item* link that is displayed on the sidebar of every Wikidata page, and the newly-created item will be added to the knowledge base[3]. Additional steps are involved in the creation of new properties. Properties must first be proposed using a specific template. Once a property proposal template is applied to the property proposals page, a discussion is opened and other editors are invited to comment on the proposal. Property proposal discussions stay open for a minimum of seven days. Editors vote in support of or opposition to the property, and once there is sufficient support an editor with the Property Creator[4] privilege will create the new property and it will become part of Wikidata.

An overview of the properties that are currently in use in Wikidata with relevance for software is provided in Table 1.

[1] Each of these properties has a identifier assigned to it. The property identifiers all begin with prefix *P* plus a string of numbers.

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

[3] https://www.wikidata.org/wiki/Special:NewItem

[4] https://www.wikidata.org/wiki/Wikidata:Property$_{c}reators$

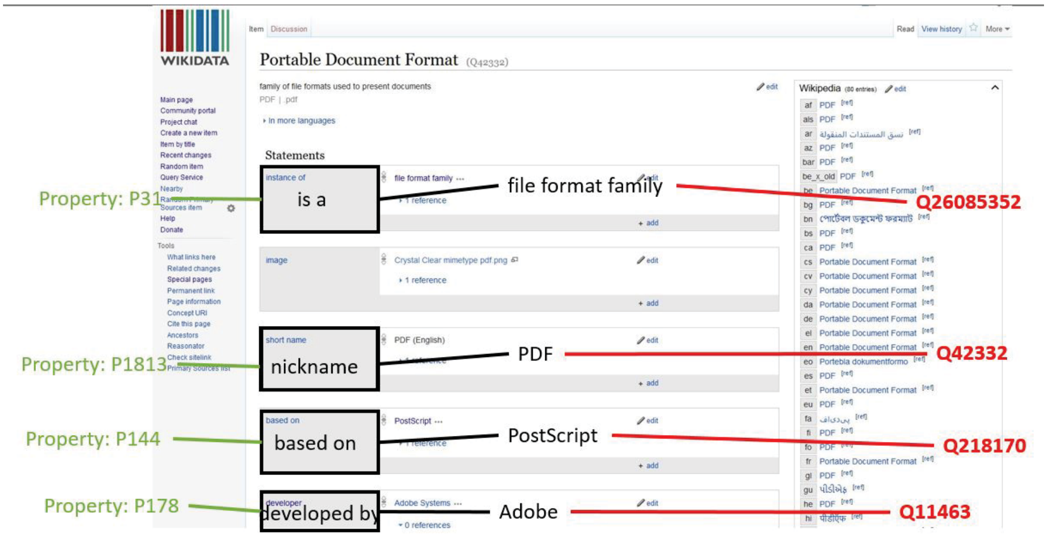


Figure 2: Wikidata item and data organization.

Wikidata Property	Data Type	Property ID
license	item	P275
software version	string	P348
official website	string	P856
depends on software	item	P1547
developer	item	P178
readable file format	item	P1072
writable file format	item	P1073
file extension	string	P1195
platform	item	P400
programming language	item	P277
operating system	item	P306
PRONOM id	external id	P2749
user manual link	string	P2078
bug tracking system	string	P1401
image	Commons Image File	P18
logo image	Commons Image File	P154

Table 1: Wikidata Properties for software

Wikidata Property	Data Type	Property ID
license	item	P275
software version	string	P348
official website	string	P856
developer	item	P178
PRONOM id	external id	P2748
LoCFDD id	external id	P3266
file extension	string	P1195
media type	string	P1163
replaces	item	P1365
replaced by	item	P1366

Table 2: Wikidata Properties for file formats

The most common property used to connect software items to file format items is the use of the properties P1072 *readable file format* and P1073 *writable file format*. This allows the relationship between the software and the file format to be machine-readable. Using structured data from Wikidata in an computational system for digital preservation allows us to use this data in a way that helps the computer program determine what the next step in a workflow (such as presenting a list of software choices to a user based on the user's file format) should be. Statements using P1072 and P1073 are made on the items related to software. In Table 2 the properties related to file format items are listed.

These tables list only a subset of relevant properties. For more exhaustive lists of useful properties, examples of how they are used, please see the wiki project[5]. There are many properties that we need to be able to extend the data model for the domain of digital preservation. For example, if we consider how the PREMIS project has modeled their semantic unit *environment*, we can see that this concept is not yet represented as a Wikidata property [12]. People who have expertise in the metadata issues and data modeling issues for the domain of digital preservation who also participate in Wikidata are in a strong position to advocate for data models in Wikidata that are compatible with previous models of the domain such as PREMIS [11] and PRONOM.

Data modeling in Wikidata happens through the use of properties and the use of class-subclass relationships [5, 32]. Models for domains within Wikidata are being extended as the project matures. For example, the modeling of the molecular biology domain takes place within Wikidata's WikiProject for molecular biology [14, 23].Collaboratively modeling the a domain requires interaction and communication among participants. Muller-Birn et al. found that the Wikidata interface is a barrier for contributors who'd like to

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

engage in conceptual modeling, but that it supports instance-level curation more adequately [25].

We participated in collective data modeling activities through WikiProject Informatics[6]. WikiProjects are sets of wiki pages that groups use to perform focused, collaborative work [44]. One avenue through which we participated in the Wikidata community is the property proposal process. The property proposal process involves using the property proposal template to describe what you would like to add to Wikidata. Figure 3 is a screenshot of a property proposal template from Wikidata.

The template provides the outline structure for the proposal. To complete property proposal, an editor must think through the planned data type, how this property will be combined with other entities in Wikidata, and some examples for others to review.

Once a property has been proposed, other Wikidata editors consider the proposal and provide comments or ask questions about the proposal. The discussion portion of this process helps ensure that multiple editors beyond the editor who created the property proposal reviews the proposal. Reviewers are looking out for redundant properties, duplicate properties, properties that would conflict with other properties, and ensure that all parts are included (such as a formatter URL if the proposed property is an external id). As of March, 2017, there are 3228 properties available for use in Wikidata[7]. Dozens of properties are created each week[8].

