Report on the international conference "Mutual Learning Workshop for Improving Cultural Heritage Bibliographical Data" (October 12–14, 2022, Prague)

https://doi.org/10.19195/0080-3626.66.15

On October 12–14, 2022, the conference "Mutual Learning Workshop for Improving Cultural Heritage Bibliographical Data" took place, organised by the Bibliographical Data Working Group (BD WG) of the DARIAH-EU consortium. This group consists of representatives of cultural heritage institutions as well as researchers, who, during this meeting, discussed re-using, enriching and automated generating of bibliographical data in an interdisciplinary perspective. The event was held in in the Upper Conference Hall at Ústav pro českou literaturu Akademie věd České republiky (Institute of Czech Literature of the Czech Academy of Sciences — ICL CAS) in Prague (Na Florenci 1420/3).

Thanks to the funding from DARIAH-EU, the stay and participation of many speakers from abroad was possible, which made the conference truly international and interdisciplinary. There were 24 participants from 11 countries (Belgium, the Czech Republic, Finland, France, Germany, Greece, Hungary, the Netherland, Poland, Spain, Switzerland) who gave 20 speeches in total. The conference was held in an in-person format but was also accessible online via Zoom (with some participants delivering remote speeches).

* * *

The conference was preceded by an official meeting of the BD WG's members (October 12th, 10:00–12:00 CEST), who discussed the group's main goal for 2023 and 2024 and focused on outlining the general idea described as "Understanding society and culture through bibliographical data. Scaling up bibliographical data research in the social sciences and humanities". The goal will be specified in 2023 (possibly during an international conference), at which time opportunities and challenges for bibliographical data science, open bibliodata, best practices for documentation of bibliodata resources, and bibliodata LOD-ification using open-access software will be discussed. A project team for book-sprint about the bibliodata related workflows in the structure of SSHOC Workflows is also scheduled to be established in 2023. The results will be presented at the DARIAH Annual Conference in 2024. The last item on the agenda is the publishing of a peer-reviewed book containing the BD WG's report *An Analysis of the Current Bibliographical Data Landscape in the Humanities: A Case for the Joint Bibliodata Agendas of Public Stakeholders* (Bibliophical Data Working Group, 2022).

The conference was divided into several panels. The first panel — *Introductory speeches* (13:30–14:35 CEST) — presented an outline of what the ICL CAS, DARIAH-EU and the BD WG have in common. The *Welcoming word* was given by Michal Kosák (Statutory Deputy Director of the ICL CAS, who spoke on behalf of absent Petr Šámal, Director of the ICL CAS), who said that when he started his job in the ICL CAS over 20 years ago, it was unimaginable to him that an international bibliographic conference could take place in Prague. Fortunately, this situation has changed because of the Czech Literary Bibliography and the BD WG.

Next, Sally Chambers (Director of DARIAH-EU and Project Coordinator at Koninklijke Bibliotheek van België — Royal Library of Belgium) gave her online speech *Cultural heritage (bibliographical) data as humanities research data*, in which she presented her connections with bibliographic data in her professional life. She described DARIAH as a pan-European digital research infrastructure (encompassing knowledge, tools, data and networks) for arts and humanities research, its members and partners, its strategic goals and foundations — all against the background of diversity and richness of metadata standards (see Riley, Becker, 2009–2010), strongly underlining the core ideas of the GLAM (see Candela, Sáez, Escobar Esteban, Marco-Such, 2022) and describing collections as data (according to the FAIR rule — Findable, Accesibile, Interoperable, Reusable; see Tasovac, Chambers, Tóth-Czifra, 2020). She also announced that the DARIAH-EU Annual Report 2021 had been published (presenting existing activities and results as well as new challenges for a post-pandemic world). Chambers also described specific goals from the Strategic Action Plan (STRAPL) for 2023 and 2024 (exploring, defining and advancing DARIAH-EU's place in the information ecosystem, developing and defining the concept of the workflow, ensuring sustainability of the SSH Open Marketplace).

