WikiframeVG: A SPARQL Template-based Framework for Wikidata Graph Exploration and Visualization

Darnelle Melvin[1,*], Andre Hulet[1] and Cory Lampert[1]

[1] University Libraries. University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV, 89154, USA

Abstract

We present WikiframeVG (Wikiframe Visual Graph), an open-source community initiative and tool which helps Wikidata editors and users explore knowledge generated from organized Wikidata sprints. WikiframeVG adopts a community driven, SPARQL template-based approach towards Wikidata graph exploration. We define the Wikiframe end user and project scope, describe Wikiframe templates, its data architecture and site architecture, introduce the user interface, present a discussion, and conclusion.

Keywords

Wikidata, Visualization Tools, Content Selection, Data Presentation

1. Introduction

In recent years, interest in galleries, libraries, archives, museums, and special collections (GLAMS) has grown in Wikidata and Wikibase contributions and consumption [1-4]. Over time, Wikidata [5] has become a large multilingual, crowdsourced, public knowledge graph of high-quality general facts and a central hub of persistent identifiers [6, 7]. However, there remains a barrier for those who lack technical or domain expertise, knowledge on how to write SPARQL [8] for graph query or construction [9-11], let alone visualize or make sense of such a giant cloud of data [12]. One solution to this problem is a graph-based discovery interface that extracts Wikidata subsets and presents these nodes and edges into frames of interest. In this initial prototype, the frame of interest shall focus on the GLAMS community. Here, we propose Wikiframe Visual Graph (WikiframeVG),[2] a tool which helps Wikidata editors and general browsers visualize and explore knowledge generated from organized Wikidata sprints. WikiframeVG is an ecosystem built upon open source technologies that provides SPARQL templates, property mappings, and configuration files allowing users a no-code approach towards Wikidata graph exploration.

In this paper we define the Wikiframe end user and project scope (Section 2). Then describe Wikiframe templates (Section 3), its data architecture (Section 4), site architecture (Section 5), and introduce the user interface, including data previews, search techniques and graph explorations (Section 6). We then present a discussion (Section 7) and end with conclusions and final thoughts (Section 8).

2. End Users and Project Scope

The WikiframeVG end users are a diverse group of researchers and practitioners consisting of developers, metadata librarians, digital repository managers, archivists, oral historians, digital humanists and library technologists. These interdisciplinary participants come primarily from the library and information science, museum studies, or digital humanities fields and tend to have

Wikidata'23: Wikidata Workshop at ISWC 2023

* Corresponding author.

darnelle.melvin@unlv.edu (D. Melvin); andre.hulet@unlv.edu (A. Hulet); cory.lampert@unlv.edu (C. Lampert)

0000-0002-4614-3504 (D. Melvin); 0009-0006-8772-9735 (A. Hulet); 0000-0002-9467-5214 (C. Lampert)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

[2] https://wikiframe.library.unlv.edu/

CEUR-WS.org/Vol-3640/paper15.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

a wide range of technical experience. This segment includes community members interested in incorporating SPARQL into existing and new workflows [13] and participants in both the Program for Cooperative Cataloging (PCC) Wikidata Pilot[3] and LD4 Wikidata Affinity Group.[4] The PCC Wikidata Pilot program ran from September 2020 through December 2021, where 76 PCC and non-PCC institutions collaborated to help facilitate the communities move towards identity management. The pilot was sponsored by the PCC - Task Group on Identity Management in NACO [14].

To support the linked open data community, WikiframeVG was built upon open data and designed for both expert and general use cases. Expert users have domain knowledge and may know how to write SPARQL statements. Potential use cases for these types of users include creating multiple options to visualize data (e.g. tables, maps, vertices and edges, or timelines), providing mechanisms to combine linked open data with other datasets, and integrating new ways to visualize similar things [15]. However, most potential users fall under the category of *lay users*. Helmich [16] describes these users as novice or expert web navigators, but users who do not possess knowledge of Semantic Web technologies such as RDF, SPARQL, or ontologies. These users are our target audience and require better tools to present, organize and visualize structured data.

The project scope for Wikiframe development came to fruition after a series of informal conversations concluding the 2021 LD4 Conference on Linked Data [17]. Out of these talks, community members realized a need to create a visualization tool based on a series of core requirements:

- build it to utilize structured data in the public domain;
- build it using open-source technologies;
- create the ecosystem where no SPARQL experience is needed; and
- build it, govern, and sustain it as a community of practice.

These four requirements act as guiding principles for all current and future Wikiframe designs and enable the community to build the application using a Wikiframe template approach.

