Journal of Open Humanities Data

Open Bibliographical Data Workflows and the Multilinguality Challenge

RESEARCH PAPER

VOJTĚCH MALÍNEK
TOMASZ UMERLE
EDWARD GRAY
IVAN HEIBI
PÉTER KIRÁLY
CHRISTIANE KLAES
PRZEMYSŁAW KORYTKOWSKI
DAVID LINDEMANN
ARIANNA MORETTI
CHARLOTTE PANUŠKOVÁ
RÓBERT PÉTER
MIKKO TOLONEN
ALDONA TOMCZYŃSKA
ONDŘEJ VIMR

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The aim of the paper is to present and analyze workflows for bibliographical data curation and research that were created during the 'Open Bibliodata Workflows' project realised by the Bibliographical Data Working Group from the DARIAH ERIC consortium. These workflows are available via SSH Open Marketplace. Its role in the SSH infrastructural system is subsequently shortly introduced. Bibliodata-related workflows are needed at different levels of data creation and research, both for specific software features or data sources as well as for consolidating methodological aspects of bibliographical data curation. Set of five workflows showcasing various models of bibliodata related workflows is discussed afterwards. First of these workflows, *From Library Data to Research Data* describes conversion of library data into a dataset for data-based research. The other four are centred around leveraging existing tools and services. *AVOBMAT: how to analyze and visualize bibliographical data and texts* showcases a tool for combining text analysis and metadata-based research. *Metadata crosswalk for citation data production in OpenCitations* is a step-by-step instruction for using the OpenCitations infrastructure, a state-of-the-art service for sharing open citation data. *LODification of bibliographical data: Zotero to Wikibase migration* illustrates current dynamic developments concerning metadata in the field of Linked Open Data. Finally, the National Information Processing Institute from Poland (OPI PIB) prepared a workflow *Studies on science and higher education system in Poland using the RAD-on platform*, discussing how to use their dataset for research.

Analysis of these workflows reveals particular needs to address the multilinguality challenge in the bibliodata field. On the level of curation this challenge is met with application of international standards for bibliographical data processing that on many

CORRESPONDING AUTHOR:
Vojtěch Malínek
Institute of Czech Literature, Czech Academy of Sciences, Prague, Czech Republic
malinek@ucl.cas.cz

KEYWORDS:
bibliographic data; workflows; multilingualism; data analysis; data driven research

TO CITE THIS ARTICLE:
Malínek, V., Umerle, T., Gray, E., Heibi, I., Király, P., Klaes, C., Korytkowski, P., Lindemann, D., Moretti, A., Panušková, Ch., Péter, R., Tolonen, M., Tomczyńska, A., & Vimr, O. (2024). Open Bibliographical Data Workflows and the Multilinguality Challenge. *Journal of Open Humanities Data*, 10: 27, pp. 1–14. DOI: https://doi.org/10.5334/johd.190

occasions do not prioritise harmonization of multilingual datasets. The main curatorial techniques on how to solve multilingual issues in bibliographical data are briefly outlined. When we are tackling research questions the multilinguality challenge is even more prominent. Hence we are closing this article with a proposal for a preliminary workflow for processing multilingual bibliodata.

(1) INTRODUCTION

This paper is an output of the Open Bibliographical Data Workflows project funded through the DARIAH Theme 2022 programme from the DARIAH ERIC. DARIAH's Bibliographical Data Working Group took this opportunity to prepare and disseminate a set of workflows in key areas related to bibliographical data (or 'bibliodata') curation and research.[1] Our paper begins by discussing the challenges and opportunities of publishing workflows via the SSH Open Marketplace for which a set of five case study workflows has been prepared. The set of workflows presented here deals with both bibliodata curation and research and showcases various aspects of bibliodata related agendas and their practical implementation, ranging from converting library data to research data (see 2.1) and introducing a novel data analysis toolkit (2.2), to sharing of open citations (2.3), bibliodata LODification (2.4), and providing a national service for bibliometrics (2.5). The paper then proceeds to the 'multilinguality challenge' that the bibliodata community is facing and how this challenge could be addressed by workflow creation and dissemination. Multilinguality in the context of bibliodata is understood here as creation of a bibliographic dataset which contains data in multiple languages in a way that facilitates seamless data processing and analysis while respecting linguistic diversity. With this in mind, we propose differentiating between the needs of bibliodata curation and research, arguing that more attention is needed to raise awareness and share best practices for processing multilingual datasets in bibliodata research workflows, given that some solutions for multilingual challenges have already been adopted by the curatorial community. A preliminary version of the workflow for 'multilingual harmonization' of bibliodata is proposed that could be further discussed and published on the Social Sciences and Humanities Open Marketplace (SSH Open Marketplace) to supplement existing bibliodata-related workflows.

(1.1) WHAT ARE SSH OPEN MARKETPLACE AND SSH OPEN MARKETPLACE WORKFLOWS?

Workflows are important from a methodological point of view as they enumerate and describe the variety of steps to be followed while working on specific bibliodata-related tasks. They could play a key role in documenting specific use cases, serving to fix concrete instructions on how the same dataset might be adapted to meet different research needs, or how a service can be used. Making these workflows publicly available undoubtedly meets the needs of the larger DH community, supports implementation of open science policies, and facilitates interinstitutional and interdisciplinary cooperation in the field (Oberbichler et al. 2021). Furthermore, the publication of workflows contributes to FAIRification of scholarly communication by providing a structured and formalized, findable and openly accessible resource for the academic community. This goal is best achieved by publishing workflows via the SSH Open Marketplace, which also serves to embed them within a wider and denser network of DH resources.