This panel concluded with the speech *Investigating relationships of bibliographical metadata and textual analysis to foster interdisciplinary collaboration* delivered by Tomasz Umerle and Vojtěch Malínek (Co-chairs of the BD WG), who presented intersections between textual data enrichment (digitised text collections) and automated metadata generation, or — in other words — intersections of NLP and LOD processing. This makes automated metadata generation a joint challenge for all members of the BD WG. The discussion after the panel revealed that publishers were very important stakeholders of bibliodata (by creating, using, crowdsourcing them), however, no publisher was present in the BD WG. This issue needs to be rectified in the future.

The introductory section was followed by the only thematic panel of that day, *Bridging the gap between bibliodata curation and research: FRBR-based models and works authority file* (14:50–17:00 CEST), which started with the speech *No, the library catalogue does not follow FRBR and here are some reasons* prepared by José Calvo Tello (Niedersächsische Staats- und Universitätsbibliothek Göttingen — Lower Saxony State and University Library in Göttingen) and Nanette Rißler-Pipka (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen — Society for Scientific Data Processing in Göttingen; GWDG) but presented only by the first author. The case of *Pride and Prejustice* by Jane Austin and certain statistics generated from the K10plus catalogue (70 mln records) revealed the existence of two more levels (Title, Local property) than in the FRBR structure (Author, Work, Expression, Manifestation) and showed which data were missing in this library catalogue (using Gemeinsame Normdatei — GND) compared to Wikidata and the Virtual International Authority File (VIAF). The authors also presented Linking Literary Works in Bibliodata (LiWoB) as an enriched data model for library catalogues.

The second speech titled *Changes in cataloguing of literary works related to the introduction of National Library of Poland Descriptors* was delivered by Kamil Pawlicki (Biblioteka Narodowa — National Library of Poland). The discussed changes were caused by the Library dropping pre-coordinated National Library of Poland Subject Headings (in use until 2016) in favour of descriptors (collected in the dictionary: dbn.bn.org.pl), which are post-coordinated. It allowed to organise the catalogue of this library with facets and created new access points to information: Cultural area (386 MARC field), Audience (385 MARC field), Form and type (380 MARC field: fixed list of 38 descriptors), Genre (655 MARC field). Additionally, it enabled marking certain new aspects of subject descriptions, such as subject descriptions of literary works (e.g. real characters, fictional characters, dynasties, historic events, intertextual references, places, social environment, institutions mentioned in novels, poems, etc.).

Ondřej Vimr (ICL CAS) then presented his speech titled *FRBR, library catalogues, bibliographies, or what do we need to carry out research into literature*, in which he talked about the kinds of information that needs to be put into bibliographies for scholars and researchers working on global literary studies (comparing national literatures and translations), who's interest lie in spatial and temporal connectivity, integration scale, agency, mediation and remediation in literature in macro- and microperspectives. He conducted a case study of the Ph.D. thesis by Stefanie Kremmel (*The Transculturality of "The Communist Manifesto"*; University of Vienna) to demonstrate compiling and using a bibliography in global literary research at a micro scale — it is a "biography" of translation, which shows how and in what form *The Communist Manifesto* (translated into over 50 languages) was republished, recycled, cited, used covertly and openly. He then proposed some requirements for translating bibliographies that should be fulfilled to make compiling this kind of biographies easier or even make them possible at all.

At the end of the day, Péter Király (GWDG) presented the speech *FRBR from the perspective of book history: Problems, and suggestions* (prepared in collaboration with János Káldos, an independent researcher). On the basis of Darnton's (1982) and Adams' and Baker's (1993) models of book history as well as works preceding *The Winter's Tale* by William Shakespeare ("traces of read") and ontologies of Bibliotheca Eruditionis 1500–1700 (Hungarian), The French Book Trade in Enlightenment Europe (Australian) and MEDIATE (Dutch) databases, he listed technological challenges in book history studies related to using a FRBR model in library catalogues, and proposed some grassroots solutions to be apllied in this area.

