Remediations of Polish Literary Bibliography: Towards a Lossless and Sustainable Retro-Conversion Model for Bibliographical Data

Maciej Maryl, Piotr Wciślik

Greenblatt, S. J. (1991). *Marvelous Possessions. The Wonder of the New World*. Chicago: Chicago University Press.
Irving, D. R. M. (2010). *Colonial Counterpoint: Music in Early Modern Manila*. Oxford, New York: Oxford University Press.
León-Portilla, M. (ed.) (2011). *Cantares mexicanos*, 2 vols. Mexico: UNAM, Fideicomiso Teixidor.
Schmidt, L. E. (2000). *Hearing Things: Religion, Illusion, and the American Enlightenment*. Cambridge (Massachusetts): Harvard University Press.
Smith, B. R. (2002). How Sound is Sound History? A Response to Mark Smith, *The Journal of the Historical Society*, **2**(3-4): 307-15.
Sterne, J. (ed). (2012). *The Sound Studies Reader*. London, New York: Routledge.
Tomlinson, G. (1995). Ideologies of Aztec song, *Journal of the American Musicological Society*, **48**(3): 343-79.

Remediations of Polish Literary Bibliography: Towards a Lossless and Sustainable Retro-Conversion Model for Bibliographical Data

Maciej Maryl
maciej.maryl@ibl.waw.pl
Institute of Literary Research of the Polish Academy of Sciences, Poland

Piotr Wciślik
piotr.wcislik@ibl.waw.pl
Institute of Literary Research of the Polish Academy of Sciences, Poland

Remediation 1.0.: "Printed database"

Polish Literary Bibliography (PBL) is a specialized bibliography which aims to map the totality of literary and cultural life in postwar Poland. It references primarily literary works and literary scholarship, however its entries also cover the related literary critique, adaptations, theatre performances, cinematography, radio and television broadcasts, as well associated events such as conferences or awards. At the heart of PBL lies its subject classification which orders the entries to reflect the domains, hierarchies and entities of Polish literary world, i.e. its ontology in the classical sense. PBL has been developed since 1954 and today covers the period 1944-2001. For most of its history it has existed in print, however since 2000 the data has been collected in the existing digital database which currently covers the period 1988-2001 what gives app. 600 000 records.

The vicissitudes of PBL remediations could be accurately captured through an urban planning metaphor. What definitely strikes every visitor to a large Moroccan city is a great contrast between *medina*, the traditional old town with centuries-long history, and *Ville Nouvelle*, new district built under the French Protectorate in the first half of the 20th century. The former reminds a maze with endless narrow streets, and buildings which are stuck densely next to each other with no visible order, whereas the latter is the essence of modern architecture with wide boulevards, large buildings and streets laid out in a grid pattern.

The current online database is quite exemplary for early bibliographical and cataloguing projects (in Poland as elsewhere) in that it is geared towards remediating the print form of the PBL instead of taking advantage of the new medium (cf. Antelman, Lynema and Pace 2006, 128). It is a tailor-made relational database developed in Oracle whose data model is built on a plethora of dataspaces for different types of records, accompanied by various catalogues of creators, contributors, associated institutions and subject headings. The former set reflects PBL's main entities: literary works in monographs and journals, adaptations in cinematography, radio and television and associated events. The latter represents an early digital take on the index card catalog, the traditional tool of the bibliographer. Furthermore each record has a special markup in order to assure that its display at the frontend follows the structure of the paper edition.

The result of this remediation is a *medina*-like database, very rich and complicated but not fit for modern uses. It makes perfect sense for people who built it, yet at the same time it is difficult to navigate by those lacking the local knowledge - be it a human or the machine. As it often happens with relational databases,[1] it does not comply with any of the common standards in terms of record structure or data formats, what eventually leads to serious problems with both preservation and interoperability of collected data.

Towards remediation 2.0

The aim of the research project we are currently pursuing (*Polish Literary Bibliography – a knowledge lab on contemporary Polish culture*) is to reestablish the PBL database project on Linked Open Data principles for its better reuse within and beyond the bibliographic domain (see e.g. Roszkowski 2013; Coyle 2010). However, we want to do better than the French colonizers of Morocco. The modernisation of PBL will be reflexive insofar as it will reconcile the OWL and the PBL's unique ontology of the literary world expressed through the structure of its entries and metadata. The main task of the current phase of the project is development and application of the new data model. This task involves (1) the choice of vocabularies

and ontologies and (2) rendering of the subject classification structure.