The SSH Open Marketplace is a discovery portal which pools and contextualizes resources for SSH research communities. These resources, broken into five categories (tools and services, training materials, datasets, publications, and workflows), allow the SSH Open Marketplace to showcase solutions and research practices for every step of the research data life cycle. Particularly important are the built-in links contextualizing resources with one another, allowing users to see, for instance, not just the tool itself but related training materials for

1 Within the implementation of the aforementioned project, an intensive two-day-lasting booksprint was organised for which the specialists with various expertise with bibliographical data curation and research have been invited. The project team includes 13 data curators, data researchers and data analysts which are responsible for operation of large bibliographical data related infrastructures or being experienced in bibliodata-driven research or with knowledge on how to arrange bibliodata specific software tools. During the project, six bibliodata related workflows were prepared and subsequently published at the SSH Open Marketplace.

learning how to use it, and publications that illustrate how other scholars have used the tool. Another aspect of contextualization is the reliance on domain-specific vocabularies, such as TaDiRAH, that allow users to describe their resources with rich metadata (Borek et al. 2021). This contextualization, alongside strong curation and community outreach, allows the SSH Open Marketplace to facilitate discoverability and findability of research services and products that are essential for the sharing and re-use of research workflows and methodologies.

The SSH Open Marketplace, developed during the Horizon 2020 Project SSHOC, acts as a thematic entry door into the European Open Science Cloud (EOSC) and is maintained by DARIAH, CLARIN, and CESSDA. The development work behind the SSH Open Marketplace was inspired by previous ventures such as the DiRT directory (Dombrowski 2014), TAPoR, and Standardization Survival Kit (Riondet & Romary 2018), and indeed the initial data aggregation came in part from these sources (Gray et al. 2021). Aside from this initial data aggregation, the SSH Open Marketplace also developed a sustainability plan and governance scheme that integrated the lessons learned from these previous experiences – notably the 'directory paradox' (Dombrowski 2021) – in order to integrate community feedback and provide solid infrastructure and human resource support from the ERICs in the interest of maintaining platform viability (Petitfils et al. 2021). As part of this effort, an editorial board was established to actively maintain the platform and curate its content.[2]

A key asset of the SSH Open Marketplace is Workflows, which build upon the built-in contextualization of the Marketplace to show a step-by-step approach to a research scenario.[3] Workflow is an ideal way to share one's research resources, and harness the power of the SSH Open Marketplace to contextualize tools and services with publications, datasets, and training resources, thus presenting a research activity from A to Z in a way that is reproducible and easy to follow.

(2) BIBLIOGRAPHICAL DATA WORKFLOWS

The Open Bibliographical Data Workflows project aimed to prepare and publish a set of bibliodata-related workflows, because of the need to share best practices, popularize existing services and datasets, and facilitate the future reproducibility of research. These workflows represent an array of examples, from more general (Workflow 1) to more specific 'use cases' (Workflows 2–5) with regard to tools and services. At the same time each workflow discusses different types of bibliodata to ensure that the needs of the largest possible user group within the bibliodata community are addressed.

(2.1) WORKFLOW 1: FROM LIBRARY DATA TO RESEARCH DATA (TOLONEN ET AL. 2023)

Library data is without question the most frequent type of bibliographical data, and is naturally a key data source for data-driven research of this type of data.

Library catalogues have been identified as a crucial resource for studying different aspects of print culture spanning from literature to intellectual history and to informatics (Abramitzky & Sin 2014; Lahti et al. 2019; Roig Sanz & Fólica 2021; Vimr & Rosiński 2022). At the same time, using them requires addressing challenges of data quality, representativeness, coherence, completeness, and interpretation. An important aspect of bibliographic data science workflow is that it is imagined as a multilingual and transnational way of approaching extensive data for SSH research over long periods of time. Data from national libraries, for example, cover hundreds of years of data and many languages. Looking at the possibility of combining different library datasets from multiple countries, researchers often face the challenge of dealing with language dependencies, deduplication, and structural harmonization.

Available bibliographic metadata is seldom readily amenable to quantitative analysis. Biases, inaccuracies, and gaps hinder the productive use of bibliographic metadata collections for research (Coleman 2020). Varying standards, conventions, and languages pose challenges for

2 See the list of Editorial Board Members here: https://marketplace.sshopencloud.eu/about/team (last accessed: 6 December 2023).

3 See the list of workflows already available in the SSH Open Marketplace: https://marketplace.sshopencloud.eu/search?categories=workflow&order=label (last accessed: 6 December 2023).

data integration. The purpose of harmonization is to turn catalogue information into a dataset that can be used in quantitative humanities research. The core of this type of work consists in iterating between data harmonization, conducting analysis built on different use cases, and validating data against other sources. The central position of these three activities within the whole iterative process is schematically presented in Figure 1.