3. Wikiframe Template

The Wikiframe template is a SPARQL statement that defines the scope of data for an element of an application theme, such as the metadata describing resources and materials for GLAMS organizations. A template creates a classification scheme that may or may not be directly supported by the Wikidata ontology and available entities. In other words, whether Wikidata defines an archival "collection" entity or not, one can write SPARQL that defines what a given organization means by "collection." In our case, we used *instance of (wdt:P31)* object values to define "collection," based on the data modeling methods of UNLV staff, (e.g., *human (wd:Q5), oral history (wd:Q558929), photo archive (wd:Q27032363)*, *manuscript collection (wd:Q42939539), etc.*). The template also includes key/value pairs for essential properties statements and their corresponding qualifiers of the class, such as properties: *archival resource key (wdt:P8091), archives at (wdt:P485),* and *oral history at (wdt:P9600)*. Examples of qualifiers include: *subject named as (pq:P1810), inventory number (pq:P217),* and *described at url (pq:P973)*. Every time the template is executed in Wikiframe, it caches all items within the class and their essential properties, which are then used by the search engine.

4. Data Architecture

[3] https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot/Participants
[4] https://sites.google.com/stanford.edu/ld4-community-site/groups#h.dm8tfdn1yc6h

To improve the speed of retrieval and to moderate the load on the Wikidata Query Service,[5] data returned by templates are cached in the Wikiframe back-end database. MySQL[6] was selected for this purpose, and is one of several database options that could have been used. Each result set is cached in its own table in a flat, unnormalized structure, and Wikiframe application logic treats each table as a distinct class of data. It was decided not to store results as semantic triples due to the time and complexity needed to create a performant and easily updatable application that relies purely on SPARQL for retrieval. The templated application classes simplify the structure and possible combinations of result sets, making it relatively easy to re-create semantic relationships on-the-fly using program iteration techniques.

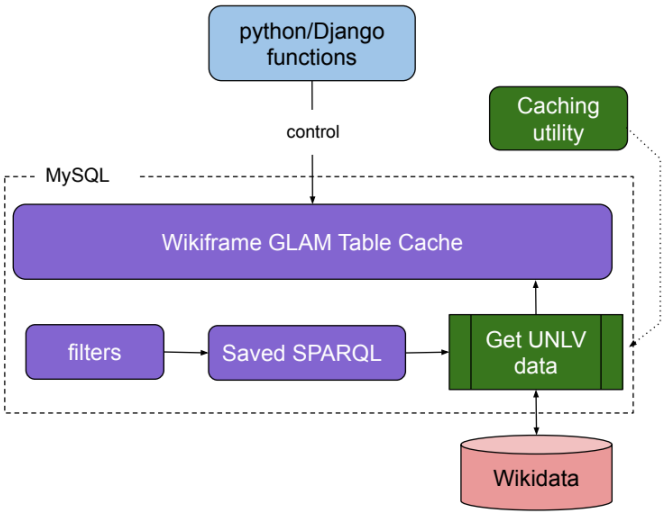

Figure 1: WikiframeVG data architecture

Figure 1 shows the main components and data flow within Wikiframe. All search and retrieval logic is written in Python[7] and uses the Django web application framework[8] to control database access and perform any needed updates to the database structure. Configuration tables store the SPARQL statements for execution against the Wikidata Query Service, and Wikidata is retrieved either by an automated utility or from a utility page within Wikiframe.

While templates control the application theme's data scope, a filter tagging system controls the organizational data scope. The templates must include statements that retrieve only a single organization's data from the set of all data available within the theme. In our case, we use a filter to set the *on focus list of Wikimedia project (wdt:P5008)* property value to UNLV's project identifier, *WikiProject PCC Wikidata Pilot/University of Nevada, Las Vegas (wd:Q100202113)*. Filters are defined and stored in their own configuration table. They consist of a filter key and a corresponding property identifier. The filter key is saved in the template SPARQL in place of the actual identifier using the tag format, `~[filt-key-name]~`. At runtime this tag is replaced with the identifier value stored in the configuration table. This approach allows organizations to reuse a Wikiframe theme as-is, simply by placing their own identifier (in this case, it would be their code for *(wdt:P5008)*) in the configuration table.

The Wikiframe cache is refreshed every 24 hours by an automated utility. The utility submits the SPARQL template query to the Wikidata query service for each application class: Collections, Oral Histories, People, Corporate Bodies, and Subjects. Each result is returned as a JSON document which is parsed and saved to a MySQL table for the given class. Cache tables are then queried by the Wikiframe search engine using Django's Object Relational Mapping framework.

