Maciej Maryl, Piotr Wciślik
Turning the Polish Literary Bibliography into a Research Tool: Challenges, Standards, Interoperability
Research based on bibliographical records

Research based on bibliographical records

Jockers 2013, 45.
Research based on bibliographical records

Figure 19. Australian novels by gender of author, percentages, 1945 to 2009 (five-yearly averages)

Figure 22. Men and women in the top twenty most critically discussed Australian authors, percentages, 1945 to 2006 (by decade)

Bode 2014, 132 i 134
Research based on bibliographical records

Long and So 2013a, 150.
Research based on bibliographical records

Long and So 2013b, 274.
Research based on bibliographical records

Moretti 2013, 183.

Open Data Citation Workshop, Prague 24.10.2016
Research based on bibliographical records

Moretti 2013, 201.

Figure 17: Ethical content of abstractions in short titles
Research based on bibliographical records

Figure 7: British hegemonic forms, 1760–1850

The anomaly constituted by the epistolary novel’s slump in the 1770s is only apparent, and easily explained: what declines in these years is the publication of all novels, and in fact, as figure 8 shows, epistolary novels were then even more hegemonic on the market—i.e. in 1790, when an impossible 71 per cent of new titles were novels in letters.

Moretti 2005, 15.
Research based on bibliographical records

Moretti 2005, 19.

For sources, see ‘A Note on the Taxonomy of the Forms’, page 53.
Research based on bibliographical records

- Online bibliographies with research components:
  - AustLit
  - NWBIB
  - Orlando
  - NEWW (Women Writers VRE)
Towards literary "Big Data"

- Big and dirty or small and clean?
- Better data $\rightarrow$ better answers
- Bigger data $\rightarrow$ bigger problems
- Working with data collected as a reference resource, not research material
- Goal: forging a database ready to answer unpredicted questions
Polish Literary Bibliography (PBL)

- Annotated bibliography of Polish literature (since 1954)
- Register of all articles, notes and other materials concerning literary life
- Bibliographical records concerning Polish Literature:
  - Authors
  - Works
  - Reviews
  - Articles
  - Contests and awards
  - TV, radio, theatre and cinema adaptations
  - Events
  - Exhibitions
  - Semantic annotations (e.g. ’literary theory’; ’exile literature’; ’psychology’; ’education’; ’censorship’)

Open Data Citation Workshop, Prague 24.10.2016

chc.ibl.waw.pl/PBL
• **Project:**
  - Polish Literary Bibliography as a research tool on contemporary Polish culture
  - 3-year project funded by the National Programme for the Development of Humanities (2015-2018)
  - ICT partner: PSNC (Poznań Supercomputing and Networking Center)

• **Goals:**
  - Creating a database of ~4 million bibliographical records on Polish literature (1939-2004)
  - Integration of heterogenous datasets
  - Integration with Linked Open Data Cloud
  - Data mining and visualisation tools for modeling processes of literary life
Remediating bibliography
How to get there?

- Choosing the right data standard
- Remediating the input tool (and habits)
- Converting the existing online database
- Retrodigitising printed volumes
- Integration with other bibliographies
Choosing the data standard

Schema.org INSTEAD of bib-dedicated ontologies and data models (FRBR, RDA, BIBFRAME)

BIB-DEDICATED ONTOLOGIES

- Not well equipped to handle theatre, cinematographic, radio and television instances of literary works (except FRBRoo)
- Too complex to handle by metadata producers (including FRBRoo)
- Validity still TBC by community of practice

SCHEMA.ORG

- The internet standard
- Robust enough to handle PBL data model
- Rendering Work/Expression/Manifestation somewhat difficult, but doable
- Not unprecedented in the library domain

DARIAH-EU

Open Data Citation Workshop, Prague 24.10.2016
Schema.org: standard vocabulary for various encodings
Schema.org: bibliographical extension

Schema.org is a set of extendible schemas that enables webmasters to embed structured data on their web pages for use by search engines and other applications. For more details, see the homepage.

This is the front page for the Bibliographic Extension, whose short name is bib.

You are viewing the Bibliographic Extension within schema.org. It defines terms such as Audiobook, Thesis, ComicStory, and workTranslation. For more details see the W3C BibExt Community Group's wiki.

This should be considered a pre-final preview release; final changes may be made after wider community review.

Terms defined or referenced in the 'bib' extension.

