Using Wikidata in the European Literary Bibliography: A Reproducible Approach

GLAM institutions (Galleries, Libraries, Archives, and Museums) have been exploring new ways to make available their digital collections. Wikidata has emerged as a leading approach with which to enrich their digital collections [1]. In parallel, new trends such as Labs and Collections as data promote the publication of digital collections suitable for computational use [2] as well as the use of reproducible code in the form of Jupyter Notebooks [3].

The European Literary Bibliography (ELB) is a project of the Institute of Czech Literature (Czech Academy of Sciences) and the Institute for Literary Research (Polish Academy of Sciences). It intends to open bibliographic data for literary studies at the European level holding resources from several institutions.

Following the Collections as data principles and focusing on the ELB, this work provides a reproducible framework including several steps for publishing and reusing digital collections based on literary bibliographies made available by GLAM institutions [4]. It also presents a collection of DH research scenarios to show how data can be explored and reused. This work is the result of an ATRIUM Transnational Access Scheme Grant. The results are available in the form of a repository of reproducible code.[1]

A reproducible framework to transform bibliographic metadata to Collections as data

This section presents the framework to publish and reuse digital collections in the form of Collections as data [4], as shown in Figure 1.

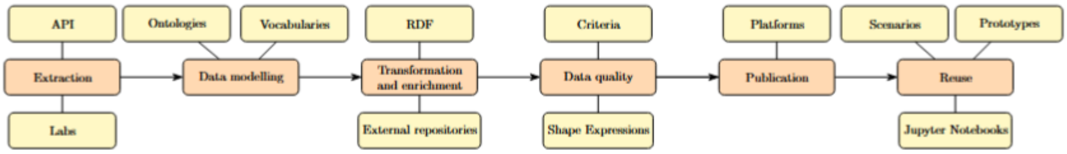

Figure 1: Reproducible framework for publishing and reusing collections as data.

Data extraction refers to the selection of data relevant to a specific topic (e.g., author, organization or theme). *Data modelling* aims at ensuring machine-readable bibliographic metadata, using ontologies and vocabularies. The *transformation and enrichment* step refers to the transformation of the data into Linked Open Data (LOD) using RDF to describe metadata as triples as well as the use of Wikidata to enrich the metadata. The *data quality* step ensures the high quality of the RDF data. The publication requires the inclusion of additional documentation including aspects such as provenance and licensing. Finally, the published datasets can be *reused* in various ways (e.g., prototypes or research scenarios defined by DH scholars).

Defining research scenarios

[1] https://gitlab.pcss.pl/dl-team/atrium/tna01/-/tree/candela?ref_type=heads

After applying the proposed framework to the ELB and exploring new uses of the data, a selection of research scenarios were defined to illustrate data reuse and integration using Wikidata as a main repository with which to enrich the metadata: i) comparative analysis of provincial vampire novels in Spain; ii) republican writers who emigrated during the Spanish Civil War; and iii) geographical distribution of publications about specific Spanish writers. Limitations were identified in terms of scope and completeness to meet researchers' needs

Conclusions

This work advances the publishing of digital collections in computationally usable forms describing how Wikidata can be used to explore new ways of analysis of the data. Future research directions include extending and implementing the research scenarios, and applying and adapting the framework to other domains.

Bibliography

[1] Candela, G., Cuper, M., Holownia, O. *et al*. A Systematic Review of Wikidata in GLAM Institutions: a Labs Approach. TPDL (2) 2024: 34-50

[2] Candela, G., Gabriëls, N., Chambers, S., *et al*. (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] Candela, G., Chambers, S., Sherratt, T. An approach to assess the quality of Jupyter projects published by GLAM institutions. J. Assoc. Inf. Sci. Technol. 74(13): 1550-1564 (2023)

[4] Candela, G., Rosiński, C., & Margraf, A. (2024). A reproducible framework to publish and reuse Collections as data: the case of the European Literary Bibliography. https://doi.org/10.5281/zenodo.14106707