The next day of the conference (October 13th) started with the panel *Bibliographical data and text collections — automated metadata generation* (9:40–11:30 CEST). The first speech was delivered online by Osma Suominen (Kansalliskirjasto — National Library of Finland), an information systems specialist, who talked about *Automated generation of subject headings using Annif and Finto AI*, his experiences with developing and evaluating these tools, their implementations in existing catalogues and databases throughout Europe, and who also outlined his future plans. Finto is a thesaurus and ontology service (composed of an automated subject indexing tool and an API service — Finto AI) and Annif was the best solution for automated indexing, because existing tools do not understand the Finnish and Swedish languages, do not cooperate with YSO (General Finnish Ontology), and, frankly, their upkeep costs are too high. Its prototype was launched in 2017, and since 2018 it has been developed (undergoing two evaulations in 2019 and 2021) into a community-oriented open source software, which is multilingual, autonomous in indexing vocabulary, and it supports different indexing algorithms. Annif could also be applied to personal private data, can be tested on private texts (on the service's website: annif.org). It this respect, the wiki documentation on GitHub as well as tutorials uploaded on YouTube are very helpful. Suominen was among several researches who wrote many academic articles about the development and evaluation of Annif.

The second speech, titled *Automated generation of keywords from the NLP perspective*, was given by Agnieszka Mikołajczyk-Bareła (Voicelab.ai; CLARIN) and it focused on using keywords extracted from abstracts or other short texts as subject headings or descriptors within the keywords generation model called plT5kw[1].

Next, Ioanna Grypari (Αθηνά — Athena Research Center) gave the speech *Implementation of fields of science and sustainable development goals automated classification systems in OpenAIRE*, in which she presented the OpenAIRE Research Graph — a collection of metadata (describing different kinds of objects in the research lifecycle and a relationship between them) harvested from many sources and users. It 1uses the FoS (Field-of-Science) Classification System (as a three-level taxonomy of sciences, where levels 1 and 2 come from the OECD classification extended with — as an additional, third level — ScienceMetrix classification labels, manually linked into first two levels) and the SDG (Sustainable Development Goals) Classification System (based on key phrases brought from SDG and matched to articles' titles and abstracts). Both systems have their classifiers, enriching metadata collected in OpenAIRE.

At the end of this panel, Mikko Tolonen (Helsingin yliopisto — University of Helsinki) and David Rosson (Technische Universität Berlin — Technical University of Berlin) said a few words about text similarity as a potential source of metadata in their paper titled *Show & tell of prototypes of the Reception Reader and Semantic Search interfaces developed in High-Performance Computing for Historical Discourse Detection project*[2] and presented its possibilities using *Macbeth* by William Shakespeare.

The next panel on that day, *Bibliographical data and text collections — metadata creation and enrichment in digitisation projects and webarchives* (13:10–15:10 CEST), started with the paper *Semantic enrichment of historical newspapers in NewsEye: A digital investigator for historical newspaper* read by Antoine Doucet (La Rochelle Université — La Rochelle University) about challenges, goals, organisation and results of a European project (which lasted from May 2018 to January 2022)[3] connecting libraries and universities from 4 countries (Austria, Finland, France, and Germany). Historical newspapers from the 1850–1950 period are a "goldmine" for SSH scholars — millions of pages digitised around Europe in multiple sources and languages and, thanks to semantic enrichment of metadata and full-texts, NewsEye provides access to this treasure-trove, whereas a vast majority of digital libraries (collecting many types of documents) only enabled accessing the information on a basic, undetailed level, which makes it difficult to retrieve the information given in newspapers.

[1] An online demo is available at: nlp-dem-1.voicelab.ai [accessed: 10.11.2022]. See also: Pęzik, Mikołajczyk-Bareła, Wawrzyński, Nitoń, Ogrodniczuk (2022).

