Using Wikidata in the European Literary Bibliography: A Reproducible Approach

Gustavo Candela | University of Alicante | Spain | gcandela@ua.es
Cezary Rosiński | IBL PAN | Poland | cezary.rosinski@ibl.waw.pl
Arkadiusz Margraf | PSNC | Poland | margraf@man.poznan.pl

ATRIUM

INTRODUCTION

GLAM institutions (Galleries, Libraries, Archives, and Museums) are increasingly making their resources available as machine-readable "Collections as Data." They leverage diverse materials—texts, images, postcards, metadata, bibliographic databases, audio, video, maps—and modern techniques (network analysis, Named-Entity Recognition, Entity Linking) to facilitate reuse, including applications in NLP and Machine Learning. Initiatives such as digital Labs and "Collections as Data" projects, along with community groups like the Bibliographical Data Working Group within DARIAH, promote open licensing and collaboration between data curators and researchers.

KEY GLAM AREAS

- Standardized documentation & checklists
- Semantic Web & Linked Open Data
- Bibliographic initiatives & data dumps
- Quality assessment & tooling

BACKGROUND

The **European Literary Bibliography** (ELB, literarybibliography.eu) is an ecosystem for processing, integrating, enriching, presenting, and sharing open literary bibliographical datasets internationally. It is developed jointly by Czech Literary Bibliography and Polish Literary Bibliography. ELB is open to bibliographers, data curators, data experts, and lay users, and intends to enhance the understanding and creative exploration of the European literary landscape, promoting literary metadata integration and reuse.

The **ATRIUM project** (Advancing fronTier Research In the arts and hUManities, atrium-research.eu) offers **Transnational Access** (TNA) scheme. This initiative is designed to support Arts and Humanities researchers by providing access to expert knowledge, mentorship, and tools from leading Data Management organisations.

FRAMEWORK

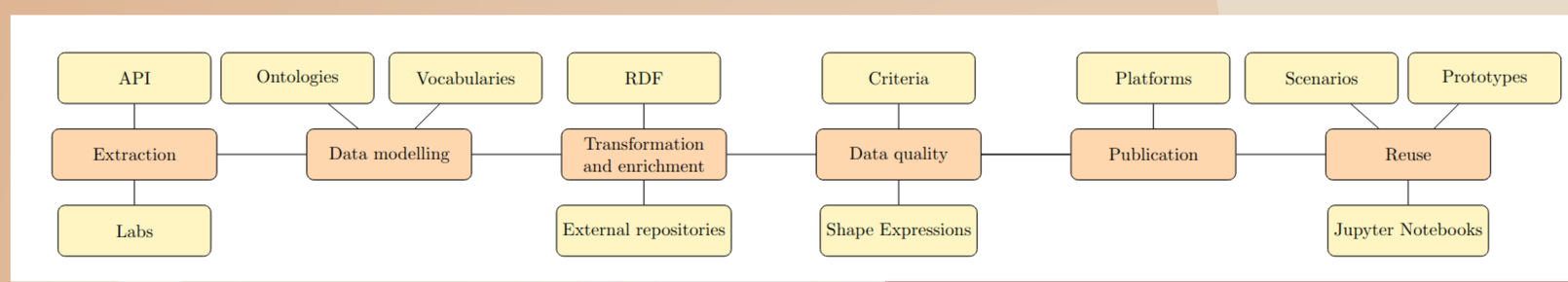

Figure 1. Framework employed to transform bibliographic metadata into Collections as Data.

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., data processing via Jupyter Notebooks and prototypes or research scenarios defined by DH scholars).

WIKIDATA ENRICHMENT

```
SELECT ?place ?placeLabel (SAMPLE(?image)
as ?img) (SAMPLE(?coordinates) as ?c)
WHERE {
 VALUES ?place { wd:Q104740 wd:Q1085
wd:Q110732 wd:Q11743615 wd:Q126781
wd:Q1881080 wd:Q208113 wd:Q211 wd:Q213
wd:Q239 wd:Q27509 wd:Q320007 wd:Q3491317
wd:Q414 wd:Q48175 wd:Q48214 wd:Q503146
wd:Q52898 wd:Q54187 wd:Q554293 wd:Q5622
wd:Q580 wd:Q595937 wd:Q7264994 wd:Q7943
wd:Q796 wd:Q81720 wd:Q992834 wd:Q103217
wd:Q183 wd:Q22693876 wd:Q270 wd:Q421819
wd:Q5690 wd:Q696835 wd:Q761 wd:Q8079842
wd:Q38 wd:Q39 wd:Q43 wd:Q111653833
wd:Q16957 wd:Q211274 wd:Q212 wd:Q268
wd:Q4472076 wd:Q90 wd:Q142 wd:Q184
wd:Q1401995 wd:Q31487 wd:Q159 wd:Q7835
wd:Q37 wd:Q36 }.

 ?place wdt:P625 ?coordinates.
 OPTIONAL {?place wdt:P18 ?image}.

 SERVICE wikibase:label { bd:serviceParam
wikibase:language "en" }
} GROUP BY ?place ?placeLabel ?img
```

Figure 2. SPARQL query on the place data of the Adam Mickiewicz dataset.

