Wikidata in the GLAM sector: a knowledge sharing approach

Gustavo Candela
University of Alicante

Abstract

Wikidata has been playing a leading role in the last years in GLAM for several purposes: i) enrichment of digital collections; ii) data quality assessment; and iii) to provide examples of reuse and rich visualizations. In order to contribute and advance the Wikimedia Movement towards the 2030 strategic direction, this research proposes a systematic review of Wikidata in the GLAM sector according to four dimensions: reuse, enrichment, extraction and data quality. This work intends to identify best practices and guidelines for GLAM institutions that can be extended to other domains.

Introduction

GLAM Institutions have been making available their collections for decades. Recently, new approaches have emerged to foster the reuse and adoption of computational access such as Labs and Collections as data [1,2,9]. These encourage the use of Wikidata to enrich their content as well as the use of the Semantic Web and Linked Data. Some examples include:

- Rijksmuseum [3]
- https://data.bnf.fr/
- https://id.loc.gov/
- National Library of Scotland [4]
- https://data.cervantesvirtual.com

However, there is still room for improvement in several aspects such as the adoption of the role of Wikidata librarian in GLAM, the use of Wikidata for data quality assessment purposes and the reuse of the metadata to create rich visualizations and share reproducible code [5,6].

In this research we propose a systematic review of Wikidata in the GLAM sector according to four dimensions: reuse, enrichment, extraction and data quality. This work intends to identify best practices and guidelines for GLAM institutions regarding the use of Wikidata that can be extended to other domains. It seeks to contribute and advance the Wikimedia Movement towards the 2030 strategic direction.

This work will start on September 1, 2024 and will conclude by June 30, 2025.

Related work

Recent initiatives such as the International GLAM Labs Community and Collections as data have promoted the use of Wikidata in the GLAM sector to enrich the digital collections [1,2]. Digital Humanities have also tackled the use of Wikidata such as the case of DARIAH in Europe and the Asociación de Humanidades Hispánicas in Spain [5]. In addition, these promote the adoption of computational methods and the publication of reproducible code in the form of Jupyter Notebooks [6]. Previous works have been based on gathering together a selection of researchers from several backgrounds for writing sprints such as the IMPACT 10th

Anniversary Workshop in Alicante and Collections as Data in Vancouver.

These efforts provide an extensive demonstration of how Wikidata can be used in the GLAM sector. Nevertheless, to the best of our knowledge, none of this research uses a systematic review approach to identify best practices for the reuse and enrichment of Wikidata. This work can be useful to encourage GLAM institutions to adopt best practices and guidelines.

Methods

The method proposed to describe Wikidata as a knowledge service in the GLAM sector works in four steps that are described below. Figure 1 shows an overview of the method.

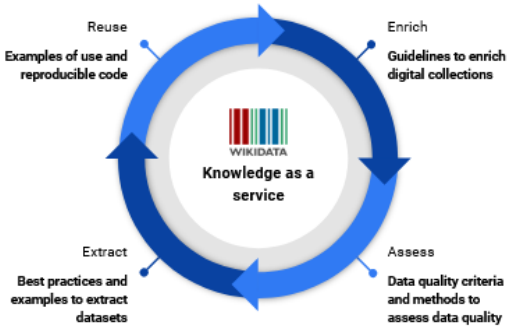

Figure 1. Overview of the method proposed in this work to describe Wikidata as a knowledge service in the GLAM sector.

Reuse. This step will review visualization examples based on the Wikidata SPARQL endpoint. It will also provide a selection of Jupyter Notebooks illustrating how to access and retrieve metadata from Wikidata [6,7].

Enrich. This step will analyze and describe the methods to enrich digital collections by means of Wikidata [2,4,6].

Assess. This step will describe existing methods to assess data quality based on Wikidata. Some examples are based on SPARQL and Shape Expressions [8].

Extract. This step will describe how to extract and create a dataset from Wikidata using SPARQL commands such as CONSTRUCT [2].

Expected output

These are the outputs expected for this research:

- 1 scientific publication for the research community (e.g., Journal of Information Science or Journal of Open Humanities Data)
- 1-2 Conferences (e.g., EuropeanaTech, TPDL, WikidataCon, DATeCH, DARIAH Annual Event)
- *Wikidata and GLAM Statements* for the community and GLAM institutions
- Wikidata in GLAM workshop (in person). The participants will include 8 to 10 librarians, data scientists and digital humanities researchers working in the GLAM sector.

Risks

Some potential risks to perform this project are listed below.

