Abstract. Wikidata is evolving as the hub of Linked Open Data (LOD), with its language-neutral URIs and close adherence to Wikipedia. Well defined URIs help the data to be interoperable and linkable. This paper examines the possibilities of utilizing Wikidata as the means of a vocabulary resource for promoting the use of linkable concepts. Digital curation projects are vibrant with varying demands and purposes, which makes them less suitable for adopting any common vocabularies or ontologies. Also, developing and maintaining custom vocabularies are expensive processes for smaller projects in terms of resources and skill requirements. In general, Wikidata entities are well documented with Wikipedia entries, and Wikipedia entries express the conceptual and hierarchical relations in detail with provisions to modify or create. The authors explain the concept of using Wikidata as a vocabulary source with a proof of concept module implementation for Omeka-S, a widely adapted open source digital curation platform. This paper is expected to show some practical insights on reliable and reasonable vocabulary development for social informatics as well as cultural heritage projects, with a notion to improve the quality and quantity of linkable data from digital curation projects.

Keywords: Metadata · Wikidata · URI · Vocabulary · Semantic web · Digital curation · Omeka

1 Introduction

Wikidata is evolving as the hub of Linked Open Data (LOD), with its language-neutral URIs and close adherence to Wikipedia and as the core project of Wikimedia’s data management strategy [9]. This paper examines the possibilities of utilizing Wikidata as a vocabulary resource for promoting the use of linkable
concepts. The authors evaluate this proposal in the context of social informatics (SI) scenarios. SI projects are vibrant with varying demands and purposes, which makes them less suitable for adopting any common vocabularies or ontologies. Also, developing and maintaining custom vocabularies are expensive processes for smaller projects in terms of resources and skill requirements. In general, Wikidata entities are well documented with Wikipedia entries, and Wikipedia entries express the conceptual and hierarchical relations in detail with provisions to modify or create. These detailed articles act as documentation for the Wikidata entity and help the domain experts to use precise entities than abstract concepts. Wikidata as a vocabulary is not a novel concept, but there are less practical and easy to use integration for less skilled users and communities. This paper explains the concept of using Wikidata as a vocabulary source and a proof of concept implementation by developing an open source module for Omeka-S\(^1\), a semantic web digital curation system, to easily find and use Wikidata URIs and use it as a linkable and comprehensive vocabulary.

1.1 Significance of Wikidata in Linked Data

Wikidata Project is growing along with the acceptance of Wikipedia. Wikidata is tightly bound to Wikipedia, and that makes it a sustainable project. Community driven model prompts it to be integral and well maintained with accurate content and reviewable contributions. Wikipedia eventually became the de-facto standard of open knowledge. Wikipedia collects and maintains multilingual unstructured textual information through community contributions. Whereas the Wikidata process and present structured information from unstructured Wikipedia entries. Wikidata also provides multilingual information, but still maintain language-neutral cool URIs [10].

Similar to Wikipedia, Wikidata is organized as pages. Every subject is called an entity, and every entity has its own editable page. There are two types of entities; they are items, and properties. Wikidata organise information as entities and these entities can be either an item or a property. Items are individuals and classes, but properties are similar to RDF properties. Wikipedia article in any language has a Wikidata item, which represents the subject of the article [2]. Wikidata accepts corrections and contributions from its community the same way as it does with Wikipedia. The corrections or changes can be real-time, which makes it acceptable to changes and additions faster than any other knowledge systems. Wikidata is not just a datasource, but a platform as well [12] and in the last couple of years, Wikidata is moving closer to the center of the LOD cloud by surpassing similar Wikipedia driven projects.