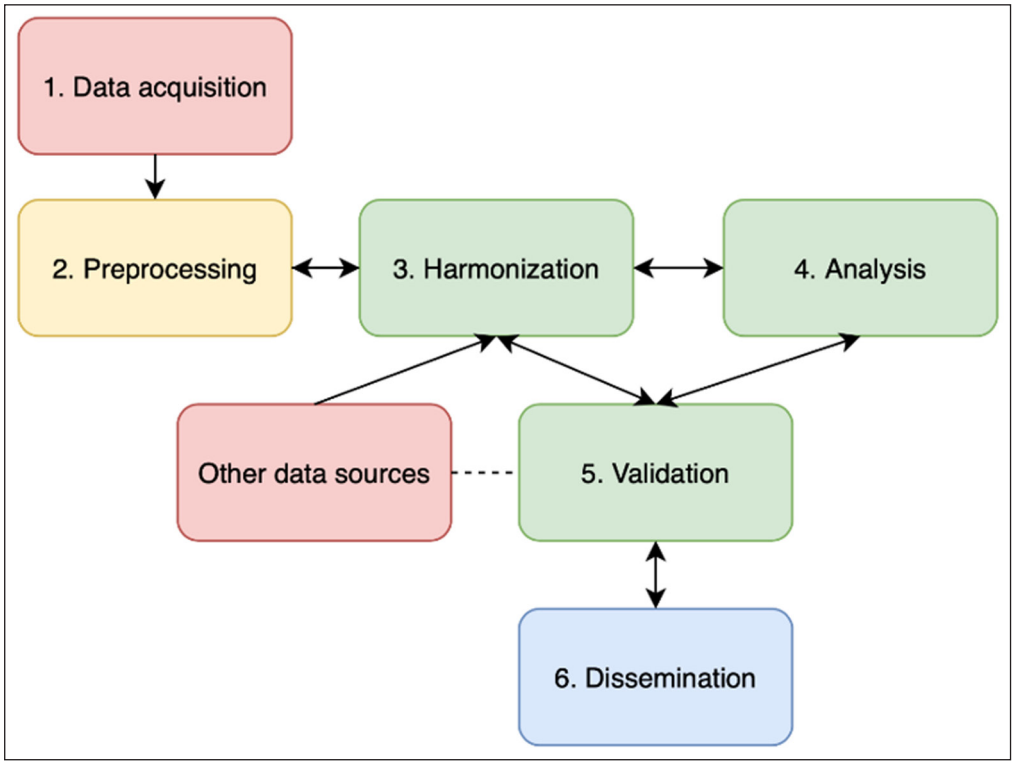

Figure 1 Bibliographic Data Science: From Catalogue to Research Data Workflow.

For harmonization, it is common to use external data sources on authors, publishers, and places to enrich and verify bibliographic information. Examples of harmonization include the removal of spelling errors, inconsistent transliteration or transcription, disambiguated and standardized terms, augmented missing values, and developed custom algorithms that can convert the raw MARC notation to numerical page count estimates, for instance. For library catalogue data we can use largely identical algorithms across most metadata collections. However, we often have to deal with various challenges posed by different languages, especially if we want to work across catalogues from different countries. Automation, scalability, and quality control are critical, as the data collections may contain information on millions of documents.

It is important to understand that harmonization is an iterative process that combines harmonization, analysis, and validation of data. On a deeper level, data quality is assessed by way of such generic quality dimensions as completeness of bibliographical information, conformity, appropriateness, and consistency (Cichy & Rass 2019; Király 2019). The slogan 'fitness for use', frequently mentioned in quality assessment literature, reflects the manner in which we always measure how a functional requirement could be satisfied by the data. Since these requirements may differ between a particular library and researcher, what may count as high quality data in one context may be low quality in another.

We might imagine this as an open science ecosystem made up of different metadata collections, where work on one of them eases the use of another. It is also helpful to think in terms of ecosystems because the harmonization step often depends on other linked sources such as authority files or other bibliographical sources. Some have described this approach, through which library metadata catalogues become research data, as 'bibliographic data science' (Lahti et al. 2019). In the case of a workflow that describes the conversion of raw data from library catalogues into research data, it is important to account for, document, and control the process. This can be described as an open science initiative because it takes questions of reproducibility and data quality seriously.

(2.2) WORKFLOW 2: AVOBMAT SOFTWARE: HOW TO ANALYZE AND VISUALIZE BIBLIOGRAPHICAL DATA (PÉTER 2023)

The aim of creating the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) multilingual research toolkit was to combine bibliographical data and textual analyses in several languages by using current NLP techniques and methods in a transparent and reproducible workflow, focusing especially on multi-language comparative analysis with time-based components (Péter et al. 2020; Péter et al. 2022; Péter et al. 2024). This methodological approach makes it possible to ask more complex research questions than would otherwise be possible when dealing separately with either textual or bibliographic data analyses. The majority of NLP-based text analysis applications make little use of bibliographic data. As for the multilingual aspects of the workflow, the outcome of preprocessing and analytical modules depends on the language(s) of the uploaded texts. There are two ways to assign a language to a document: researchers can manually choose a language for the full dataset (out of 52 available languages) or select the automatic language detection option. As regards the latter, the system chooses a language independently for each document; in the preprocessing phase, based on the selected language, users can choose a number of other features, including stopword, punctuation filtering, and lemmatization. For stopword and punctuation filtering, the spaCy library is used, to which additional stopword and punctuation lists can be added. The research toolkit implements spaCy language models (small, large, and transformer) for lemmatization; in the case of languages without spaCy models (e. g. Czech, Bulgarian, Romanian), LemmaGen (Juršic et al. 2010) is implemented.

Users can investigate the bibliographic data of texts preprocessed in AVOBMAT in various ways (e.g. diachronic charts, network analysis, gender analysis of authors) by using distant and close reading methods. They can also explore semantically enriched metadata, drawing, for example, on named entity recognition of the documents. Metadata enrichment includes the identification of author gender in 55 languages (male, female, unknown gender or without author). Users can upload a list of male and female forenames supplementing and replacing the ones found in the dictionaries of the programme.[4] AVOBMAT currently identifies the Parts-of-Speech tags in 16 languages and produces different interactive visualizations and statistical tables of results. It also disambiguates and links recognized named entities to Wikidata, VIAF, and ISNI. The advantage of the latter, in the case of multi-language collections such as the European Literary Text Collection corpus or European Literary Bibliography, is that the recognized named entities such as names, locations, and organizations – appearing in distinct forms in different languages – are disambiguated and interconnected with one another, using Linked Data principles, via the relations stored in the growing multilingual Wikidata knowledge base.