Focused engagement with the property proposal workflow in Wikidata allows participants to take active roles in the data modeling activities. The negotiation that occurs among editors through the property proposal workflow helps ensure that properties are designed efficiently.

3.2 Our Wikidata work to date

Members of our team have been proposing properties for external IDs from other databases. For example, we proposed and saw created, P2748 *PRONOM file format identifier*, P2749 *PRONOM software identifier*, and P3266 *LoC FDD ID.* These three properties each have a data type of external id, the property is constructed using the base URL for the external resource. For example, the formatter URL for P3266 *LoC FDD ID* is shown in Figure 4 [9].

We also worked to use these properties to create claims around items representing file formats. By undertaking data curation of the Wikidata entities within the domain of digital preservation, we discovered ways to use properties to model relationships between file formats and other entities. If we are not yet able to express a certain relationship, we make note of properties that do not yet exist which we would like to propose. For example, we are considering proposing a property to express the concept "part of configured software environment" to bring together sets of software used together for a purpose, such as presenting a particular emulated computing environment via a framework such as Emulation as a Service[10]. A property such as "part of configured software environment" could then be used to connect items for a base operating system, software that has been installed, etc. The data could then be reused in a system where that information is displayed to a user who needs information about what functionality can be expected from a particular configured environment.

Understanding the data model is very important when attempting to get data out of Wikidata. There are several options for how to get data out of Wikidata. It is possible to get structured data on a per item basis in a selection of data formats[11]. It is also possible to make a copy of a dump of the data. Another option is to request data through Wikidata's application programming interface[12] (API). It is also possible to request data from the Wikidata Query Service, a SPARQL endpoint. The name SPARQL is a recursive acronym for SPARQL Protocol and RDF Query Language [13]. SPARQL queries RDF triple stores in order to identify and return sets of triples that meet the criteria specified in the query. The fact that Wikidata maintains a SPARQL endpoint allows for powerful, flexible queries to be written to get data out of Wikdiata [8, 16, 23]. In the following section we will provide some examples of how the way the data is modeled has implications for how we can construct queries to get data out of Wikidata.

4 OPEN MODELING CHALLENGES

Through participation in the collaborative data modeling activities in Wikidata, we have observed several open questions that are relevant for the community in general. We consider these to be open questions because we have observed different segments of the community responding to these challenges using a variety of strategies, but agreement about how to harmonize the strategies is not yet resolved. We will illustrate these discussions with examples from Wikidata and examples of queries written for the Wikidata Query Service.

4.1 The Bonny and Clyde problem

One issue that arises when trying to use Wikidata as a format registry comes from its inception: originally the Wikidata entities were derived from Wikipedia pages. At the beginning of the Wikidata project there existed simple matches between pages on Wikipedias and Wikidata. Wikidata served as a hub to allow the alignment between the different language versions of a given page. Nearly all Wikipeida articles in all language versions have associated Wikidata items [14]. The notability criteria [13] indicate that this match (having at least one valid sitelink) as the major acceptability factor for the existence of a item. However, from a format registry perspective, some notions need to be separated. The main examples come from the different compression algorithms (gzip, bzip2, etc.) for which there is usually a single Wikipedia page describing not only the file format, but also the algorithm and a reference implementation. This leads to a single entity with multiple meanings. As an example, the bzip2 entity (Q273563) was described [14] in 2014 as an instance of file format and free software, and properly linked to 19 different sitelinks. Currently only two possibilities exist to bypass this limitation: (1) either create new entities to reflect the different natures. In the previous example, this implies adding the

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

[7] https //www.wikidata.org/wiki/Special ListProperties/

[8] For a complete list of Wikidata properties visit: https //www.wikidata.org/wiki/Wikidata:List_of_properties

[9] https //www.wikidata.org/wiki/Property P3266

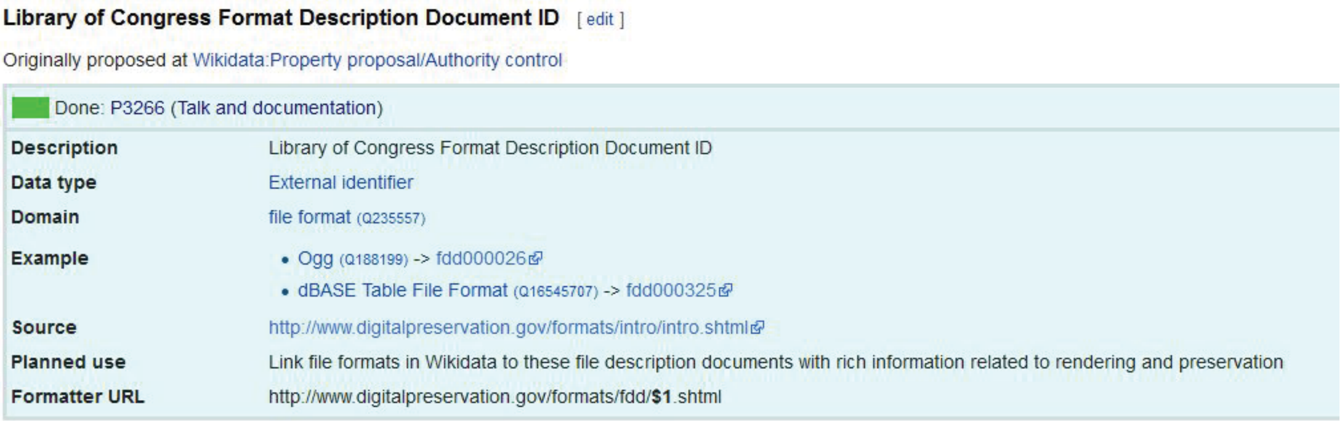
[10] http://bw-fla.uni-freiburg.de/

[11] https://www.wikidata.org/wiki/Wikidata Data_access

[12] https://www.wikidata.org/w/api.php

[13] https://www.wikidata.org/wiki/Wikidata Notability

[14] https://www.wikidata.org/w/index.php?title=Q283563&direbctionn̄ext&oldid1̄84272967