[2] See: app-kaiku.rahtiapp.fi/semantic [accessed: 10.11.2022].

[3] See: newseye.eu [accessed: 10.11.2022].

In relation to this speech, Matteo Romanello (Université de Lausanne — University of Lausanne) presented the work of 15 people: *Impresso — media monitoring of the past*, a fully open access tool[4] intended to help correcting OCR mistakes, finding the frequency of any word in an article of certain type or on a certain subject, finding articles mentioning a given person, similar or associated items and building one's own subcollection of retrieved articles. This brilliant work presents a lot of challenges (with stored files, messy data, noisy OCRed texts, visualisation and exploration), described by Romanello (who also outlined solutions implemented to combat them). At the end of the speech, unique possibilities of Impresso were demonstrated in a short movie presenting a case study of the reception of the Battle of Arhnem in Switzerland and Luxembourg newspapers and comparing it with the results obtained from a similar search for the Battle of Stalingrad.

The panel closed with the paper *Bibliographical metadata and web archives: Some solutions for metadata management in the field of web archiving* presented by Márton Németh (Országos Széchényi Könyvtár — National Széchényi Library), who answered the following questions: what is the subject of the description in web archives? what kind of metadata is needed? in what format? how can this data be produced? and what can this data be used for? WaybackMachine, the most popular search engine in this area, omits much of internet content, especially from national internet domains. National libraries are trying to fix national (e.g. British, Czech, Hungarian) web archives, but it doesn't solve the lack of full-text search capability, lack of relevance, etc. The fact that the living web can be defined as one single document — huge, ever-changing, and unlimited — only makes the answer to the following question more difficult: which domain boundaries, collection methods, levels of descriptions, types of metadata, users' needs should be chosen or taken into account? Some standards and recommendations could help with this (ISO 14873:2013, ISO 28500:2017, Dublin Core familiar *Descriptive metadata for web archiving*, RDA familiar *Metadata application profile for description of websites with archived version*). Németh then demonstrated how metadata records used in the National Széchényi Library[5] try to remove these obstacles by retrieving bibliographic information about websites as well as the prospects for the future.

The last panel of that day was *Open science in bibliodata re-use: Linked open data, data modeling and science evaluation* (15:30–17:00 CEST). It started with David Lindemann's (Euskal Herriko Unibertsitatea — University of the Basque Country) speech titled *LOD-ifying bibliographical data using free software: CLB-LOD Wikibase* about an experiment with the Czech Literary Bibliography MARC dataset exported and presented with Wikibase with some preliminary considerations. Firstly: applying the FAIR Open Data and 5-Star-Linked-Data guidelines should make the workflows accessible to anyone and enable linking via Wikidata. This idea is grounded in the fact that while MARC is not understood by non-bibliographers, the wiki-environment is. What is more, Wikibase is better for data analytics and its data could be federated with Wikipedia. It was the third experiment of this kind (Lindemann was previously involved in exporting Zotero and SQL into Wikibase), and, just like the previous ones, lead to a succesful outcome.

[4] See: impresso-project.ch [accessed: 10.11.2022].

[5] See: webarchivum.oszk.hu [accessed: 10.11.2022].

The next speech was *LexMeta: A linked open metadata model for bibliographical data of dictionaries* given by Penny Labropoulou (Αθηνά) and David Lindemann, who presented a process of extending the content of the LexBib Konwledge Graph and the Catalogue of Lexicography and Dictionary Research (including bibliographical items, events, persons, dictionaries), using the Wikibase software while implementing the RDF/OWL and SKOS objectives. Dictionaries are difficult to describe because they can exist either in a document form (books, CD-ROMs, pdf files — considered as a publication for citing) or function as lexical resources/datasets (computational lexica, ontologies, terminological lists — considered as easily accessible content). The final effect was enhanced with classifying entities using the LexVoc classification, distribution properties (access, format, size, publisher, licence etc.), and their relations with other entities.