[5] https://query.wikidata.org/
[6] https://www.mysql.com/
[7] https://www.python.org/
[8] https://www.djangoproject.com/

5. Site Architecture

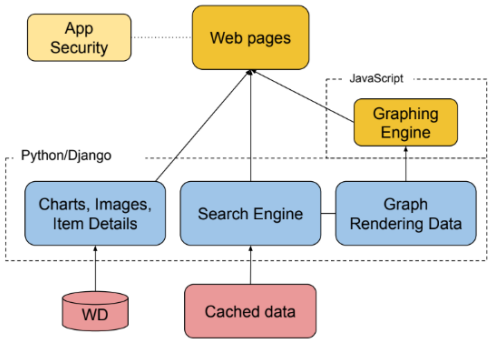

Figure 2: WikiframeVG site architecture

As shown in Figure 2, Wikiframe application logic is organized into four main components. Some data, namely for charts, images, and all item details are retrieved live from Wikidata. These are small record sets that don't need to be cached and, in the case of item details, would unnecessarily complicate caching. The search engine accepts keywords and phrases and graph node identifiers, then runs appropriate queries, reconstitutes semantic relationships based on data mappings stored in the source code, and passes that data to the graphing engine, Vis.js.[9] For instance, the relationship type mapping for the Person class allows relationships according to occupation, field of work, place of birth, and place of death, while the Corporate Body class mapping shows relationships for "instance of" and the main subject. Users can then turn available relationships between individual results on and off in the graph. Relationship and query mappings can be updated via a configuration table and the mappings file.

The search execution process also generates graph rendering data, which Django passes to the web page, where it is consumed by application JavaScript functions that render and annotate the graph. This enables simultaneous search results lists and their accompanying graph, which is a key aspect of our system use case: users should receive both authoritative results and see the semantic relationships between them that drive further insight into the contents of GLAMS information systems.

Table 1
WikiframeVG: Application Statistics

Measure		Value
User	Unique browser session per week, avg.	42.5
Server Utilization (6 cores, 16 GB RAM)		
	Weekly CPU utilization, avg. pct.	3
	CPU utilization, lifetime max. pct.	73.17
	RAM use, lifetime avg. pct.	8.73
	RAM use, lifetime max. pct.	9.73
Wikidata Cache		
	Cache size, MB	2.8
	Total rows	5017
Search Engine Performance		
	Average response time for Collections cache, seconds	.24

[9] https://visjs.org/

Table 1 shows performance statistics to date for WikiframeVG. As this application is still in the proof of concept stage and has not yet been widely shared or publicized, the number of connections has been low, and the small size of the UNLV data cache has enabled fast response times and low server resource utilization with a minimal server specification. Due to the small size of the cache, no index tuning has yet been performed. CPU load spikes when the caching utility refreshes the cache, so caching will be optimized to support higher load scenarios.