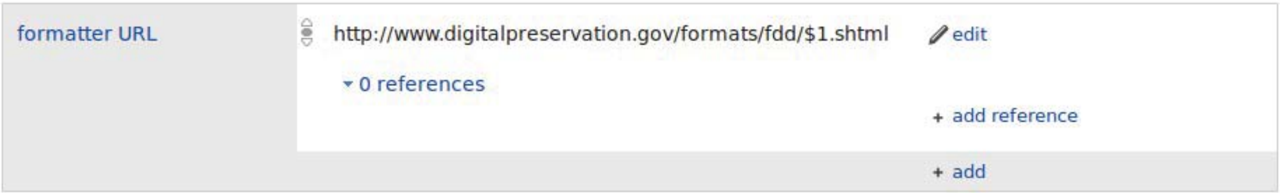
Library of Congress Format Description Document ID [edit]

Originally proposed at Wikidata:Property proposal/Authority control

🟩 Done: P3266 (Talk and documentation)

Description	Library of Congress Format Description Document ID
Data type	External identifier
Domain	file format (Q235557)
Example	• Ogg (Q188199) -> fdd000026 • dBASE Table File Format (Q16545707) -> fdd000325
Source	http://www.digitalpreservation.gov/formats/intro/intro.shtml
Planned use	Link file formats in Wikidata to these file description documents with rich information related to rendering and preservation
Formatter URL	http://www.digitalpreservation.gov/formats/fdd/**$1**.shtml

Figure 3: A complete property proposal template on Wikidata



formatter URL

http://www.digitalpreservation.gov/formats/fdd/$1.shtml

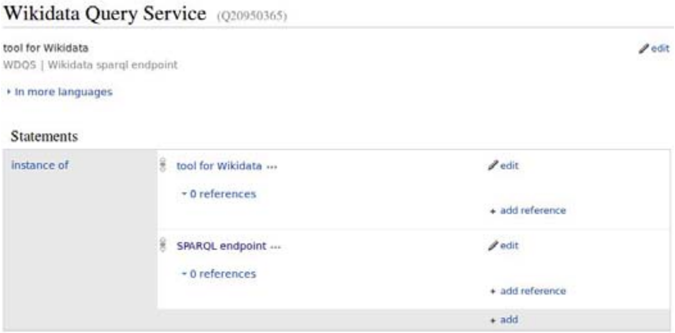
Figure 4: The formatter URL is built into all properties with the data type of external id.

"bzip2 Archive" (Q27866052) entity linked to the software entity by "readable/writable file format"(P1072 and P1073) properties. But this breaks the direct access to the Wikipedia pages describing the file format and goes against the notability criteria which may lead to a future removal of the entity ; or (2) take advantage of the multicultural aspect of Wikipedia and find one variation that has already separated the concepts to initiate the right model. An example of such an occurrence is the GZIP program where the Portuguese Wikipedia has a specific GZ file format page leading to the differentiation of the tool (Q283647) and the file format (Q10287816). There remains the issue of generalizing such distinction in the different languages but the "Article Placeholder" plugin might solve this [18].

In any case, in order to extend the model and to be able to describe more precise information, the issue of enhancing the relationships between the entity and the interlinks requires a solution. This relationship is important because the Wikipedia pages are the main way to expose the structured information, primarily through the use of information boxes. The Wikipedia page is often the best way to gather corrections and additions from Wikipedians. Even though no satisfying solution exists currently, the problem is well known as the "Bonnie and Clyde issue" [15] and various attempts are made which should lead to an option that will meet global agreement.

4.2 Properties refined by qualifiers

Properties are used in Wikidata to represent some attribute of an item to form a statement, as depicted in Figure 2. Certain types of statements can be expressed by different properties in combination with a one or more qualifiers. In Figure 5 we see a screenshot of some of the claims for item Q20950365, *Wikidata Query Service*. We can see that Property 31 *instance of* is used to claim that Wikidata Query Service is an instance of a SPARQL endpoint.



Wikidata Query Service (Q20950365)

tool for Wikidata

WDQS | Wikidata sparql endpoint

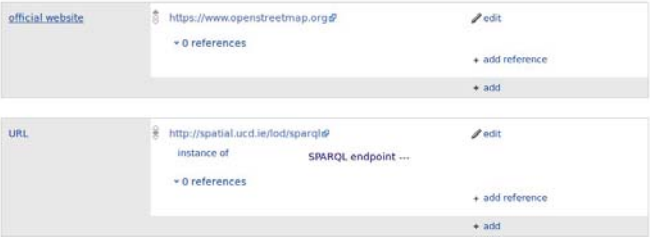
Statements

instance of: tool for Wikidata; SPARQL endpoint

Figure 5: The Wikidata item for SPARQL Wikidata Query Service

In contrast, Figure 6 is a screenshot of some of the claims for item Q936 *Open Street Map*. For this item a claim using Property 2699 *URL* is used to indicate the web location of OSM's SPARQL endpoint. This claim is refined through the addition of a Property 31 *instance of* qualifier to indicate that this URL is the URL for OSM's SPARQL endpoint. As we see from the differences between

[15] https://www.wikidata.org/wiki/Help:Handling_sitelinks_overlapping_multiple_items



official website
https://www.openstreetmap.org
edit
0 references
+ add reference
+ add

URL
http://spatial.ucd.ie/lod/sparql
edit
instance of SPARQL endpoint
0 references
+ add reference
+ add

Figure 6: The Wikidata item for Open Street Map

how the claims have been structured, the use of qualifiers to refine properties is a source of flexibility in how data can be modeled in Wikidata.

Some members of the Wikidata community hold the opinion that P31 should not be used as a qualifier and have stated their rationales on the discussion page for the property[16]. In situations where different subsets of the community decided to model data in different ways there is often a period of time where both strategies are observed and then a larger discussion is held to determine if one strategy is privileged over another. The discussions take place in the property talk namespace and are archived alongside the property proposals so that they may be revisited or reconsidered at a later point, if needed.

