30 January 2024

Building the European Literary Bibliography: Challenges and Strategies in Integrating and Presenting Multilingual Data

Short paper abstract

Ondřej Vimr, Institute of Czech Literature, Czech Academy of Sciences,
https://orcid.org/0000-0002-9364-0685, vimr@ucl.cas.cz
Cezary Rosiński, Institute of Literary Research, Polish Academy of Sciences,
https://orcid.org/0000-0002-6136-7186, cezary.rosinski@ibl.waw.pl

Library catalogues and literary bibliographies have been identified as essential resources for studying various facets of book production, from literature to intellectual history and informatics (Lahti et al. 2019). However, the bibliographical metadata landscape presents challenges, characterized by diverse data producers each with their unique standards and goals (Umerle et al. 2022). Data integration and reuse for research necessitate a structured approach. This paper discusses the strategies and challenges in creating the European Literary Bibliography (https://literarybibliography.eu/en/), an ongoing international project aimed at establishing an ecosystem for processing, integrating, enriching, presenting, and sharing multilingual bibliographical datasets. This ecosystem, open to bibliographers, data curators, data experts, and lay users, intends to enhance the understanding and creative exploration of the European literary landscape, promoting literary metadata integration and reuse.

The Problem

Many library catalogues and digital bibliographies share the MARC21 cataloguing standard but are usually curated and stored independently. Integrating and reusing these isolated datasets requires addressing challenges related to data availability, licensing, quality, completeness and more. Although the data might seem open and accessible at first glance, various issues arise during its reuse. For example, each data producer follows their standards, which are not always in line with the FAIR principles, and their datasets are often poorly documented, lack semantic links, and display national variations in the usage of the MARC21 format, which is inherently complex.

Essential for analyzing extensive international humanities and social science data, bibliographic data is inherently multilingual and transnational. Data from national libraries spans centuries and

multiple languages, presenting challenges like language dependencies and deduplication when merging datasets from different countries.

Moreover, enriching traditional bibliographical data with external sources like Wikidata or geotagging services is appealing and rational. These additions offer new perspectives and contextualize the data geographically or sociologically and are particularly helpful for data presentation. At the same time, however, they demand the development of further services due to the limited straightforward integration paths in current datasets.

The Approach

The paper will detail the workflow and the current state of the European Literary Bibliography project, including bibliographical data aggregation, data harmonization, data enrichment, data LOD-ification, and data presentation.

Data aggregation involves collecting data from official sources via APIs or data dumps and selecting literary materials based on criteria like the Universal Decimal Classification. Data harmonization primarily involves standardizing the data format to MARC21 and using authority databases to correctly match and assign persistent identifiers (PIDs) for entities like individuals and institutions. An important aspect of harmonization is the automatic translation of data to enable aggregation at the terminology level, facilitating the integration of multilingual datasets.

Data enrichment involves using PIDs to gather information not originally included in bibliographic records. This is achieved, utilizing databases such as VIAF, Wikidata, Geonames, and ISSN, while focusing on both structured identifiers and unstructured strings. The next step, data LOD-ification, leverages Linked Open Data to extract additional information, often beyond the scope of a field bibliography, such as relevant dates and places for persons or the location of institutions. Linked Open Data also enables the production of name labels in all interface languages facilitating data presentation. It is important to remember that for presentation purposes the harmonised data have to be supplemented by textual information in the appropriate target language.

Finally, data presentation builds on all the previous steps, allowing for statistical analysis of the bibliographic data, visualization of relationships among people, institutions, and places, and mapping of bibliographic information onto interactive maps through geotagging. This invites users to explore the data from various perspectives, in line with the growing interest in the digital, spatial, and sociological aspects of literary production and consumption.

Currently, the European Literary Bibliography holds data from four national libraries and two bibliographical collections across Europe (Czech, Polish, Finnish, and Spanish) as well as the

Dialnet scientific database. It contains over 4.5 million bibliographic records, information on over 140,000 individuals, 15,000 places and more.

Bibliography

Lahti, Leo, Jani Marjanen, Hege Roivainen, a Mikko Tolonen. 2019. „Bibliographic Data Science and the History of the Book (c. 1500–1800)". *Cataloging & Classification Quarterly* 57(1): 5–23. https://doi.org/10.1080/01639374.2018.1543747

Umerle, T., Colavizza, G., Herden, E., Jagersma, R., Király, P., Koper, B., Lahti, L., Lindemann, D., Łubocki, J. M., Malínek, V., Milanova, A., Péter, R., Rißler-Pipka, N., Romanello, M., Roszkowski, M., Siwecka, D., Tolonen, M., & Vimr, O. (2022). *An Analysis of the Current Bibliographical Data Landscape in the Humanities. A Case for the Joint Bibliodata Agendas of Public Stakeholders*. Zenodo. https://doi.org/10.5281/zenodo.6559857