6. User Interface: Data Preview, Search, and Graph Exploration

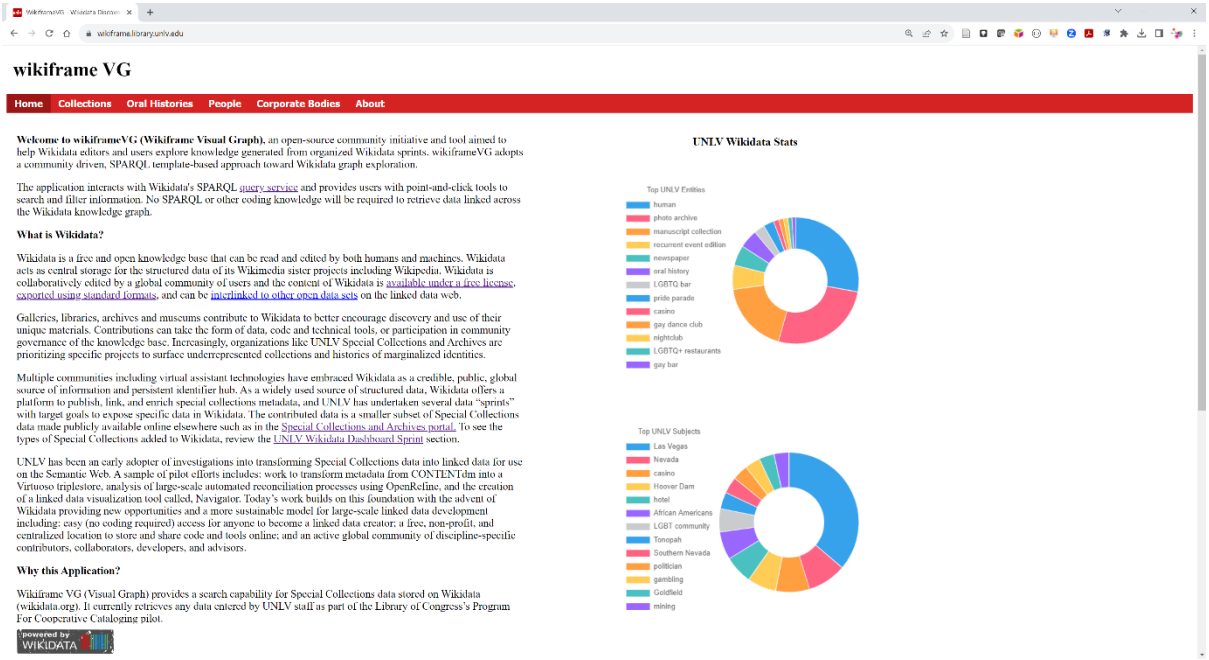

Figure 3: WikiframeVG data preview view

The current version of Wikiframe not only provides a mechanism to preview information about the Wikidata within frame, but also provides a visual way to search and explore a predefined subset of the knowledge graph. When loading the homepage, a live SPARQL query sends a request to the Wikidata Query Service, and the results are dynamically visualized on the right-side pane. As shown in Figure 3, the top ring chart provides a count of the number of *instance of (wdt:P31)* object values while the lower chart counts occurrences of the *main subject (wdt:P921)* object values. These two charts provide a clear overview of what the local Wikiframe instance contains and what it does not.

The Wikiframe user interface incorporates *exploratory search techniques* [18], where a user starts with a basic general search criterion, then from the initial results traverses the graph to nodes of interest for additional information. The interface also incorporates *faceted browsing* [19] views where the *instance of (wdt:P31)* object values are represented as dedicated classes in search pages (e.g. Collections, Oral Histories, People, Corporate Bodies) or used within a page as a facet filter in which one or more class values can be selected for search.

6.1. Collections

From the collections page, users can explore Wikidata pertaining to manuscript and photographic collections. The index powering the search box is optimized to explore archival collections by title, material type, or other descriptive details. For example, by using keywords like "correspondence", or "mining" or "papers", the results return a graph and list of collections where the collection nodes are clustered by subject. Another strategy is to use partial or full names of known collections (e.g. type "Clinton" for the Clinton Wright Photographs). On the

collections page, Wikiframe users can also use a facet filter to construct a query by selecting the checkbox next to the itemized list of subject headings.

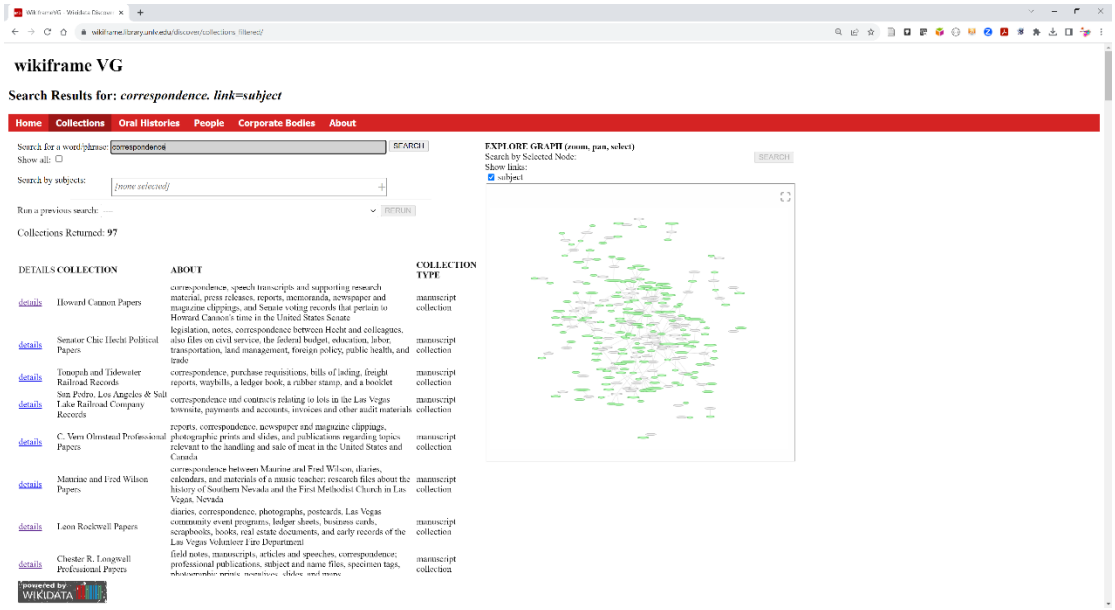

Figure 4: WikiframeVG collections page: Result list and graph rendering from keyword search