This flexibility in how data can be modeled impacts how data can be queried from Wikidata. For example, a query[17] for all items that are described as P31 *instances of* Q2621192 *SPARQL endpoint* will return a single result[18], that of item Q20950365 *Wikidata Query Service.* In contrast, if we structure the query differently[19] to ask for all items where P31 *instances of* Q2621192 *SPARQL endpoint* is used as a qualifier, we get 22 results.

A third option for how to structure the query[20] is a request for any item in which there is a P31 statement in a predicate position, thus returning all 23 results. While it is the case that this third query allows us to gather all 23 results, it highlights the necessity of deep familiarity with the data model, and how different portions of the community are working, in order to write a SPARQL query that will not miss relevant data.

The current state of modeling in Wikidata is that many editors are working independently and people are experimenting with many different options for how domains might be modeled. We have noticed a need for additional tools that could promote awareness [31] among editors of how sub-communities (often affiliated through WikiProjects) are currently working with data models for their domains. An example of such a tool is the navigation box shown in Figure 7. This navigation box was created by members of WikiProject Source Metadata[21] to provide an overview of the properties most relevant to the work of the sub-community. This navigation box provides a glanceable display [9] that anyone wanting to quick gain familiarity with the subset of properties that will be of use while curating data related to bibliographic metadata. Rather than scroll through the list of nearly 4,000 properties currently in use, users can quickly find a short list presented via the navigation box. Many WikiProjects use tables or navigation boxes to present relevant properties, however not all users visit WikiProjects regularly and may not encounter these lists if they are not already familiar with work practices of wikis [24].

Another promising strategy to address consistent data modeling for a given domain is the creation of a portal on top of Wikidata, such as that of the WikiGenomes project [30]. The WikiGenomes portal provides the relevant set of Wikidata entities so that basic researchers who would like to engage in data curation can interact with Wikidata via the portal interface. The strategy of creating a domain-specific portal is one approach to increasing the consistency of how data is being structured as the portal system design can guide users to structure data according to the current set of best practices as understood by the developers of the WikiGenomes portal.

The open modeling challenges, such as the "Bonny and Clyde" problem, involve mappings between Wikipedias and Wikidata, the divergence in the use of properties or properties mixed with qualifiers, and the ongoing data modeling to extend Wikidata properties in additional domains will require creative solutions. Certain sub-communities are addressing these issues through the development of domain-specific portals which may encourage more consistency in how statements are constructed, at least within the domain. Those interested in contributing to discussions of how else we might address these open challenges are invited to join the Wikidata community.

Contributing to a project in which not only the content (structured data), but also the enabling software (MediaWiki and Wikibase), are open to inspection, revision, or extension. The fact that free software [34] is used to enable this commons-based peer production system means that the infrastructure of Wikidata itself is subject to discussion and improvement. Some members of the Wikidata community actively extend the infrastructure by creating Wikidata-specific software tools that can be used to interact with the system[22]

4.3 Querying the knowledge base

Mechanisms to get data out of Wikidata are an important component of the infrastructure of Wikidata. The Wikidata Query Service is not only a SPARQL endpoint, but also a set of tools that can generate different visualizations of the data such as timelines, bubblecharts, and maps [26].

Wikidata uses a distinct data type for properties for external identifiers to other systems or repositories[23]. Data types indicate the kind of values that are appropriate to use in combination with the property [14]. As external identifiers are added to Wikidata, we can begin to see a crosswalk of authorities emerge within the knowledge base. For example, in the domain of digital preservation many people are familiar with PRONOM IDs, known as PUIDs, for

[16] See https //www.wikidata.org/wiki/Property_talk P31#P31_as_qualifier

[17] Query url http://tinyurl.com/kw4g39r

[18] The examples described in this section performed the way described here as of March, 2017. Due to the fact that the Wikidata knowledge base is edited by users it could change at any point in the future.

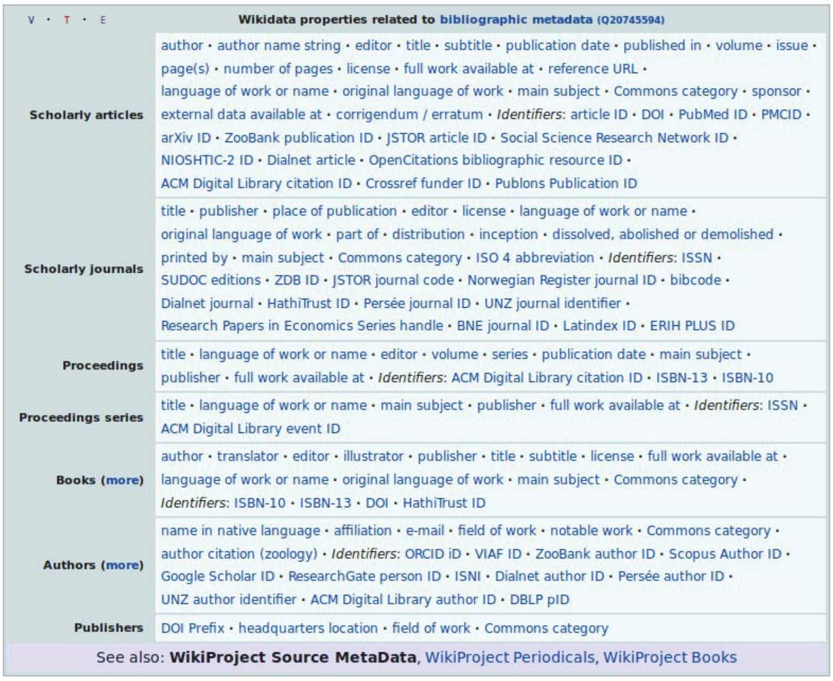
[19] See the alternative formulation of the query at http://tinyurl.com/godxqkk

[20] See a third option for how to structure the query at http //tinyurl.com/lautbhp