(2.3) WORKFLOW 3: METADATA CROSSWALK FOR CITATION DATA PRODUCTION IN OPENCITATION (MORETTI & HEIBI 2023)

OpenCitations (Peroni & Shotton 2020), managed by the Research Centre for Open Scholarly Metadata at the University of Bologna, is a public service infrastructure organization that advocates for open science principles in bibliographic and citation data exposition. The data published by OpenCitations is first collected from several data sources, curated, remodelled to be expressed in compliance with the OpenCitations Data Model (OCDM) (Daquino et al. 2020), and exposed using Semantic Publishing and Referencing (SPAR) Ontologies and Semantic Web technologies (Berners-Lee & Kagal 2008). To guarantee unrestricted access to data, the collections are produced in easily reusable and interoperable formats, namely RDF, SCHOLIX (Burton et al. 2017), and CSV, and provided under a Creative Commons CC0 1.0 public domain licence. In sum, all data released by OpenCitations complies with both Force11 FAIR principles (Wilkinson et al. 2016), and recommendations by I4OC (Shotton 2017) that citation data should be structured, separable, and open.

4 When identifying the gender of a given author, the software also takes into consideration the detected language of the document since certain forenames can have different genders in different languages. In the revised version of programme, the user will have the opportunity to add forenames, which deliberately break gender norms (Bobbie short for both Robert and Roberta), as well as those of non-gender binary persons to the 'unknown' category at the preprocessing phase to avoid misgendering.

Citation data is crucial in calculating the impact of scientific and scholarly research, but also in verifying the soundness of theories, and making studies replicable. In SSH disciplines today, there is a noticeable delay in awareness regarding open science issues (Peels & Bouter 2018). For this reason, SSH researchers should be provided with the appropriate tools for facilitating the exchange of data concerning references among publications.

The proposed workflow offers a procedure for extracting citations and bibliographic data from a dataset and reshaping it in conformity with a new data model (Heibi et al. 2019). As concerns metadata, however, adopting a language-agnostic approach and a broad character encoding scheme – barring any restrictions imposed by data models – enhances the preservation of bibliodiversity over global uniformity in a dominant language.

This workflow describes a procedure for ingesting data into the OpenCitations infrastructure (Peroni & Shotton 2020), with the logical consequence that data will be represented in the OpenCitations Data Model (OCDM) (Daquino et al. 2020). Nevertheless, it also exposes general good practices for metadata crosswalks, fostering free and open-source tools, and disseminating an easily replicable procedure for metadata conversion between different data models.

The workflow consists of five steps:

1) Data source selection: identification of a data source, exposing structured data where bibliographic entities are identified by persistent identifiers (PIDs).
2) Development of a software plug-in for data conversion, extending the OpenCitations converter software to manage the bibliographic data of a new source; this is done by validating identifiers, abstracting data content, and re-structuring it according to the end data model (OCDM).
3) Production of metadata and citation data collections, structured according to OCDM.
4) Ingestion of the metadata collection in OpenCitations META (Massari et al. 2024), using META software to integrate new data into the infrastructure and map each persistent identifier to an OMID, and using the OpenCitations identifier to uniquely identify bibliographic records for deduplication purposes.
5) Production of OpenCitations citation data: starting from the collection of citations expressed as links between the PIDs of the entities involved, the OpenCitations META database is used to map each PID with its corresponding OMID and produce a dataset of OMID-OMID citation links; the collection is then published in RDF, CSV, and SCHOLIX format on open platforms under a CC0 waiver licence.

Looking forward, this workflow provides an easily replicable procedure for metadata crosswalks between different models, while also serving as a predisposition to help deal with the complexities of multilingual bibliographic data acquisition. Researchers and institutions working with non-English-language data sources may use this methodology as a model, as it may be adapted and customized to meet their individual needs.

(2.4) WORKFLOW 4: LODIFICATION OF BIBLIOGRAPHICAL DATA: ZOTERO TO WIKIBASE MIGRATION (LINDEMANN & KLAES 2023)

Research in the SSH may require a flexible software solution for bibliodata aggregation, curation, storage, transformation, and dissemination, while also ensuring interoperability with external tools and resources, and conforming to FAIR principles. Achieving these goals often requires major efforts to harmonize data, regarding both conceptual data modelling and data format migration. To date, there are very few universal standards among bibliodata providers for sharing metadata, most notably MARC21. Longstanding efforts to standardize inter-library data sharing have focused on addressing the needs of libraries and not on providing bibliodata to the wider research community (Possemato, Lionetti & Gazzarini 2021).

To a certain extent, bibliodata aggregation across multiple data sources, management, storage, and dissemination can be achieved using freely available reference management systems like Zotero. However, options for data curation, enrichment, and alignment to other resources remain very limited, especially when scaling up to larger datasets and spanning multiple descriptive

languages (see also 3.2). To address this need, *LODification of bibliographical data: Zotero to Wikibase migration* is a workflow for converting bibliodata collections from Zotero to a custom Wikibase: a free and open source software solution for storing, modelling, editing, and querying Linked Data, by which literal, field-based bibliodata information can be turned into a network of identifiable entities and standardized properties. The most well-known Wikibase instance is Wikidata, currently the largest free and open Knowledge Graph (Vrandečić & Krötzsch 2014). It contains millions of entities, describing, for example, scholarly publications and their authors. Custom Wikibase instances for managing bibliodata may or may not follow the manner in which the Wikidata community models entities, especially as this regards the semantics of bibliodata-describing properties. Use-case-specific modelling decisions are thus still possible, while otherwise it is preferable to follow Wikidata's modelling choices to ease upload to or federation with Wikidata, and to benefit from Wikidata tools such as Scholia (Nielsen, Mietchen & Willighagen 2017) or Author Disambiguator.