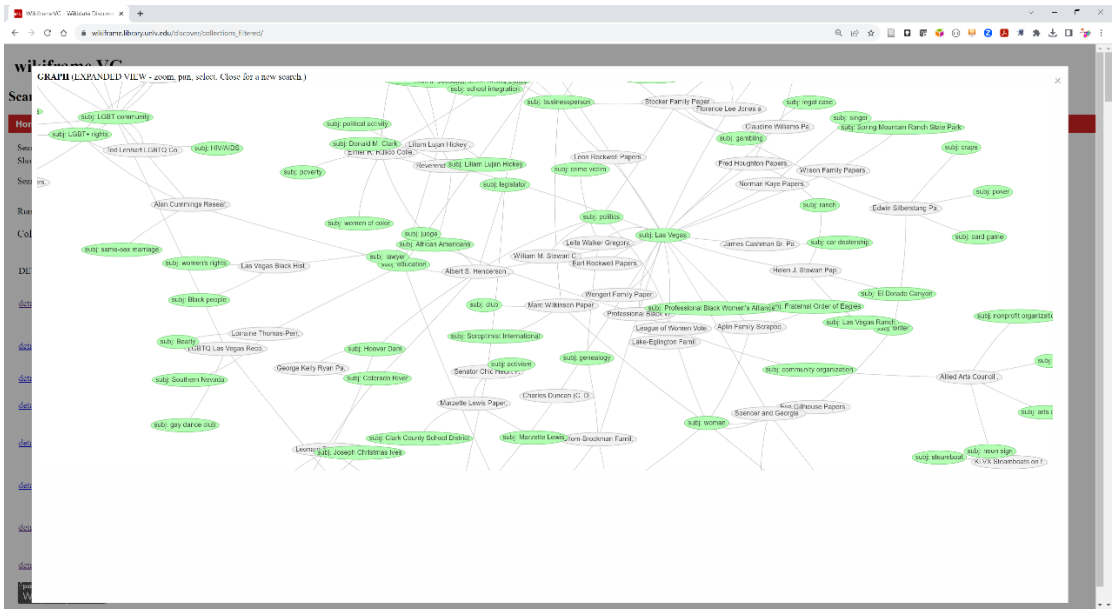

Figure 5: WikiframeVG collections page: Segment of the graph rendering in expanded view

6.2. Oral Histories

On the oral histories page, users can use the search box or facet filter to discover oral histories of interest, by selecting one or more subjects from the facet filter tool. The search box index is optimized to explore oral histories by title, a person's name, or some descriptive detail. To find an oral history based on an individual, a user can type all or part of a name, like "Hurtado" which will return multiple oral histories from Mary and Alfred Hurtado. It is also possible to combine keywords search and facet filter settings. For example, by typing the phrase "African American", then selecting one or more subjects, will return oral histories related to African Americans. It is

worth noting that some combinations of keywords and facet filter selections can at times return zero results.

6.3. People

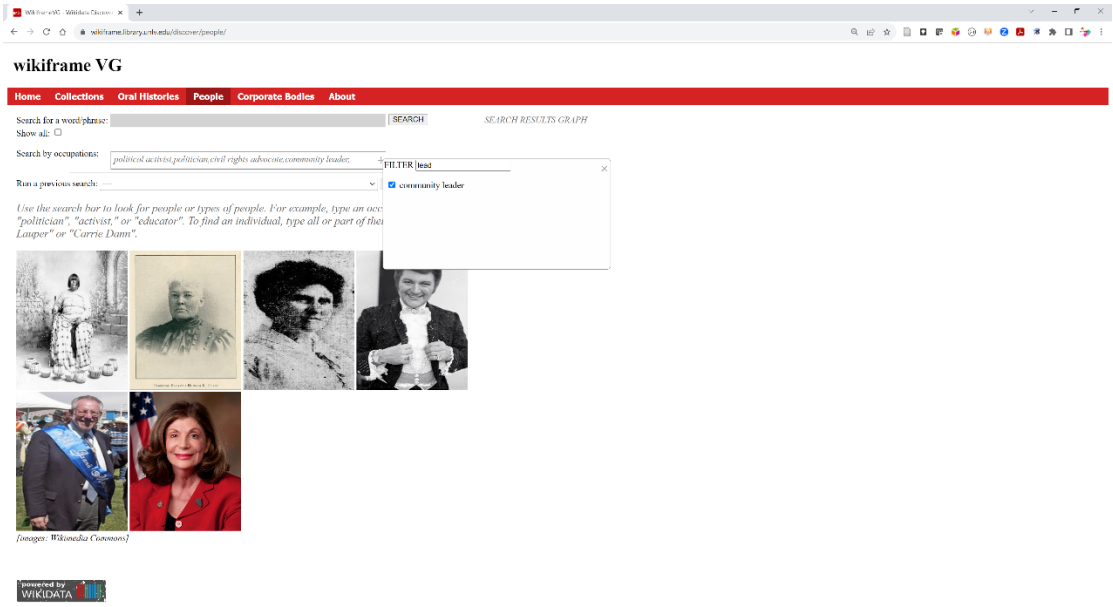

Figure 6: WikiframeVG people page: Query construction using occupation facet filter