1.2 Wikidata as a Vocabulary Source for Digital Curation

Wikidata can be treated as a controlled vocabulary resource in places where we can use it to replace literals. Multi-lingual labels in Wikidata permits the use of

\(^1\) https://omeka.org/s/.
any Wikidata entity to be mapped with a persistent URI irrespective of the language. In the context of Social Informatics projects, the stakeholders can always find the relevant resources from Wikipedia and use the corresponding Wikidata entity as a URI, with suitable labels. Concepts are always linked to the same URI resources irrespective of the language or domain of the project. A higher level of interoperability with precise concept mapping can be obtained from this approach. Wikidata, so as Wikipedia is a domain-independent knowledge graph and the URIs serve valid machine-actionable as well as human-interpretable resources. The machine actionability gives the possibility of using these URIs to further automated knowledge extraction processes.

2 Related Works

Product ontology described a way to develop ontologies for e-commerce definitions using Wikipedia [7]. Product ontology uses Wikipedia, and Wikipedia based URIs in Product Ontology namespace [8]. This paper proposes similar attempts using Wikidata entities instead of Wikipedia and emphasizes a clear language-neutral URI concept with a more LOD cloud oriented perspective. Another related work is using Wikidata as a semantic framework for the Gene Wiki. This attempt created a fully open and extensible data resource for human and mouse molecular biology and biochemistry data [1]. Gene Wiki resource enriches all the Wikipedias with structured information and serves as a new linking hub for the biological semantic web.

Organisations like Europeana encourages users to adapt linkable URIs from Wikidata as a vocabulary to semantically enrich the open data. Using URIs as identifiers will enhance the reusability and interoperability of the data. Studies were conducted on describing digital objects within the context of Europeana with Wikidata [3]. Various projects uses Wikidata as an authority record for obtaining persistent concept URIs to use within ontologies and linkable datasets [5,6]. General purpose ontologies like schema.org makes such process simple, seamless and efficient. There were studies to evaluate the linkage between Library of Congress Subject Heading (LCSH) and Wikidata [11] for vocabulary alignment. Wikidata was also used as a work authority for video games [4].

OpenRefine, an open-source data cleansing and transforming tool utilizes Wikidata reconciliation APIs to align datasets to Wikidata items.

This paper proposes more broader approaches than being domain specific, by developing a plugin for Omeka-S, a general purpose digital curation system.

2 http://www.productontology.org/.
3 https://pro.europeana.eu/page/get-your-vocabularies-in-wikidata.
4 https://schema.org/.
5 https://openrefine.org/.
6 https://wikidata.reconci.link/.
2.1 Manually Finding URIs from Wikipedia and Wikidata

The most simple and straightforward method to use Wikidata is to find the corresponding entities from Wikipedia. Every article in Wikipedia has an entity mapped in Wikidata. These Wikidata articles can act as good documentation for the Wikipedia entry. However this process is cumbersome for maintaining bigger digital curation projects, especially in digital humanities, cultural heritage and social informatics.

Domain experts can always create new articles within Wikipedia to obtain more precise URIs in Wikidata namespace. When a specific Wikidata entry is not exist for any given concepts, the users can create it by editing Wikidata. If any specific properties are not available, users have to propose new properties in the Wikidata ontology, which is a systematic process, and the decision is taken by the Wikidata maintainers. This option to create new articles promotes bi-directional growth by increasing the number of articles as well as in return, obtaining precise concept URIs.

3 Methods

Using Wikidata URIs to describe resources is getting acceptance from the Semantic Web communities. Developing and integrating a ready to use and extensible system will help the adaptation process smoother for communities with limited resources. Omeka S is versatile and easy to use Semantic Web digital curation systems. We anticipate that implementing our proposal within Omeka will be beneficial for many. Omeka S is widely used in cultural heritage and digital humanities curation. Omeka S is developed with PHP, and it is easy to install, configure, and maintain. Omeka Team also provides a module name ValueSuggest7 to search and auto-populate vocabularies within Omeka S. Our implementation is based on ValueSuggest module, which is under GPLv3. ValueSuggest module provides an excellent framework to develop a vocabulary search tool that can be used within Omeka ecosystem.