(1) Vocabularies and ontologies (in the narrow sense used in information science) are needed to disambiguate the RDF triples (subject-predicate-object expressions). Here we need to balance two criteria. First the vocabularies and ontologies must enable widest possible sharing in the data cloud. Second they must be granular and complex enough in order to reflect the PBL data model, since adding too many heterogeneous elements would be counterproductive. The above applies to both metadata elements and their values.

Whereas the choice of value vocabularies was rather straightforward, using the geonames and Virtual International Authority File (VIAF) for disambiguating geographical, personal and corporate names, the choice of the meta-ontology,[2] or the vocabulary describing the metadata elements of the current PBL data model was much more difficult. It would be only natural to opt for one of the ontologies dedicated for describing bibliographic records, such as Functional Requirements for Bibliographic Records (FRBR) and its Resource Description and Access (RDA) and Bibliographic Framework Initiative (BIBFRAME) vocabulary variants (cf. Coyle 2016). Indeed, both contain a crucial distinction between "works" (a certain intellectual creation as such, regardless its edition, format or medium) and "instances" (expressions and manifestations of this intellectual creation) which in PBL is paramount for referencing editions, adaptations and critiques of a literary oeuvre of a particular author. For example, a review of *Don Quijote* refer to either Cervantes' literary achievement in general or to the newest translation of the Spanish original into Polish.

However, the FRBR-based ontologies are either not well equipped to handle theatre, cinematographic, radio and television instances of literary works, or (as in the case of FRBRoo) too complex to be easily handled by metadata producers in their everyday practice (Coyle 2016, 153).[3] Therefore, we opted for a solution that is more generic but robust enough - the schema.org ontology. However contestable due to its rather restricted vocabulary when it comes to describing books, this solution is not unprecedented in the bibliographic domain.[4]

This process of mapping is by no means mechanical. In many cases the PBL original methodology and the solutions of the first remediation entailed conceptual challenges, which will be addressed in more detail in our presentation. For instance, one needs to solve the tension between a minute bibliographic description on one hand, and the standard vocabulary on the other. Expressions entailing similar yet slightly different properties of the book such as "woodcut engravings"; "illustrations"; "drawings"; "reproductions"; "pictures"; "prints" need to be fit into the elements of the formal vocabulary of schema.org, properties such as "illustrator" and "artform."

(2) The second challenge of the new data-model involves the PBL subject classification structure. Here the option of using one of the existing and well-established subject headings/authority files published as Linked Data, such as the Library of Congress or German National Library Subject Headings was rather out of question, given the methodological uniqueness of PBL. Instead we will strive to create our own Linked-Data ready classification scheme while at the same time providing a partial mapping to existing resources.

To realize the scope of this challenge one needs to bear in mind that PBL has been an ongoing project for the last sixty years. During this time, not only literary life and its study have evolved (cf. the emergence of the online literary life, Maryl 2015), but also certain state entities disappeared (e.g. Yugoslavia or the Soviet Union). Given that the future database will be populated through retroconversion of the paper records in addition to the existing database records, we cannot take the current classification for granted, but also accommodate its historical evolution. A non-intrusive way to account for the historicity of PBL would be to add timestamps to subject headings. Whether a synthetic data-reconciliation layer is possible requires further analysis.

Conclusions

In the concluding remarks we will concentrate on the expected benefits of translating PBL into LOD.

- PBL datasets can be enriched through integrating other Linked Data collections (e.g. geographical data on places relevant to literary life).
- Data exchange protocols can be established between PBL and other bibliographies published as Linked Data.
- PBL data can be used for data-driven research in the humanities on such fields as reception history or transfer studies.
- The methodology and the production pipeline developed in this project can be reused for retroconversion of other disciplinary bibliographies.

Acknowledgment

This work was supported by Polish Ministry of Science and Higher Education through the National Programme for the Development of the Humanities (grant number: NR 0061/NPRH3/H11/82/2014).

Bibliography

Antelman, E., Lynema, A. and Pace, K. (2006). Toward a Twenty-First Century Library Catalog *Information Technology & Libraries*, **25**(3): 128–39.

Coyle, K. (2010). Understanding the Semantic Web: Bibliographic Data and Metadata, *Library Technology Reports*, **1**: 5-31.

Coyle, K. (2016). *FRBR Before and After*. Chicago: ALA Editions.

van Hooland, S. and Verborgh, R. (2014). *Linked Data for*