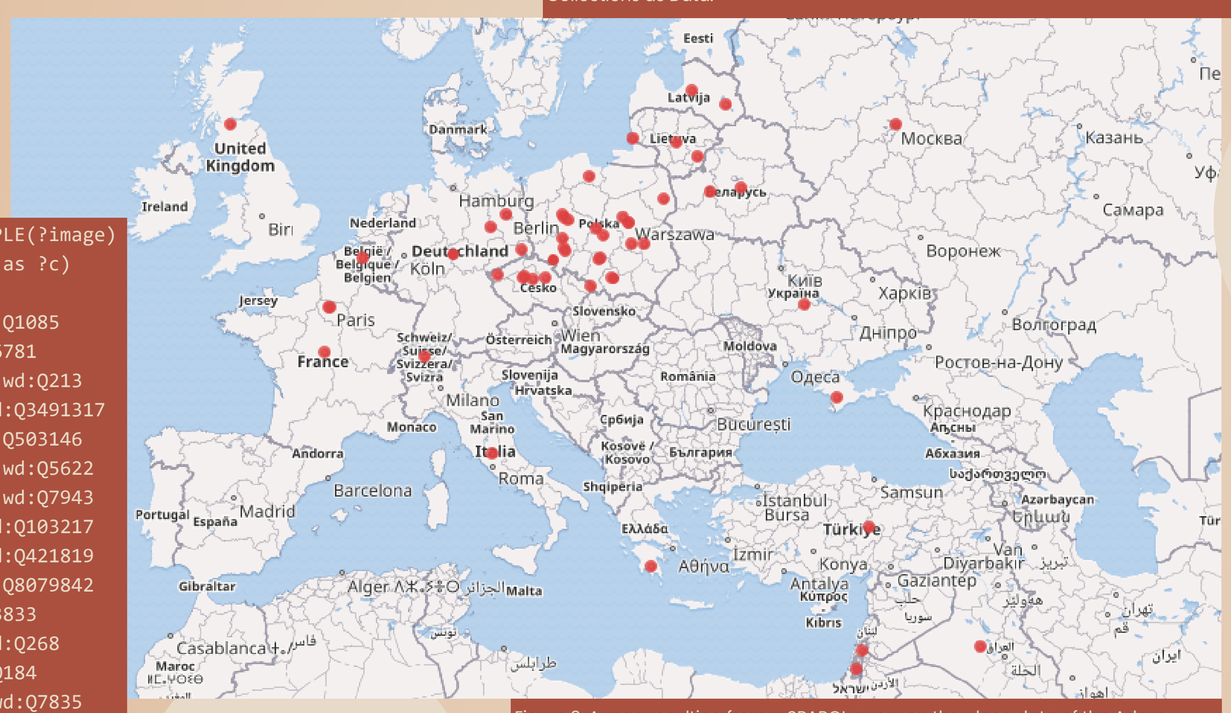

Figure 3. A map resulting from a SPARQL query on the place data of the Adam Mickiewicz dataset

REFERENCES

- 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
- Candela, G., Gabriëls, N., Chambers, S., Pham, T.-A., Ames, S., Fitzgerald, N., Hofmann, K., Harbo, V., Potter, A., Ferriter, M., Manchester, E., Irollo, A., Keer, E. V., Mahey, M., Holownia, O., & Dobreva, M. (2023). A Checklist to Publish Collections as Data in GLAM Institutions. Global Knowledge, Memory and Communication. https://doi.org/10.1108/GKMC-06-2023-0195
- Candela, G., Rosiński, C., & Arkadiusz, M. (2025). ATRIUM Transnational Access Scheme - A reproducible framework to publish and reuse Collections as Data: The case of the European Literary Bibliography. https://zenodo.org/records/15075805
- Chambers, S., Walsh, M., Caswell, M., Harder, G., Okumura, M., Corrin, J., Baeza Ventura, G., Antonijevic, S., Knazook, B., Narlock, M., Bailey, J., Neudecker, C., Downie, J. S., Layne-Worthey, G., van Strien, D., Irollo, A., Whitmire, A., Lee, J., Berry, D., … Ridge, M. (2023). Position Statements -> Collections as Data: State of the field and future directions. https://doi.org/10.5281/zenodo.7897735
- Vimr, O., & Rosiński, C. (2024). Building a European Literary Bibliography. Challenges and Strategies in Integrating and Presenting Multilingual Data. https://doi.org/10.5281/zenodo.11454445

RESEARCH SCENARIOS

The ELB exemplifies how **GLAM data** can be integrated with **scholarly communication**. The research scenarios illustrate how the proposed framework can be applied to address specific challenges faced by GLAM institutions when their data are used for **scientific purposes**. Scenarios focus on **linking bibliographical resources with external authority databases**, such as Wikidata, to increase the possibility of inferring and viewing data. These are the 3 scenarios that we propose in this work:

1. Comparative analysis of provincial Spanish novels featuring vampires
2. Republican writers who migrated during the Spanish Civil War
3. Geographical distribution of publications dedicated to particular Spanish writers

RESULTS

Our main contributions are:

1. A **framework** to publish and reuse Collections as Data based on bibliographic metadata
2. The **results** obtained using the proposed infrastructure, including **reproducible code**
3. A **selection of DH research scenarios** illustrating how the data could be explored and reused.

These contributions are intended to leverage the adoption of Collections as Data by GLAM institutions providing detailed documentation and code, as well as additional use cases based on bibliographic metadata.

Dataset	Classes	Properties	No. triples	No. external links
Adam Mickiewicz	7	19	47,631	863
Miguel de Cervantes	5	19	44,389	749
National Library of Spain	7	19	10,515,016	67,563

Table 1. Results obtained by applying the framework to selected datasets from the European Literary Bibliography

VISIT OUR REPOSITORY

CC BY

Funded by the European Union

European Literary Bibliography
Universitat d'Alacant
Universidad de Alicante
DIGITAL HUMANITIES CENTRE
PSNC