- Unequal representation for the participants of the workshop. Attendees will be personally invited to participate considering a balance in field of expertise, gender, etc.
- Short list of participants for the workshop. The participation in the workshop will be ensured by the wide network of contacts in GLAM institutions.
- External hazards risk such as a pandemic for the workshop. In this

case, the workshop will be organized in remote mode.

Community impact plan

This research will be disseminated and promoted in several multilingual Wikimedia and GLAM channels:

- Wikimedia Spain
- International GLAM Labs Community
- Collections as data initiative
- Impact Centre of competence - https://www.digitisation.eu/
- Biblioteca Virtual Miguel de Cervantes

Please, note that the author is a member of Wikimedia Spain. This plan demonstrates a commitment to three relevant areas in the GLAM sector of computational access, reuse and publication of digital collections, reflecting alignment of community interests.

Evaluation

The evaluation of this research will be based on: i) the publication and availability for the community of the outputs proposed; ii) the number of people attending the workshop; iii) attendance to international conferences as a speaker to present the results of this work; and iv) the dissemination of the work through the channels proposed (e.g., number of blog posts). In addition, the research publication will also be reviewed by the journal editors and reviewers.

Budget

A total amount of USD 43,125:

- 25,000 *Wikidata in GLAM* workshop
- 2,500 research publication as open access
- 5,000 conferences
- 5,000 equipment
- 5,625 institutional overhead

Prior contributions

- Research articles [2,6,7,8]
- Attendance as a speaker to WikidataCon 2017 and 2019
- #DatatónCervantes event - Wikimedia Spain
- #Wikihackatón event - Wikimedia Spain
- Wikidata and libraries workshop
- LOD and SPARQL workshop
- Wikidata workshop - Wikimedia UK and the National Library of Scotland

References

[1] Mahey, M., Al-Abdulla, A., Ames, S., Bray, P., Candela, G., Chambers, S., Derven, C., Dobreva-McPherson, M., Gasser, K., Karner, S., Kokegei, K., Laursen, D., Potter, A., Straube, A., Wagner, S-C. and Wilms, L. with forewords by: Al-Emadi, T. A., Broady-Preston, J., Landry, P. and Papaioannou, G. (2019) *Open a GLAM Lab.* Digital Cultural Heritage Innovation Labs, Book Sprint, Doha, Qatar, 23-27 September 2019.

[2] Candela, G., Gabriëls, N., Chambers, S., Dobreva, M., Ames, S., Ferriter, M., Fitzgerald, N., Harbo, V., Hofmann, K., Holownia, O., Irollo, A., Mahey, M., Manchester, E., Pham, T.-A., Potter, A. and Van Keer, E. (2023), "A checklist to publish collections as data in GLAM institutions", *Global Knowledge, Memory and Communication*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/GKMC-06-2023-0195

[3] Chris Dijkshoorn, Lizzy Jongma, Lora Aroyo, Jacco van Ossenbruggen, Guus Schreiber, Wesley ter Weele, Jan Wielemaker: The Rijksmuseum collection as Linked Data. *Semantic Web* 9(2): 221-230 (2018). https://doi.org/10.3233/SW-170257

[4] National Library of Scotland. (2023). The National Librarian's Research Fellowship in Digital Scholarship 2022-23. Data Foundry. https://data.nls.uk/projects/the-national-librarians-research-fellowship-in-digital-scholarship-2022-23/

[5] Asociación de Humanidades Digitales Hispánicas. (2023). *Compartir Pantalla: Linked Open Data y SPARQL*. https://humanidadesdigitaleshispanicas.es/compartir-pantalla-con-gustavo-candela-14-de-noviembre-de-2023/

[6] Candela, G., Chambers, S., & Sherratt, T. (2023). An approach to assess the quality of Jupyter projects published by GLAM institutions. *Journal of the Association for Information Science and Technology*, *74*(13), 1550–1564. https://doi.org/10.1002/asi.24835

[7] Candela, G., Sáez, M. D., Escobar Esteban, Mp., & Marco-Such, M. (2022). Reusing digital collections from GLAM institutions. *Journal of Information Science, 48*(2), 251-267. https://doi.org/10.1177/0165551520950246

[8] Candela, G. (2023). An automatic data quality approach to assess semantic data from cultural heritage institutions. *Journal of the Association for Information Science and Technology*, *74*(7), 866–878. https://doi.org/10.1002/asi.24761

[9] Padilla, T., Scates Kettler, H., Varner, S., & Shorish, Y. (2023). Vancouver Statement on Collections as Data. Zenodo. https://doi.org/10.5281/zenodo.8342171