This day closed with a remote entry by Giovanni Colavizza (Universiteit van Amsterdam — University of Amsterdam) and his paper *Wikipedia citations: Opening up Wikipedia as an altmetric resource*. The English-language Wikipedia is a very important source of information because it is a bridge between science and ordinary users. Articles published there have a lot of content and sources cited, thus the cited data has to be extracted, classified and validated. Based on his own experience and research, Colavizza provided us some notes on what needs to be done in the scope of entities cited in Wikipedia for turning it into better resource.

The last day of the panel (October 14th) was comprised of a single panel — *Opening metadata collections resource for the needs of research and scientific communication* (9:45–11:30 CEST), during which three speeches were delivered. The first one, titled *Quantitative analysis of the Swedish National Bibliography data: Case study*, was given by Ylva Sommerland (Kungliga biblioteket — National Library of Sweden), a cataloguer. Sommerland demonstrated how Nationalbibliografin (Swedish National Bibliography) changed its statistical information in 2021: from a formal report (*Nationalbibliografin i siffror*) presenting national bibliography in numbers to a common poster folded into a brochure (*Utgivningspuls*) explaining current trends in the Swedish book publishing trade.

The next speech, *The National Library of France: A national library's perspective on open bibliographic metadata*, was given by Mathilde Koskas (Bibliothèque nationale de France — National Library of France), who is the Chair of the IFLA Bibliography Section. She presented library metadata (catalogue data which are bibliographic and authority records) as open data (with different Creative Commons licences: CC0 — no copyright, CC-By or ODC-By — open with attribution). It is very important from the point of view of public sector information. Koskas then showed how the National Library of France provides data (as openly as possible)[6], and presented conclusions which were met with approval by the.

The last speech, *Reusing an open metadata collection to model and enrich a corpus for NLP-based research: The case of 19/20MetaPNC*, was delivered by Agnieszka Karlińska, Cezary Rosiński and Tomasz Umerle (Instytut Badań Literackich Polskiej Akademii Nauk — Institute of Literary Research, Polish Academy of Sciences). The Polish literary corpora review — a Polish subcollection in the European Literary Text Collection, Polish

[6] See: data.bnf.fr [accessed: 10.11.2022]; bibliographienationale.bnf.fr [accessed: 10.11.2022]; bnf.fr/fr/bnf-datalab [accessed: 10.11.2022].

19th- and 20th-century Novels (see Kubis, 2021) — was presented as a resource that needed to be both properly balanced and representative (by a properly defined collection and determined proportions across text categories) and described a new metadata-enriched corpus (because there is no representative and balanced collection of Polish novels that could serve as a referential corpus), called Metadata-enriched Polish Novel Corpus from the 19th and 20th Centuries (19/20MetaPNC), which included the content of 1000 Polish novels with plots from after 1815, published between 1864 and 1939[7]. This raised a few research questions, which were examined and answered by the team.

The conference was then closed with the *Summary and joint statement*, in which Mikko Tolonen, Vojtěch Malínek and Tomasz Umerle talked about how we benefited from the three conference days. Tolonen started it with the words: "The country is good when children are laughing and playing — says a Finnish proverb". Although Tolonen made up this proverb, it is a good representation of the moment in history we find ourselves in. We know more and more about data and we are able to collect and process them better and better. Now it is time to use this asset more and start to "laugh and play" with it: exchange, reuse, visualise, enhance, enrich, and discover the knowledge not available to retrieve from analogue processed data. We have metadata and we have texts and, what is more, we are beginning to put them together. During the conference we saw great presentations delivered by researchers in libraries and digital humanities. What connected all speakers and their audience, no matter where they came from and what they did, was the intellectual attention paid to the automated creation of metadata. Admittedly, we could have talked a little bit more about the reuse and enrichment of data. Gaps in the data landscape still exist, and the first thing to do is to bridge them and then start to think about the next steps. Our challenge for the future is to do more as a community. Now, when a "country" of datasets becomes "good," we need to go further. We need to think what we could do better together. Because we could achieve more with automation, but automation does not bring us together and does not provide the necessary quality, since AI cannot replace cataloguing and bibliographing by people — it can only support it. Libraries have their traditional responsibilities, but the present tasks them with a new ones — and judging which ones to prioritise can be very difficult. Therefore we need to invest in librarians because things are changing, new technologies are coming, and we need to train people to use this great treasure-trove: once technology make them free of some duties, they will be able to take care of other tasks.