[21] https://www.wikidata.org/wiki/Wikidata:WikiProject_Source_MetaData

[22] https //www.wikidata.org/wiki/Wikidata Tools

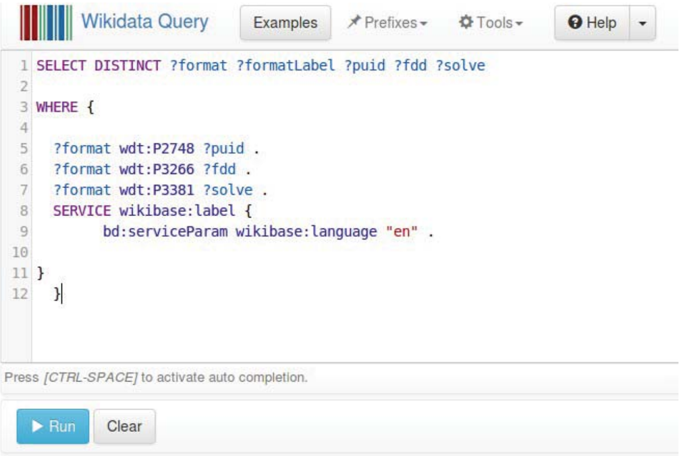
[23] https://www.wikidata.org/wiki/Wikidata Glossary#Datatypes



V · T · E	Wikidata properties related to bibliographic metadata (Q20745594)
Scholarly articles	author · author name string · editor · title · subtitle · publication date · published in · volume · issue · page(s) · number of pages · license · full work available at · reference URL · language of work or name · original language of work · main subject · Commons category · sponsor · external data available at · corrigendum / erratum · *Identifiers*: article ID · DOI · PubMed ID · PMCID · arXiv ID · ZooBank publication ID · JSTOR article ID · Social Science Research Network ID · NIOSHTIC-2 ID · Dialnet article · OpenCitations bibliographic resource ID · ACM Digital Library citation ID · Crossref funder ID · Publons Publication ID
Scholarly journals	title · publisher · place of publication · editor · license · language of work or name · original language of work · part of · distribution · inception · dissolved, abolished or demolished · printed by · main subject · Commons category · ISO 4 abbreviation · *Identifiers*: ISSN · SUDOC editions · ZDB ID · JSTOR journal code · Norwegian Register journal ID · bibcode · Dialnet journal · HathiTrust ID · Persée journal ID · UNZ journal identifier · Research Papers in Economics Series handle · BNE journal ID · Latindex ID · ERIH PLUS ID
Proceedings	title · language of work or name · editor · volume · series · publication date · main subject · publisher · full work available at · *Identifiers*: ACM Digital Library citation ID · ISBN-13 · ISBN-10
Proceedings series	title · language of work or name · main subject · publisher · full work available at · *Identifiers*: ISSN · ACM Digital Library event ID
Books (more)	author · translator · editor · illustrator · publisher · title · subtitle · license · full work available at · language of work or name · original language of work · main subject · Commons category · *Identifiers*: ISBN-10 · ISBN-13 · DOI · HathiTrust ID
Authors (more)	name in native language · affiliation · e-mail · field of work · notable work · Commons category · author citation (zoology) · *Identifiers*: ORCID iD · VIAF ID · ZooBank author ID · Scopus Author ID · Google Scholar ID · ResearchGate person ID · ISNI · Dialnet author ID · Persée author ID · UNZ author identifier · ACM Digital Library author ID · DBLP pID
Publishers	DOI Prefix · headquarters location · field of work · Commons category

See also: **WikiProject Source MetaData**, WikiProject Periodicals, WikiProject Books

Figure 7: A navigation box used by Project Source Metadata to provide an overview of relevant properties



```
SELECT DISTINCT ?format ?formatLabel ?puid ?fdd ?solve

WHERE {

  ?format wdt:P2748 ?puid .
  ?format wdt:P3266 ?fdd .
  ?format wdt:P3381 ?solve .
  SERVICE wikibase:label {
        bd:serviceParam wikibase:language "en" .

}
  }
```

Press [CTRL-SPACE] to activate auto completion.

Figure 8: Text of a SPARQL query on the Wikidata Query Service endpoint demonstrating how the external identifiers from different systems are associated with an item.

file formats and software. The Library of Congress in the United States also publishes identifiers for file formats, know as Format Description Documents (FDDs). Editors of the "Just Solve the File Formats Problem" wiki[24] have created wiki pages for many file formats, and these can be uniquely identified by their URLs. Editors have added statements using properties that have been created for PUIDs, FDDs, and File Formats Wiki identifiers. In combination we now know how these three repositories interrelate, overlap, and where they do not overlap. Coyle noted the value of such mappings and pointed out how vital these "switching stations" would be to the web of linked open data [10]. In Figure 8, we can see a SPARQL query on the Wikidata Query Service[25]. The results of this query are returned as a table. From these results we find that Wikidata item with QID *Q278934* has an English label of *shapefile*, a PUID of *x-fmt/235*, a Library of Congress Format identifier of *fdd000280*, and a File Formats Wiki page called *Shapefile*.

Beyond providing infrastructure for this type of alignment for crosswalking between repositories, another advantage of having digital preservation technical metadata in Wikidata is to allow for queries that are ask questions of this data in combination with other data from Wikidata. For example, we can ask questions about software that leverage data from other parts of the knowledge base. In Figure 9, we see a screenshot of a timeline generated as one of the result format options on the Wikidata Query Service[26]. In this query we are asking for the dates of birth for the developers of free software to be plotted on a timeline. It is possible to run this query against the Wikidata Query Service and results will be returned. This is due to the fact that information about people is also in Wikidata. In contrast, if we attempted to run this query in the context of a repository of technical metadata related to digital preservation, results would only be returned if the repository had been populated with data about the software developers. Inclusion of data about people might be considered outside of the scope of digital preservation. It is also unclear if such a stand-alone

[24] http://fileformats.archiveteam.org/wiki/Main_Page

[25] To run this query on the endpoint follow this link http://tinyurl.com/gl3nmq8.