Building on previous work (Lindemann, Klaes & Zumstein 2019; Klaes 2021; Lindemann 2021), this workflow proceeds to describe the preliminary stages of ZotWb, a new python app for enabling the migration of bibliographic records from a Zotero group library to a custom Wikibase instance, in the process cleaning and harmonizing data, and finally aligning and enriching this data with Wikidata content. Wikibase items and Zotero records remain linked to each other.

ZotWb allows a fine-grained and self-defined mapping of Zotero data item types, data fields, and creator types to Wikibase properties and classes. In several cases, however, it suggests using Wikidata-aligned entities for this purpose. The OpenRefine tool is integrated in our workflow for carrying out data cleaning, as well as alignment with Wikidata (reconciliation) and subsequent enrichment. OpenRefine offers highly specialized functions for duplicate detection based on string similarity, allowing users to harmonize variant literals describing persons, places, or institutions, and to disambiguate entities with identical labels. ZotWb includes a reconciliation service for its own custom Wikibase, in addition to the default Wikidata reconciliation service built into OpenRefine, so that literals can be matched against entity labels on the ZotWb Wikibase and/or Wikidata.[5] Users may choose to import aligned Wikidata-entities, or to exploit the alignment using federated SPARQL queries.

The ZotWb tool and workflow around it rely entirely on free software, which makes it possible for anybody to share bibliographies as Linked Data.

(2.5) WORKFLOW 5: STUDIES ON SCIENCE AND HIGHER EDUCATION SYSTEM IN POLAND USING THE RAD-ON PLATFORM (TOMCZYŃSKA & KORYTKOWSKI 2023)

A workflow concerning studies on the science and higher education system in Poland illustrates how to gather and analyze data using the RAD-on platform (Tomczyńska et al. 2023). The RAD-on platform is a realization of the idea of open government data, providing reports, analysis, and raw data on science and higher education in Poland, and featuring data collected at the national level and at all universities and research institutions in Poland since 2013. This platform has been developed for scientists and scholars, students, entrepreneurs, policymakers, and journalists looking for reliable and up-to-date information on scientific and scholarly institutions and the research they conduct. In addition to its openness, RAD-on implements the FAIR concept, associated primarily with scientific datasets used in research.

The RAD-on platform is part of the largest national research information system on science and higher education in Europe, and its users gain access to regularly updated data on Polish science, interactive maps and charts, and REST APIs, which can be used to create custom data summaries. Freely available datasets contain information on scientific and scholarly institutions, activities in the fields of science and the arts, academic staff, academic promotion procedures, and bibliographical data about scientific and scholarly publications by Polish researchers.

The workflow includes five steps in which a user can perform a custom analysis of data using RAD-on:

5 The reconciliation service for the custom Wikibase is part of the ZotWb tool; it is a version of the Wikidata service (Delpeuch, 2020).

1) The first step involves coming up with research questions which may then be answered through the querying of RAD-on's data.
2) The second step underlines the importance of identifying the relevant dataset for a particular scientific inquiry.
3) The third step presents two possible approaches for retrieving data: the first involves basic export of data, which might be preferred by users with no background in IT; the second, more advanced approach involves the use of an application programming interface (API), and might be more suitable for users with basic programming skills, such as data scientists.
4) The fourth step identifies various procedures for checking the obtained data.
5) In the last step, some useful and popular tools for data analysis are listed and examples of completed data analysis using RAD-on are provided.

The workflow for RAD-on was designed to benefit all its users, regardless of their digital literacy or familiarity with the higher education and research landscape in Poland. Its authors were confronted with the challenge of striking a balance between providing accurate data analysis descriptions and ensuring that they will be understood by scientists working outside data science and related fields.

(3) FUTURE CHALLENGES FOR THE CREATION OF NEW BIBLIODATA WORKFLOWS: THE MULTILINGUALITY CHALLENGE

The bibliodata-related workflows presented in this paper describe practices pertaining to bibliodata curation and research. One issue they all share in common is the need for data harmonization. What these workflows illustrate is that, because of their differing functions, curatorial-infrastructural and research environments place different emphasis on data harmonization. In the following sections we investigate how the harmonization of multilingual data is being tackled in both cases. We argue that bibliodata-related research needs stronger support to face this challenge, given its more varied and less formalized nature of research activities. To that end, a general multilingual harmonization workflow is proposed for further specification and case-by-case adaptation.

(3.1) MULTILINGUALITY IN CONTEMPORARY BIBLIOGRAPHICAL DATA CURATION

In this section we present the typical scenario of a 'multilingual challenge' in the curatorial environment. Multilinguality issues are discussed among the librarian and bibliographical community, because MARC21 standards – the highly predominant data format for curation of bibliographical data in the librarian sphere – and related cataloguing rules (today mostly RDA or previous AACR2) do not sufficiently reflect multilingual aspects (Balula & Leão 2019; Riva 2022). MARC21 does not define specific rules on how to manage multilingualism at the general level, presupposing only one language to be used for processing each singular bibliographical or authority record, by virtue of the fact that the field for cataloguing language is defined as non-repeatable.[6]

Multilinguality issues might be partly solved at the level of specific types of fields. Fields for subject description (6XX fields) allow the parallel use of different descriptive systems. These systems are often based on the language (code of such systems is stored in subfield 2 or in the second indicator of the 6XX field).[7] In cataloguing praxis, this option for inclusion of multilingual variants is mostly used for the topical headings (field 650).

Hence, these descriptive systems make it possible to create the bibliographical record in multiple languages and to interconnect each given value with the language used for its cataloguing. For example, in the cataloguing praxis of the Czech National Library, authority file for topical headings includes translation of each heading into English (language version is then specified

6 Cf. https://www.loc.gov/marc/bibliographic/bd040.html (last accessed: 6 December 2023).

7 Cf. list of such descriptive systems https://www.loc.gov/standards/sourcelist/subject.html (last accessed: 27 January 2024).

by subfield 2). This allows it to include into each record topical headings in both Czech and English (czenas vs. eczenas subject authority files).