Wikidata offers different APIs to query and retrieve data. The primary interface for querying Wikidata is its SPARQL endpoint8. Wikidata also provides a MediaWiki API9, which is fast and powerful for basic operations. Wikidata community provides a Wikidata reconciliation web service API for OpenRefine10. This API can be used to align datasets to Wikidata items in OpenRefine. As per the Reconciliation Service API specification11, the API has different querying, recommendations, and preview endpoint implementations. This module utilizes all these endpoints to form a querying and previewing system effectively. Due to the performance advantages, main querying interfaces are built on top of the

7 https://omeka.org/s/modules/ValueSuggest/.
8 https://query.wikidata.org/.
9 https://www.wikidata.org/w/api.php.
10 https://wikidata.reconci.link/.
11 https://reconciliation-api.github.io/specs/latest/.
Fig. 1. Overview of the modules querying flow and corresponding API services

MediaWiki API and reconciliation API than with Wikidata SPARQL API. However, for advanced scenarios and use-cases, a SPARQL based querying system will be efficient than the other two. A detailed overview of the modules querying flow and corresponding API services are illustrated in Fig. 1.

The querying is dealt through the module interface within Omeka, and the previews are obtained from the preview endpoint of the reconciliation API directly from the web browser. As a default nature, the APIs’ response is cached within the Omeka server using file cache for an hour, this improves performance and reduces response time\textsuperscript{12}. Caching helps collaborative editing for bigger projects stable and saves resources. The module provides extensible filter samples for MediaWiki API, OpenRefine reconciliation API, and Wikidata SPARQL API. A detailed matrix of default filters and interfaces are provided in Table 1. The examples are sufficient enough to cover most of the general use-cases. Users can modify or add more filters based on the given examples for specific use cases or create more complex filters to cater to their particular use-cases.

Table 1. Default filters and interfaces

| List type       | Description                  | API service   | API type     | Constraint   |
|-----------------|------------------------------|---------------|--------------|--------------|
| All             | All Wikidata items           | MediaWiki     | wbsearchentities |             |
| Entities        | All Wikidata entities        | OpenRefine    | Suggest      | Entity       |
| Properties      | All Wikidata properties      | OpenRefine    | Suggest      | Property     |
| Persons         | Persons from Wikidata        | OpenRefine    | Reconciliation | Q5           |
| Locations       | Locations from Wikidata      | OpenRefine    | Reconciliation | Q2221906    |
| Languages       | Languages from Wikidata      | OpenRefine    | Reconciliation | Q315         |
| Custom          | Custom filter example        | OpenRefine    | Reconciliation |             |
| SPARQL          | SPARQL example               | SPARQL        | Query        |             |

\textsuperscript{12} https://docs.zendframework.com/zend-cache/.
4 Results

The authors successfully implemented the concepts discussed in this paper and developed a wikidata module for Omeka-S, with which digital curators can easily search and populate relevant terms and URIs from Wikidata. This module is released as an opensource software under GNU General Public License v3. Omeka-S users can easily install and use this module within their Omeka-S instances. The entire installable source code is accessible from the project GitHub repository at https://github.com/nishad/omeka-s-wikidata. The module is listed in Omeka-S website at https://omeka.org/s/modules/Wikidata/.

Once installed and activated, preconfigured wikidata data types can be selected for properties in resource templates within the Omeka-S control panel. By default, the module provides ready to use configurations for all Wikidata items, entities, properties, places, persons, and languages. Data type selection for properties is shown in Fig. 2.

Fig. 2. Preconfigured Wikidata data types for properties in resource templates

Fig. 3. Example of searching for entities indicating a person
Users can directly type in the search terms within the property fields for the items, and the module will provide an easy to access list of search results to pick the proper term. To avoid the confusion of similar labels, further information can be viewed on mouse hovering individual results. A default search for a person is demonstrated in Fig. 3.

The module is multi-lingual by default. Wikidata names in available languages can be retrieved by setting the language field for every entry. Multi-lingual capability of the module is demonstrated in Fig. 4.