On the people page, users can also use the search box or facet filter to explore people of interest, by selecting one or more occupations from the facet filter tool. The search index is optimized to explore people by occupation or by name. You also can combine keywords and filter settings. For example, by typing the word "Vegas", then selecting one or more occupations, will result in occupations related to Las Vegas.

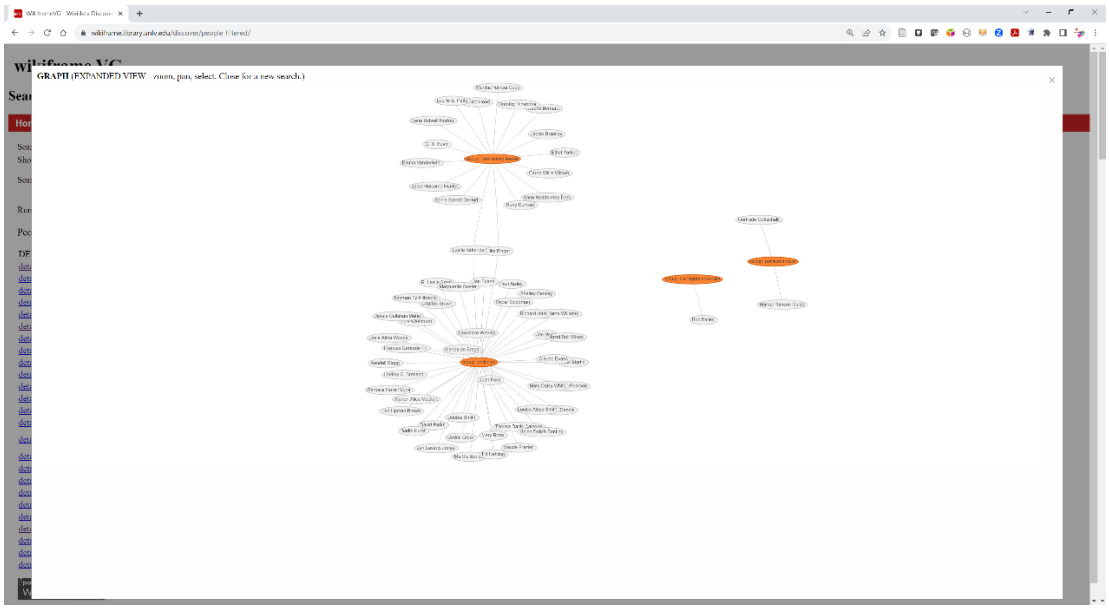

Figure 7: WikiframeVG people page: Graph rendering from query construction in expanded view

6.4. Corporate Bodies

On the corporate bodies page, users can use the search box and filter tools to look for corporate entities. The search index is optimized to explore corporate bodies by name or by type of business (i.e. *instance of (wdt:P31)*). The user also can combine keywords and facet filter settings for corporate bodies.

7. Discussion

WikiframeVG is available at: https://wikiframe.library.unlv.edu and currently retrieves only data from UNLV's Special Collections dataset. Moreno-Vega and Hogan [20] describe the challenge of retrieving results across the entire Wikidata graph from a content management and performance perspective, and they note that "investigating incremental indexing schemes would be an important practical contribution" (315). In fact, Wikiframe's frame-of-interest approach significantly curtails the need to extensively index Wikidata content by simplifying the problem of which data to index. The template construct enables asking complex questions about the Wikidata graph, but sets those key questions as constraints on the availability of data within the context they create.

The goal to visualize linked data and their relationships between special collections and archives materials has been essential to the Wikiframe project team moving this work forward. However, there are several areas of research that warrants further investigation, including limitations and opportunities, future enhancements, and governance as a community of practice.