65007 $a argentinská próza $7 ph162925 $y 20. století $2 czenas

65009 $a Argentine prose literature $y 20th century $2 eczenas

Apart from authority files, the MARC21 standard as such offers at least a partial solution for dealing with multilinguality in the fields not related to authority systems. First of all, official documentation defines field 880 as 'Alternate Graphic Representation', which enables it to represent the values from basically any field in a different script (Arabic, Latin, Chinese/Japanese/Korean, Cyrillic, Greek and Hebrew are defined for this field with a specific code).[8] Library systems (ILS) allow for data to be displayed simultaneously in all of the recorded scripts. MARC21 partly stores the information about the original title of the translated book as well, but this information is processed inconsistently (fields 240 and 765) with respect to the cataloguing rules used. In each of these approaches, it is clear that application is far from reaching its full potential, so that sometimes the name of a translator (in 245$c together with other contributors) is the only evidence that a book is a translation (no source language nor original title is recorded) (Ivaska 2020).

The subject description of a given document is mostly provided in verbal form as well as in language-agnostic classification systems such as Dewey Decimal or Universal Decimal. Each value in these systems is represented by a specific numeric code, which makes it possible to group together the terms for related entities regardless of their verbal representation (e.g. each value related to language or literature starts with 8, etc.). These classifications are regularly mapped and translated into vernacular languages, which allows librarians and bibliographers to understand the content of the record without needing to be familiar with its language.

For the purposes of internationally understandable subject descriptions, various disciplinary-specific descriptive systems are continuously developed within internationally coordinated consortia. These consortia often curate the central version of a given descriptive system, which is available mostly in English. National nodes are responsible for its translation into specific vernacular languages. With regard to the level of their globalization, internationally used thesauri are more frequent in natural and medical sciences, where research outputs are published mostly in English (cf. internationally widely broadened thesauri like MeSH or Agrovoc).

Next to the aforementioned options defined in general MARC21 standards, multilinguality issues are quite often solved via local interpretations of this format, in particular by defining specific subfields that contain information about the language of the value in a given field. Such a solution was quite recently implemented at KBR (Royal Library of Belgium), which uses a special subfield @ for specification of the language (in ISO notation).[9]

110 $a Province du Brabant wallon $g Brabant wallon $c Wavre $@ fr-BE

Contrary to the Czech and Belgian solution, the Israel National Library uses subfield 9 for specification of the script:

710 $9 heb $a תליהק ידע הוהי לארשיב

710 $9 lat $a Watchtower Bible and Tract Society of New York

Unfortunately, these rules are defined only as a local extension of MARC21, frequently undocumented, differ among various databases, or often do not exist at all.

Multilinguality issues are taken into account also for the processing of records in other types of bibliographical databases. National bibliometric databases usually register data in vernacular languages as they are used primarily for evaluation purposes at the level of a given country. Nevertheless, at least the most important information on a particular document (title, abstract, keywords) is typically available in another language (predominantly English) for the international audience.

8 Cf. https://www.loc.gov/marc/bibliographic/bd880.html (last accessed: 6 December 2023).

9 Code @ is not used in the standard, so it is not conflicting with existing subfield definitions.

For the most part, references and bibliographical citations are currently language-agnostic. In spite of this, there are now a plethora of citation norms, which predominantly do not use any language specific abbreviations at all – each element (volume, issue etc.) might be defined by its position, or by specific interpunction in the given field –, or prefer to use internationally recognized abbreviations in English or Latin.

(3.2) MULTILINGUAL HARMONIZATION FOR BIBLIODATA-BASED RESEARCH

Bibliodata-driven research performed on multilingual datasets poses a significant challenge, and researchers commonly face the task of disambiguating language-sensitive values in metadata fields. They may wish, for example, to harmonize the names of publishing houses, authors, or subject descriptions that are expressed in different languages on different databases. These processes typically aim to yield reliable statistical results, for which the harmonization and deduplication of values are crucial. What is more important, researchers may wish to combine different single-language datasets, and their research might entail different standards for disambiguation or deduplication than those that exist in curatorial settings (Tolonen et al. 2019).

Multilinguality poses an even bigger challenge to the harmonization of metadata contents when creating international datasets. Having investigated the parameters of the bibliodata with regard to the availability of multilingual elements, we propose a dedicated workflow for the preparation of multilingual bibliographical datasets. Such a workflow might not contribute only to further data reuse between bibliodata processors from different countries, but also significantly accelerate comparative research on bibliographical data, providing a path for interconnecting and harmonizing datasets processed in different languages.[10]

The main goal of such a workflow is to ensure that all language-sensitive data elements are machine-readable while respecting multilinguality as such. To do so it is critical to identify language-sensitive data elements, apply multilingual PID systems, and enrich data in a way that respects multilinguality and the machine-readability of data.

From our perspective, enrichment of the bibliographical datasets with multilingual elements should generally be based on the following steps:

1) identification of metadata fields with multilingual content
2) identification of external sources useful for multilingual description (e.g. VIAF, Wikidata)
3) application of harmonization solutions
4) enrichment of data with PIDs or thesauri while respecting multilinguality of contents
5) documentation and publication of data

Key elements for implementing such workflows in the case of both research and curatorial praxis are:

1) the use of data models and standards that allow for direct representation of multilinguality (e.g. by a specific subfield/data element),
2) further development of multilingual bibliodata infrastructure descriptive thesauri and other knowledge bases with emphasis on incorporating values in less common languages; development of such datasets requires intensive cooperation with GLAM experts and bibliodata researchers,
3) preparation of dedicated software tools for multilingual enrichment of the given bibliographical datasets.