[26] Query url: http //tinyurl.com/mstqtpg

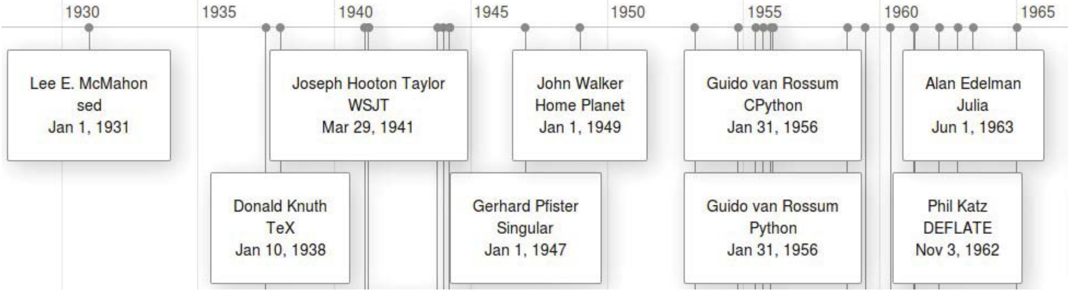


Figure 9: A SPARQL query on the Wikidata Query Service endpoint illustrating how data from the digital preservation domain can be combined with data about people.

repository for digital preservation would have the resources to maintain a SPARQL endpoint for their data.

Another example is the query seen in Figure 10[27]. In this query we are asking for the latitude and longitude for the birthplaces of developers of free software. This query demonstrates the power of storing technical metadata in a cross-domain knowledge base, and how data from the domain of digital preservation can be combined with other data to answer a broad range of questions.

Creating SPARQL queries to get data our of Wikidata allows us to present subsets of this data in a flexible manner. Such queries can be harnessed to drive future linked-data powered applications for the digital preservation domain.

5 WHY WIKIDATA?

Some of the challenges of describing the domain of digital preservation are the many divergent perspectives about how to model the domain and how to define the entities within it. In Wikidata, we have the ability to negotiate and refine the items and properties we need for our work in digital preservation, and to experiment with multiple strategies in a live system.

Due to the commons-based peer production of the Wikidata community, this is only the starting point for integration of additional data about the digital preservation domain. As more editors join the Wikidata community, an increasing number of domains are being modeled in Wikidata. New items are being created, new properties are being discussed and created, and more statements are being asserted, many with references that serve as the provenance for the claims [27].

There are several communities already working to model data within specific domains. Examples of well-established data modeling efforts are those of the Su Lab at The Scripps Research Institute and of WikiCite[28]. A group of researchers from the Su Lab have been leading efforts to model genetic information related to humans and model organisms in Wikidata [15, 23]. The Su Lab team has published documentation of their data modeling activities and created tools for their own work and made those tools available for others to use or adapt. One of the tools that the Su Lab has shared is Wikidata Integrator [43]. Wikidata Integrator is a Python library that supports the creation of Python bots to programmatically edit Wikidata.

Another active community working on data modeling is WikiCite. The participants are working together to model bibliographic relationships, and make bibliographic metadata readily available to editors who add references to Wikidata or other projects such as the different language versions of Wikipedia [36]. This group is made up of participants from many different organizations, many of whom have experience working with bibliographic metadata in several other computational workflows. Members of the WikiCite community have also created specific tools to support interacting with bibliographic metadata and Wikidata. For example, the Scholia project, a web service “creates on-the-fly scholarly profiles for researchers, organizations, journals, publishers, individual scholarly works, and for research topics” [26].

These sub-communities of the larger Wikidata project are examples of how, in a peer production system, the users are constantly updating, correcting, and implementing creative strategies to support collaborative data modeling work.

6 CONCLUSION

Through participating in the Wikidata community, we are exploring the potential for Wikidata to serve as the repository of technical metadata for the international digital preservation community. An advantage to using the Wikidata infrastructure is that this data is structured, queryable and computable [7, 29]. The value of collaborating with people domain expertise in digital preservation in creating the data models for the digital preservation is that their domain expertise is helpful in creating the underlying conceptual relationships of the knowledge base. Contributors to Wikidata create entity-level content, but also create the conceptual structure for the system [28].

Wikidata is multi-lingual knowledge base with support for more than 350 languages[14]. The multi-lingual design of Wikidata means that people from many different language communities are able to read the content of Wikidata in their own languages. For example, in Figure 11 we see a side-by-side comparison of the infoboxes for Adobe Acrobat in English Wikipedia, French Wikipedia, and Japanese Wikipedia. This is an advantage over knowledge bases that only support users of a single language. Due to the way Wikidata was designed, the work of an editor to add content to Wikidata

[27] http://tinyurl.com/mwuxhse
[28] https://meta.wikimedia.org/wiki/WikiCite

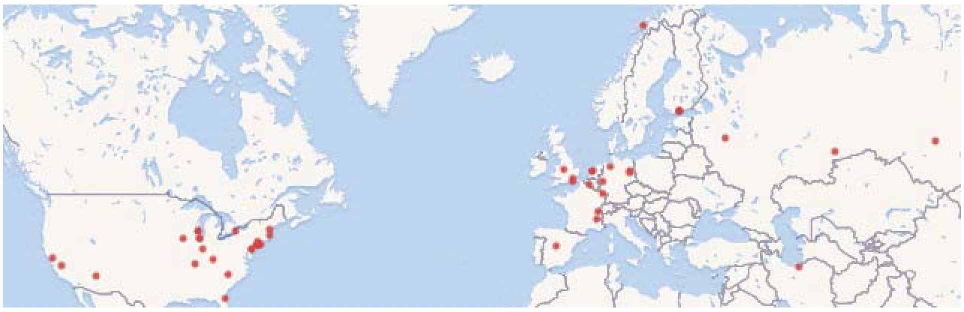
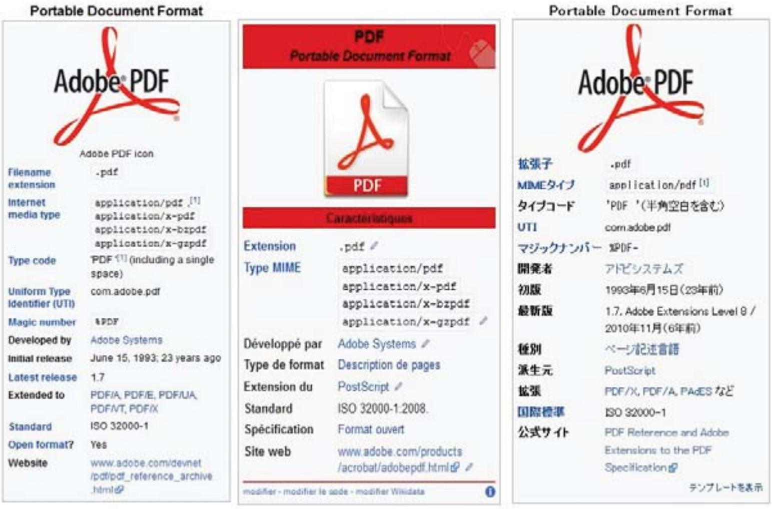


Figure 10: A SPARQL query on the Wikidata Query Service endpoint illustrating how data from the digital preservation domain can be combined with data about geography.