* * *

As a result of the meeting the BD WG aims to prepare a public statement on how producers and consumers of bibliodata can intensify a collaboration in the fields discussed during the workshop. Both kinds of materials from the conference — slideshows and recordings — are available online[8]. All of them were worthwile — especially since BD

[7] See: tinyurl.com/metapnc [accessed: 10.11.2022].

[8] Search for "Mutual Learning Workshop for Improving Cultural Heritage Bibliographical" on Zenodo: zenodo.org/search?page=1&size=20&q=%22Mutual%20Learning%20Workshop%20

WG meetings are the only ones of their kind in the scope of bibliography, data, GLAM institutions and users in this part of Europe. I cannot wait for the next ones. If you cannot wait too, you should follow the BD WG's news channel on Twitter[9]. As was stated at the beginning, the closest opportunity could be just ahead, during the Annual Digital Humanities Conference (ADHO), July 10–14, 2023, in Graz, Austria. Stay tuned!

Jakub Maciej Łubocki
ORCID: 0000-0002-1957-0682
Muzeum Narodowe we Wrocławiu

REFERENCES

Bibliophical Data Working Group. (2022). *An Analysis of the Current Bibliographical Data Landscape in the Humanities: A Case for the Joint Bibliodata Agendas of Public Stakeholders*, Zenodo. https://zenodo.org/record/6559857#.ZEjCLXZByUk,%20DOI:%2010.5281%2Fzenodo.6559857.

Candela, G., Sáez, M.D., Escobar Esteban, M.P., Marco-Such, M. (2022). Reusing digital collections from GLAM institutions. *Journal of Information Science*, *48*(2), 251–267. DOI: 10.1177/0165551520950246.

Kubis, M. (2021). Quantitative analysis of character networks in Polish 19th- and 20th-century novels. *Digital Scholarship in the Humanities*, *36*(suppl. 2), ii175–ii181.

Pęzik, P., Mikołajczyk-Bareła, A., Wawrzyński, A., Nitoń, B., Ogrodniczuk, M. (2022). Keyword extraction from short texts with a Text-To-Text Transfer Transformer. In: E. Szczerbicki, K. Wojtkiewicz, S.V. Nguyen, M. Pietranik, M. Krótkiewicz (Eds.), *Recent Challenges in Intelligent Information and Database Systems: 14th Asian Conference, ACIIDS 2022, Ho Chi Minh City, Vietnam, November 28–30* (pp. 530–542). Singapore: Spinger. DOI: 10.1007/978-981-19-8234-7_41.

Riley, J., Becker, D. (2009–2010). *Seeing Standards: A Visualization of the Metadata Universe*, Jennriley.com. https://jennriley.com/metadatamap/.

Tasovac, T., Chambers, S., Tóth-Czifra, E. (2020). *Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper*, HAL Science Ouverte. hal.archives-ouvertes.fr/hal-02961317/document.

for%20Improving%20Cultural%20Heritage%20Bibliographical%22 [accessed: 2.12.2022] and on YouTube: youtube.com/playlist?list=PLqQrH2L40gN2vxi_VLKty3IAyNh7u6xkG [accessed: 2.12.2022].

[9] See: twitter.com/bibliodatawg [accessed: 10.11.2022].