Creation of a workflow for multilingual enrichment of the bibliographical data represents only the first step towards this goal. We are suggesting that the SSH Open Marketplace, with its strategies to engage relevant stakeholders, might become an important vehicle for creating such workflows demanding international and cross-sectoral collaboration. What would be of

10 Multilinguality issues nowadays attract growing attention among humanities scholars. Recently in 2023, two dedicated multilinguality working groups have been founded: Multilingual DH within DARIAH-ERIC (January) and WG on Aligment Multilingual Vocabularies in Social Sciences & Humanities within RDA Alliance (October).

particular value is a detailed presentation of tools (e.g. platforms) emphasizing research reuse of multilingual dataset in the SSH Open Marketplace – namely inclusion of data harmonization workflows for such infrastructures that could serve as tested, scalable workflows. One example of such platforms is the European Literary Bibliography (literarybibliography.eu), a service aggregating bibliographic metadata, or projects such as Share-VDE. At the same time, of course, ensuring creation of multilingual datasets for research purposes will not solve any other existing issues around data quality or data coherence for any possible research or curational need.

(4) CONCLUSION

Describing and disseminating workflows' descriptions are an important tools for any academic community of practice, including the bibliographical data community. The role of the workflows is to promote harmonization and interoperability in a dispersed field and to provide an opportunity to better allocate resources and produce reproducible results of scientific works. To that end initiatives such as the SSH Open Marketplace are helpful as a vehicle for standardisation and dissemination.

Bibliodata-related workflows are needed at different levels of data creation and research, both for specific software features or data sources as well as for consolidating methodological aspects of bibliographical data curation at a very general level. Such workflows have to be applied in particular for cleaning, conversion, enrichment and harmonization of the bibliographical data.

Multilinguality is an important part of data harmonization both in curation and in research, yet there are no overarching, consistent workflows to deal with these challenges. This issue is especially prominent in those research activities which vary from case to case, and which are less formalized than curatorial processes. Multilinguality needs to be consistently respected on all possible levels of scholarly data processing. The 'Helsinki Initiative on Multilingualism in Scholarly Communication' advocates for supporting the infrastructure of scholarly communication in national languages, as this fosters opportunities for publishing locally relevant research and interacting with communities beyond academia, and in particular helps to interconnect locally relevant data to broader international cooperative networks. Our proposed workflow for the preparation of multilingual bibliographical datasets aims to satisfy the need for supporting and developing multilingual frameworks for data sharing.

FUNDING INFORMATION

This work has been supported by the DARIAH ERIC within Open Bibliodata Workflows project (DARIAH Theme 2022 Call), Humanities and Social Sciences Cluster of the Centre of Excellence for Interdisciplinary Research, Development and Innovation of the University of Szeged and by Ministry of Education, Youth and Sports of the Czech Republic within its activities on support of large research infrastructures (Czech Literary Bibliography III project; LM2023043).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR ROLES

Vojtěch Malínek (Project Administration, Supervision, Methodology, Writing), Tomasz Umerle (Project Administration, Supervision, Methodology, Writing), Edward Gray (Methodology, Writing), Ivan Heibi (Methodology, Writing), Péter Király (Methodology, Writing), Christiane Klaes (Methodology, Writing), Przemysław Korytkowski (Methodology, Writing), David Lindemann (Methodology, Writing), Arianna Moretti (Methodology, Writing), Charlotte Panušková (Methodology, Writing), Róbert Péter (Methodology, Writing), Mikko Tolonen (Methodology, Writing), Aldona Tomczyńska (Methodology, Writing), Ondřej Vimr (Methodology, Writing).

AUTHOR AFFILIATIONS

Vojtěch Malínek orcid.org/0000-0002-9553-5993
Institute of Czech Literature, Czech Academy of Sciences, Prague, Czech Republic

Tomasz Umerle orcid.org/0000-0002-7335-0568
Institute of Literary Research, Polish Academy of Sciences, Poznań, Poland

Edward Gray orcid.org/0000-0002-5201-1014
DARIAH-EU/IR* Huma-Num (CNRS UAR 3598), France

Ivan Heibi orcid.org/0000-0001-5366-5194
Department of Classical Philology and Italian Studies – FICLIT, University of Bologna, Bologna, Italy

Péter Király orcid.org/0000-0002-8749-4597
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, Germany

Christiane Klaes orcid.org/0000-0003-4870-4392
Technische Universität Braunschweig, Universitätsbibliothek, Braunschweig, Germany

Przemysław Korytkowski orcid.org/0000-0003-3504-7282
Faculty of Computer Science, West Pomeranian University of Technology in Szczecin, Poland;
National Information Processing Institute, Warsaw, Poland

David Lindemann orcid.org/0000-0002-8261-6882
UPV/EHU University of the Basque Country, Faculty of Arts, Vitoria-Gasteiz, Spain

Arianna Moretti orcid.org/0000-0001-5486-7070
Department of Classical Philology and Italian Studies – FICLIT, University of Bologna, Bologna, Italy

Charlotte Panušková orcid.org/0000-0002-3534-8440
Institute of Czech Literature, Czech Academy of Sciences, Prague, Czech Republic

Róbert Péter orcid.org/0000-0002-7972-4751
Institute of English and American Studies, University of Szeged, Hungary

Mikko Tolonen orcid.org/0000-0003-2892-8911
Department of Digital Humanities, University of Helsinki, Finland

Aldona Tomczyńska orcid.org/0000-0002-0832-8081
National Information Processing Institute, Warsaw, Poland

Ondřej Vimr orcid.org/0000-0002-9364-0685
Institute of Czech Literature, Czech Academy of Sciences, Prague, Czech Republic