Portable Document Format

Adobe PDF icon

Filename extension	.pdf
Internet media type	application/pdf ,[1] application/x-pdf application/x-bzpdf application/x-gzpdf
Type code	'PDF '[1] (including a single space)
Uniform Type Identifier (UTI)	com.adobe.pdf
Magic number	%PDF
Developed by	Adobe Systems
Initial release	June 15, 1993; 23 years ago
Latest release	1.7
Extended to	PDF/A, PDF/E, PDF/UA, PDF/VT, PDF/X
Standard	ISO 32000-1
Open format?	Yes
Website	www.adobe.com/devnet/pdf/pdf_reference_archive.html

PDF
Portable Document Format

Caractéristiques

Extension	.pdf
Type MIME	application/pdf application/x-pdf application/x-bzpdf application/x-gzpdf
Développé par	Adobe Systems
Type de format	Description de pages
Extension du	PostScript
Standard	ISO 32000-1:2008.
Spécification	Format ouvert
Site web	www.adobe.com/products/acrobat/adobepdf.html

modifier - modifier le code - modifier Wikidata

Portable Document Format

拡張子	.pdf
MIMEタイプ	application/pdf[1]
タイプコード	'PDF '（半角空白を含む）
UTI	com.adobe.pdf
マジックナンバー	%PDF-
開発者	アドビシステムズ
初版	1993年6月15日（23年前）
最新版	1.7, Adobe Extensions Level 8 / 2010年11月（6年前）
種別	ページ記述言語
派生元	PostScript
拡張	PDF/X, PDF/A, PAdES など
国際標準	ISO 32000-1
公式サイト	PDF Reference and Adobe Extensions to the PDF Specification

テンプレートを表示

Figure 11: A side-by-side comparison of the infobox for Adobe PDF from English Wikipedia, French Wikipedia, and Japanese Wikipedia.

benefits users across all supported languages as that data is available in each of the supported languages.

We see value in working to create this infrastructural component of the web of linked data. We are working together to create this content in a system that was designed to ensure that this structured data will belong to the public domain in order to avoid the situation of needing to purchase access to this data from a for-profit business organization [19].

Making a decision to invest in a project that is building infrastructure for the public domain may not be possible for all domains. In the domain of digital preservation, activities that contribute to public domain infrastructure will allow us all to share the work of maintaining access to this structured data. We are all free to fork this project at any time, all data is downloadable at any time, and we can reuse data from any contributor to Wikidata.

If we are successfully in curating data related to the technical metadata of digital preservation then we will have created a powerful resource that will enable others to reuse this metadata for their own systems and projects. Once we have established which are the readable file formats for a particular version of a particular piece of software and curated that data in Wikidata then, by virtue of the Creative Commons Zero (CC0) license[29], others can reuse that data rather than create it anew themselves.

Imagine a future in which the technical metadata we need for our digital preservation work is freely available to all as linked open data. Please join us as we build this future together.

7 ACKNOWLEDGEMENTS

We thank Wikimedia Deutchland[30] for supporting the Wikidata project. We thank the Wikidata community for sharing thoughts,

[29] https://creativecommons.org/choose/zero/

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

feedback, and data modeling work. Specifically, would like to thank all participants of WikiProject Informatics[31] for engaging with us as we work to describe the domain of digital preservation in Wikidata.

REFERENCES

[1] 2016. MediaWiki. https://www.mediawiki.org/wiki/MediaWiki. (2016). Visited July 1, 2016.

[2] UK National Archives. 2017. Download DROID: file format identification tool. (2017).

[3] Yochai Benkler. 2002. Coase's Penguin, or, Linux and The Nature of the Firm. *Yale Law Journal* (2002), 369–446.

[4] Yochai Benkler, Aaron Shaw, and Benjamin Mako Hill. 2013. Peer production: a modality of collective intelligence. *Collective Intelligence* (2013).

[5] Freddy Brasileiro, João Paulo A Almeida, Victorio A Carvalho, and Giancarlo Guizzardi. 2016. Applying a multi-level modeling theory to assess taxonomic hierarchies in Wikidata. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 975–980.

[6] Tim Brody, Leslie Carr, Jessie Hey, Adrian Brown, and Steve Hitchcock. 2008. PRONOM-ROAR: Adding format profiles to a repository registry to inform preservation services. *International Journal of Digital Curation* 2, 2 (2008), 3–19.

[7] Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Elvira Mitraka, Julia Turner, Tim Putman, Justin Leong, Chinmay Naik, Paul Pavlidis, Lynn Schriml, Benjamin M Good, and others. 2016. Wikidata as a semantic framework for the Gene Wiki initiative. *Database* 2016 (2016), baw015.

[8] Moira Burke and Robert Kraut. 2008. Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 27–36.

[9] Sunny Consolvo, Predrag Klasnja, David W McDonald, Daniel Avrahami, Jon Froehlich, Louis LeGrand, Ryan Libby, Keith Mosher, and James A Landay. 2008. Flowers or a robot army?: encouraging awareness & activity with personal, mobile displays. In *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 54–63.

[10] Karen Coyle. 2010. Understanding the Semantic Web: Bibliographic Data and Metadata. *ALA Library Technology Reports* 1 (2010), 5–31.

[11] Angela Dappert, Rebecca Squire Guenther, and Sébastien Peyrard. 2016. *Digital Preservation Metadata for Practitioners: Implementing PREMIS*. Springer.