![Fig. 4. Wikidata names in available languages can be retrieved by setting the language field](image)

Upon selecting the appropriate entry, the module automatically populates the URI and label in the selected language, as demonstrated in Fig. 5.

![Fig. 5. Automatically populated Wikidata URI and label](image)

5 Conclusion

Linking data from different domains will help to develop new possibilities of information-driven knowledge discovery. Linking concepts through URIs will improve the quality and availability of linkable data. In terms of semantic web
based digital curation, the impact of such projects can be increased drastically by introducing linkable URIs. More practical approaches and utilities will help implementers and users to adapt Wikidata centric URIs in their curation platforms and Content Management Systems.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Number JP18K11984.

The authors acknowledge the Omeka Team, for developing Omeka-S, a wonderful open-source Semantic Web digital curation platform software, and the ValueSuggest module.

This paper uses icon of ‘computer user’ (https://thenounproject.com/term/computer-user/1577369) by bezier master from the Noun Project, licensed under Creative Commons CCBY.

References

1. Burgstaller-Muehlbacher, S., et al.: Wikidata as a semantic framework for the Gene Wiki initiative. Database 2016, January 2016. https://doi.org/10.1093/database/baw015. https://academic.oup.com/database/article/doi/10.1093/database/baw015/2630183

2. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing Wikidata to the linked data web. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8796, pp. 50–65. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11964-9_4

3. Freire, N., Isaac, A.: Wikidata’s linked data for cultural heritage digital resources: an evaluation based on the Europeana data model. In: Proceedings of the 2019 International Conference on Dublin Core and Metadata Applications, DCMI 2019, Dublin Core Metadata Initiative, pp. 59–68 (2019)

4. Fukuda, K.: Using Wikidata as work authority for video games. In: Proceedings of the 2019 International Conference on Dublin Core and Metadata Applications, DCMI 2019, Dublin Core Metadata Initiative, pp. 80–87 (2019)

5. Gelati, F.: Archival Authority Records as Linked Data thanks to Wikidata, schema.org and the Records in Context Ontology, September 2019. https://doi.org/10.5281/zenodo.3465304

6. Gelati, F.: Personen-Datenbank Zeitgeschichte: Code Pre-Release, October 2019. https://doi.org/10.5281/zenodo.3484706

7. Hepp, M.: GoodRelations: an ontology for describing products and services offers on the web. In: Gangemi, A., Euzenat, J. (eds.) EKAW 2008. LNCS (LNAI), vol. 5268, pp. 329–346. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87696-0_29

8. Hepp, M., Siorpas, K., Bachlechner, D.: Harvesting wiki consensus: using Wikipedia entries as vocabulary for knowledge management. IEEE Internet Comput. 11(5), 54–65 (2007). https://doi.org/10.1109/MIC.2007.110

9. Malyshev, S., Krötzsch, M., González, L., Gonsior, J., Bielefeldt, A.: Getting the most out of Wikidata: semantic technology usage in Wikipedia’s knowledge graph. In: Vrandečić, D., et al. (eds.) ISWC 2018. LNCS, vol. 11137, pp. 376–394. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00668-6_23

10. Sauermann, L., Cyganiak, R.: Cool URIs for the semantic web. W3C note, W3C, December 2008. https://www.w3.org/TR/2008/NOTE-cooluris-20081203/
11. Snyder, E., Lorenzo, L., Mak, L.: Linked open data for subject discovery: assessing the alignment between library of congress vocabularies and Wikidata. In: Proceedings of the 2019 International Conference on Dublin Core and Metadata Applications, DCMI 2019, Dublin Core Metadata Initiative, pp. 12–20 (2019)

12. Ogbuji, U.: Introduction to Wikidata as a platform and data source, November 2018. https://developer.ibm.com/technologies/artificial-intelligence/articles/use-wikidata-in-ai-and-cognitive-applications-pt1/