7.1. Limitations and opportunities

Using the template structure has limitations and opportunities. The template concept raises a problem of omission: does the template in fact model all the available data for an application class? If a subset of properties and identifiers defines the class at a given time, can the application discover new properties or identifiers that might be relevant to the application scope? For instance, if a set of collections are described as *newspapers (wd:Q11032)* but newer collections are described as *publications (wd:Q732577)*, how does Wikiframe recover? In our application development process, we will investigate how the interplay of filters and new forms of description can drive a discovery process that tags in-scope Wikidata entries needing review. Outcomes might include modifications to underlying templates or adjustments to metadata descriptive practices. While metadata quality was not a primary driver in the application development, Wikiframe offers a compelling opportunity to view data in a new way and identify areas where metadata creators can work to improve overall metadata quality, for example, by observing disjointed nodes in the graph and adding properties to create links between two or more nodes.

7.2. WikiframeVG future enhancements

Wikiframe can be expanded to include views besides graph visualization and tabular results; we are particularly interested in exploring the possibilities related to geographic locations. In our case, Melvin [21] devoted a Wikidata sprint to adding Wikidata that contained information on Las Vegas-area LGBTQ people, organizations, and events, including organization and event coordinate locations. Melvin wrote SPARQL queries to showcase these collections and their relationships, showing gay businesses (gay bars, LGBTQ-friendly restaurants, coffee shops, and other community businesses) on a map of the city, raising the visibility and awareness of this

community and of UNLV's collections about it.[10] This work has become the basis for a new Wikiframe geographic visualization feature (Figure 8) that will be released in a future version.

Figure 8: WikiframeVG map prototype of corporate bodies plotted on layers organized by *instance of (wdt:P31)*

7.3. WikiframeVG as a community of practice

Wikiframe is intended to be developed collaboratively by the Wikidata community. This commitment to sharing the code for reuse and community contributions is a core principle of open-source application development. By establishing a core codebase and making the application available for others to develop additional templates, the application can be easily customized for other datasets or research communities. The goal of sharing the application openly is a key part of sustaining the work invested thus far. As more users test and interact with the tool, it can be continuously improved or modified according to community needs and priorities. Some of these enhancements may be completed by the original developer because of community feedback, while others may happen over time as different members of the community see how the tool can be used to visualize their unique data sets. The templates depend upon new data being made available on Wikidata and on the writing of new SPARQL statements; two activities that are robustly pursued within the Wikidata community. Sharing Wikiframe as a tool that can leverage this work happening in the community has resulted in strong interest and excitement as other organizations see the value of the work and want to contribute in their own ways. This distributed approach to future Wikiframe development better ensures that the project will remain active regardless of any one organization's funding or resources.

8. Conclusion

There is significant interest in the GLAMS community to harness the power of Wikidata's extensive global knowledge graph and centralized source of persistent identifiers. But to realize this potential, more organizations need tools that can overcome the barriers. These include: lack of technical expertise, lack of experience and knowledge of Semantic Web technologies, and the scarcity of available and easy to implement visualization software to fully express the richness of the data. We developed WikiframeVG with the assumption that making sense of Wikidata or relevant portions of Wikidata is a strategic project for professional disciplines and organizations,

10 https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot/UNLV#Visualize_Las_Vegas_LGBTQ+_Businesses

and we developed this proof of concept to provide a no-code approach to Wikidata graph exploration. WikiframeVG is a template-based solution that bypasses the common problems of specialized library development projects that only work for one organization's unique needs and which become obsolete quickly as the organization lacks interest in long-term sustainability. WikiframeVG is customizable, expandable, and can be collaboratively developed through community cooperation over time as user needs evolve. As more GLAM community members work with tools such as WikiframeVG and build collaborative tools, the whole community will benefit and better leverage the power of Wikidata for their researchers and users.

Acknowledgements

This research has been funded by the University of Nevada, Las Vegas. Libraries - Dean's Leadership Circle through special projects funding. Source code is made available under the Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0): https://github.com/UNLV-Libraries/wikidata-discovery-project.

References

[1] D. Melvin, C. Lampert, Ethical explorations using Wikidata and Wikidata tools to expose underrepresented special collection materials, in: A. Provo, K. Burlingame, B. M. Watson, (Eds.) Ethics in Linked Data, 1st ed., Library Juice Press, Sacramento, CA, 2023, pp. 411-435.

[2] OCLC Research Archives and Special Collections Linked Data Review Group, Archives and special collections linked data: Navigating between notes and nodes. OCLC Research, Dublin, OH, 2020. doi:10.25333/4gtz-zd88.

[3] J. Godby, K. Smith-Yoshimura, B. Washburn, K. Davis, K. Detling, C. Fernsebner Eslao, S. Folsom, X. Li, M. McGee, K. Miller, H. Moody, H. Tomren, C. Thomas, Creating library linked data with Wikibase: Lessons learned from Project Passage. OCLC Research, Dublin, OH, 2019. doi:10.25333/faq3-ax08.