Types (15)
- Article, Audiobook, Chapter, Collection, ComicCoverArt, ComicIssue, ComicSeries, ComicStory, CoversArt, Newspaper, Thesis

Properties (18)
- abridged, artist, colorist, duration, inSupportOf, inker, letterer, pageEnd, pageStart, pagination, penciler, publishedBy, publisherImprint, readyBy, translationOfWork, translator, variantCover, workTranslation

Enumeration values (1)
- GraphicNovel

Terms and conditions

Data Liberate

DARIAH-EU
# Data conversion

### Documentation of mapping to schema.org

<table>
<thead>
<tr>
<th>Klasyfikacja</th>
<th>Rodzaj Zapisu</th>
<th>Klasyfikacja o temie</th>
<th>Klasyfikacja o twórcu</th>
<th>Klasyfikacja o utworze</th>
<th>Klasyfikacja w hasie rzeczowym</th>
<th>Klasyfikacja wywad</th>
<th>Klasyfikacja wyzwan</th>
<th>Klasyfikacja wyzwan wywad</th>
<th>Klasyfikacja wyzwan wyzwan wywad</th>
<th>Klasyfikacja wyzwan wyzwan wywad wywad</th>
<th>Klasyfikacja wyzwan wyzwan wywad wywad wywad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Klasyfikacja</td>
<td>Klasyfikacja o temie</td>
<td>Klasyfikacja o twórcu</td>
<td>Klasyfikacja o utworze</td>
<td>Klasyfikacja wywad</td>
<td>Klasyfikacja wyzwan wywad</td>
<td>Klasyfikacja wyzwan wyzwan wywad wywad</td>
<td>Klasyfikacja wyzwan wyzwan wywad wywad wywad</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klasyfikacja</td>
<td>Klasyfikacja o temie</td>
<td>Klasyfikacja o twórcu</td>
<td>Klasyfikacja o utworze</td>
<td>Klasyfikacja wywad</td>
<td>Klasyfikacja wyzwan wywad</td>
<td>Klasyfikacja wyzwan wyzwan wywad wywad</td>
<td>Klasyfikacja wyzwan wyzwan wywad wywad wywad</td>
<td>Klasyfikacja wyzwan wyzwan wywad wywad wywad wywad</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Open Data Citation Workshop, Prague 24.10.2016

[Link to CHC.IBL.WAW.PL/PBL](chc.ibl.waw.pl/PBL)
Data conversion challenges

- **Work/Expression/Manifestation**
  - Selective application to reference commentary (reviews, polemics etc.)
  - Not used to reference translations: the problem of translations from secondary source

- **Contributor roles**
  - 86 different roles generated through autopsy
  - Overlaps with the note field
  - List simplified and flexible: general/obligatory and specific/voluntary

- **Linking the data with the LOD cloud**
  - places
  - people
  - subjects
Remediating the tool
Remediating the tool (annotations)
• 1939-1945 (3 vols., 1250 pages)
• 1944-88 (51 vols., 37 661 pages)
• 1976-1989 - samizdat (1 vol., text file)
Retrodigitisation challenges

- Technology
  - Parsing
  - Lemmatization
- Tracing methodological inconsistencies
- Time factor
  - Subject-classification (new approaches to literary studies, e.g. gender studies)
  - new forms of literary life, e.g. online literary life
  - changing political geography: Yugoslavia and Soviet Union no longer there

Open Data Citation Workshop, Prague 24.10.2016
Research challenges: local specificity

- **Data-collection methodology**
  - Selection of writers
  - Unrecorded debuts
  - Data censorship
- **Geography**
  - Changing borders
  - Literature in Exile
  - Polish literature = in Poland or in Polish?
  - Literary transfers
- **Shape of literary life**
  - Official vs. Underground publishing
PBL_lab VRE

Topical data collections

- UGRND (1976-1989)
- PBL (1944-88)
- PBL online (1988-)
- WWII (1939-45)

LOD

DARIAH-EU

Open Data Citation Workshop, Prague 24.10.2016
Conscious research based on bibliographical data

In order to come up with right research questions (and answers)...

...be sure to know how your data were selected, collected, censored, abridged, standardised, annotated, digitised, corrected and linked together.
References


• Long, Hoyt and Richard So (2013a) ”Network Analysis and the Sociology of Modernism” boundary 2 40(2):147-182

• Long, Hoyt and Richard So (2013b) ”Network Science and Literary History” Leonardo 3(46), 274-274.


• Moretti, Franco 2013 "Style, Inc.: Reflections on 7,000 Titles" Distant Reading, London: Verso.