[12] Angela Dappert, Sébastien Peyrard, Janet Delve, and CC Chou. 2012. Describing digital object environments in premis. In *9th International Conference on Preservation of Digital Objects (iPRES2012)*. 69–76.

[13] Bob DuCharme. 2013. *Learning Sparql*. O'Reilly Media, Inc.

[14] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing Wikidata to the linked data web. In *The Semantic Web–ISWC 2014*. Springer, 50–65.

[15] Benjamin Good. 2015. Poof it works – using wikidata to build Wikipedia articles about genes. http://sulab.org/2015/10/poof-it-works-using-wikidata-to-build-wikipedia-articles-about-genes/. (2015). Online; accessed 18 Feb 2016.

[16] Daniel Hernández, Aidan Hogan, and Markus Krötzsch. 2015. Reifying RDF: What Works Well With Wikidata? *SSWS@ ISWC* 1457 (2015), 32–47.

[17] Ali Ismayilov, Dimitris Kontokostas, Sören Auer, Jens Lehmann, and Sebastian Hellmann. 2015. Wikidata through the Eyes of DBpedia. *arXiv preprint arXiv:1507.04180* (2015).

[18] Lucie-Aimée Kaffee. 2016. *Generating Article Placeholders from Wikidata for Wikipedia: Increasing Access to Free and Open Knowledge*. Ph.D. Dissertation. University of Applied Sciences.

[19] Jaron Lanier. 2014. *Who owns the future?* Simon and Schuster.

[20] Richard Lehane. 2017. Sigfried. (2017).

[21] Peter McKinney, Steve Knight, Jay Gattuso, David Pearson, Libor Coufal, David Anderson, Janet Delve, Kevin De Vorsey, Ross Spencer, and Jan Hutař. 2014. Reimagining the Format Model: Introducing the Work of the NSLA Digital Preservation Technical Registry. *New Review of Information Networking* 19, 2 (2014), 96–123.

[22] Peter McKinney, David Pearson, David Anderson, Jan Hutař, Steve Knight, Libor Coufal, Janet Delve, Jay Gattuso, Kevin DeVorsey, and Ross Spencer. 2014. A next generation technical registry: moving practice forward. (2014).

[23] Elvira Mitraka, Andra Waagmeester, Sebastian Burgstaller-Muehlbacher, Lynn M Schriml, Andrew I Su, and Benjamin M Good. 2015. Wikidata: A platform for data integration and dissemination for the life sciences and beyond. *bioRxiv* (2015), 031971.

[24] Jonathan T Morgan, Michael Gilbert, Mark Zachry, and David McDonald. 2013. A content analysis of Wikiproject discussions: Toward a typology of coordination language used by virtual teams. In *Proceedings of the 2013 conference on Computer supported cooperative work companion*. ACM, 231–234.

[25] Claudia Müller-Birn, Benjamin Karran, Janette Lehmann, and Markus Luczak-Rösch. 2015. Peer-production system or collaborative ontology engineering effort: What is Wikidata?. In *Proceedings of the 11th International Symposium on Open Collaboration*. ACM, 20.

[26] Finn Årup Nielsen, Daniel Mietchen, and Egon Willighagen. 2017. Scholia and scientometrics with Wikidata. *arXiv preprint arXiv:1703.04222* (2017).

[27] Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From freebase to wikidata: The great migration. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1419–1428.

[28] Alessandro Piscopo, Christopher Phethean, and Elena Simperl. 2017. Wikidatians are Born: Paths to Full Participation in a Collaborative Structured Knowledge Base. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

[29] Tim E Putman, Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Chunlei Wu, Andrew I Su, and Benjamin M Good. 2016. Centralizing content and distributing labor: a community model for curating the very long tail of microbial genomes. *Database* 2016 (2016), baw028.

[30] Tim E Putman, Sebastien Lelong, Sebastian Burgstaller-Muehlbacher, Andra Waagmeester, Colin Diesh, Nathan Dunn, Monica Munoz-Torres, Gregory Stupp, Andrew Su, and Benjamin M Good. 2017. WikiGenomes: an open Web application for community consumption and curation of gene annotation data in Wikidata. *bioRxiv* (2017), 102046.

[31] Kjeld Schmidt. 2002. The problem withawareness': Introductory remarks onawareness in CSCW'. *Computer Supported Cooperative Work (CSCW)* 11, 3-4 (2002), 285–298.

[32] Andreas Spitz, Vaibhav Dixit, Ludwig Richter, Michael Gertz, and Johanna Geiß. 2016. State of the Union: A Data Consumer's Perspective on Wikidata and Its Properties for the Classification and Resolution of Entities. In *Wiki Workshop at ICWSM*.

[33] Richard Stallman. 2009. Viewpoint Why open source misses the point of free software. *Commun. ACM* 52, 6 (2009), 31–33.

[34] Richard M Stallman. 1990. The GNU manifesto. In *Computers, ethics, & society*. Oxford University Press, Inc., 308–317.

[35] Susan Leigh Star and Karen Ruhleder. 1994. Steps towards an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM, 253–264.

[36] D. Taraborelli, J. Dugan, L. Pintscher, D. Mietchen, and C. Neylon. 2016. WikiCite 2016 Report. (2016).

[37] Global Digital Format Registry Team. 2008. Global Digital Format Registry. http://library.harvard.edu/preservation/digital-preservation_gdfr.html. (2008). Visited Nov 16, 2016.

[38] Unified Digital Format Registry Team. 2012. Unified Digital Format Registry. http://www.udfr.org/. (2012). Visited Nov 16, 2016.

[39] Denny Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference companion on World Wide Web*. ACM, 1063–1064.

[40] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[41] Wikibase. 2015. Wikibase. http://wikiba.se/. (2015). Online; accessed 15 January 2016.

[42] Wikidata. 2015. DataModel. (2015). https://www.mediawiki.org/wiki/Wikibase/DataModel

[43] Wikidata. 2017. Q31743627 — Wikidata. (2017).

[44] Haiyi Zhu, Robert E Kraut, Yi-Chia Wang, and Aniket Kittur. 2011. Identifying shared leadership in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3431–3434.

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