[4] Association of Research Libraries, ARL white paper on Wikidata: Opportunities and recommendations. ARL, Washington D.C., 2019. URL: https://www.arl.org/wp-content/uploads/2019/04/2019.04.18-ARL-white-paper-on-Wikidata.pdf

[5] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57(10) (2014) 78-85. doi:10.1145/2629489.

[6] J. Neubert, Wikidata as a linking hub for knowledge organization systems?: Integrating an authority mapping into Wikidata and learning lessons for KOS mappings, in: Proceedings of the 17th European Networked Knowledge Organization Systems Workshop co-located with the 21st International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017), CEUR-WS.org, Thessaloniki, Greece. 2017, pp. 14-25. URL: https://ceur-ws.org/Vol-1937/paper2.pdf

[7] J. Neubert, Linking knowledge organization systems via Wikidata, in Proceedings of the 2018 International Conference on Dublin Core and Metadata Applications. DCMI, Porto, Portugal. 2018, pp. 1-3. doi:10.5555/3308533.3308534.

[8] L. Feigenbaum, G. T. Williams, K. G. Cark, E. Torres, SPARQL 1.1 Protocol (W3C Recommendation). URL: https://www.w3.org/TR/sparql11-protocol/

[9] K. Smith-Yoshimura, Analysis of international linked data survey for implementers, D-Lib Magazine 22(7/8), (2016) doi:10.1045/july2016-smith-yoshimura.

[10] K. Smith-Yoshimura, Analysis of 2018 international linked data survey for implementers, Code4Lib Journal 42, (2018) URL: https://journal.code4lib.org/articles/13867

[11] J. Helmich, T. Potoček, J. Klímek, M. Nečaský, Towards easier visualization of linked data for lay users, in: Proceedings of the 12th International Conference on World Wide Web, ACM, Budapest, Hungary. 2017, pp. 700-709. doi:10.1145/3102254.3102261.

[12] M.C. Schraefel, D. Karger, The pathetic fallacy of RDF, in: International Workshop on the Semantic Web and User Interaction. Springer Berlin, Heidelberg, Athens, GA. 2006. URL: https://eprints.soton.ac.uk/262911/1/the_pathetic_fallacy_of_rdf-33.html
[13] D. Melvin, What came first the SPARQL or the egg?: Managing SPARQL informed Wikidata productions for special collections and archives. URL: https://osf.io/zq25v/
[14] Wikidata Pilot. PCC Identity Management Home, https://wiki.lyrasis.org/display/pccidmgt/Wikidata+Pilot
[15] J. Klímek, J. Helmich, M. Necaský, Use cases for linked data visualization model, in Proceedings of the Workshop on Linked Data on the Web, LDOW 2015, Co-Located with the 24th International World Wide Web Conference (WWW 2015), CEUR-WS.org, Florence, Italy, 2015, pp. 1–2. URL: https://ceur-ws.org/Vol-1409/paper-08.pdf
[16] J. Helmich, T. Potoček, J. Klímek, M. Nečaský: Towards easier visualization of linked data for lay users, in: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, ACM, Amantea, Italy, 2017, pp. 1-9. doi:10.1145/3102254.3102261
[17] 2021 LD4 Conference, 2021. URL: https://sites.google.com/stanford.edu/2021ld4conf/home
[18] R. Mirizzi, A. Ragone, T. Di Noia, E. Di Sciascio, Semantic Wonder Cloud: Exploratory search in DBpedia, in: F. Daniel, F. M. Facca (Eds.), Current Trends in Web Engineering. ICWE 2010. Lecture Notes in Computer Science, volume 6385. Springer, Berlin, Heidelberg, 2010, pp. 138-149. doi: 10.1007/978-3-642-16985-4_13.
[19] Y. Tzitzikas, N. Manolis, P. Papadakos, Faceted exploration of RDF/S datasets: a survey, Journal of Intelligent Information Systems 48, (2017) 329-364. doi: 10.1007/s10844-016-0413-8.
[20] J. Moreno-Vega, A. Hogan, GraFa: Scalable faceted browsing for RDF graphs, in Proceedings of the 17th International Semantic Web Conference, Springer. Monterey, CA, 2018, pp. 301-317. doi:10.1007/978-3-030-00671-6_18.
[21] D. Melvin, Proactive strategies to improve underrepresentation in public knowledge graphs: A Wikidata sprint in UNLV Special Collections highlighting LGBTQ+ Las Vegas. URL: https://osf.io/2ge9h/