REFERENCES

Abramitzky, R., & **Sin, I.** (2014). Book Translations as Idea Flows: The Effects of the Collapse of Communism on the Diffusion of Knowledge. *Journal of the European Economic Association*, *12*(6), 1453–1520. DOI: https://doi.org/10.1111/jeea.12093

Balula, A., & **Leão, D.** (2019). Is multilingualism seen as added-value in bibliodiversity?: A literature review focussed on business and research contexts. *ELPUB 2019 23rd edition of the International Conference on Electronic Publishing*. Jun 2019, Marseille, France. DOI: https://doi.org/10.4000/proceedings.elpub.2019.17

Berners-Lee, T., & **Kagal, L.** (2008). The Fractal Nature of the Semantic Web. *AI Magazine*, *29*(3), 29–34. DOI: https://doi.org/10.1609/aimag.v29i3.2161

Borek, L., Hastik, C., Khramova, V., Illmayer, K., & **Geiger, J. D.** (2021). Information Organization and Access in Digital Humanities: TaDiRAH Revised, Formalized and FAIR. Information between Data and Knowledge. *Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021)*. Regensburg, Germany, 8th–10th March 2021. pp. 321–332. Glückstadt: Verlag Werner Hülsbusch. DOI: https://doi.org/10.5283/epub.44951

Burton, A., Fenner, M., Haak, W., & **Manghi, P.** (2017). *Scholix Metadata Schema for Exchange of Scholarly Communication Links*. DOI: https://doi.org/10.5281/zenodo.1120261

Cichy, C., & **Rass, S.** (2019). An Overview of Data Quality Frameworks. *IEEE Access*, *7*, 24634–24648. DOI: https://doi.org/10.1109/ACCESS.2019.2899751

Coleman, C. N. (2020). Managing Bias When Library Collections Become Data. *International Journal of Librarianship*, *5*(1), 8–19. DOI: https://doi.org/10.23974/ijol.2020.vol5.1.162

Daquino, M., Peroni, S., Shotton, D., & **Massari, A.** (2020). *The OpenCitations Data Model*. 836876 Bytes. DOI: https://doi.org/10.6084/M9.FIGSHARE.3443876.V7

Delpeuch, A. (2020). Running a Reconciliation Service for Wikidata. In L.-A. Kaffee, O. Tifrea-Marciuska, E. Simperl & D. Vrandečić (Eds.), *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference (ISWC 2020)*. Virtual Conference, November 2–6, 2020. Available at: https://ceur-ws.org/Vol-2773/paper-17.pdf

Dombrowski, Q. (2014). What Ever Happened to Project Bamboo? *Literary and Linguistic Computing*, 29(3), 326–339. DOI: https://doi.org/10.1093/llc/fqu026

Dombrowski, Q. (2021). The Directory Paradox. In A. McGrail, A. D. Nieves & S. Senier (Eds.), *People, Practice, Power: Digital Humanities Outside the Center*. University of Minnesota Press. pp. 83–98. http://www.jstor.org/stable/10.5749/j.ctv2782dmw.9

Gray, E., Larrousse, N., Petitfils, C., Barbot, L., Fischer, F., Ďurčo, M., Illmayer, K., Condordia, C., Konig, A., Van Uytvanck, D., & **Buddenbohm, S.** (2021). D7.4 Marketplace – Data population & curation (v1.0). *Zenodo*. DOI: https://doi.org/10.5281/zenodo.5783358

Heibi, I., Peroni, S., & **Shotton, D.** (2019). Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics*, *121*, 1213–1228. DOI: https://doi.org/10.1007/s11192-019-03217-6

Ivaska, L. (2020). Identifying (indirect) translations and their source languages in the Finnish National Bibliography Fennica. *Electronic Journal of the KäTu Symposium on Translation and Interpreting Studies*, *13*, 75–88. Available at: https://research.utu.fi/converis/portal/detail/Publication/48815433

Juršic, M., Mozetič, I., Erjavec, T., & **Lavrač, N.** (2010). Lemmagen: Multilingual Lemmatisation with Induced Ripple-down Rules. *Journal of Universal Computer Science*, *16*(9), 1190–1214. DOI: https://doi.org/10.3217/jucs-016-09-1190

Király, P. (2019). *Measuring metadata quality*. PhD dissertation. University of Göttingen. DOI: https://doi.org/10.13140/RG.2.2.33177.77920

Klaes, C. (2021). *Linked OpenData-Strategien zum Identity Management in einer Fachontologie – Prototypische Entwicklung eines Workflows zur Aufbereitung und zum Interlinking von Personennamen*. Masterarbeit. Universität Hildesheim. Available at: http://nbn-resolving.org/urn:nbn:de:gbv:hil2-opus4-13789

Lahti, L., Marjanen, J., Roivainen, H., & **Tolonen, M.** (2019). Bibliographic Data Science and the History of the Book (c. 1500–1800). *Cataloging & Classification Quarterly*, *57*(1), 5–23. DOI: https://doi.org/10.1080/01639374.2018.1543747

Lindemann, D. (2021). Zotero to Elexifinder: Collection, curation, and migration of bibliographical data. *SiKDD 21 Slovenian KDD Conference*. October 4th, 2021, Ljubljana. DOI: https://doi.org/10.5281/zenodo.6896969

Lindemann, D., & **Klaes, C.** (2023). LODification of bibliographical data: Zotero to Wikibase migration. *Social Sciences & Humanities Open Marketplace*. Available at: https://marketplace.sshopencloud.eu/workflow/P0siWJ

Lindemann, D., Klaes, C., & **Zumstein, P.** (2019). Metalexicography as Knowledge Graph. *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Leipzig (Open Access Series in Informatics). pp. 19:1–19:8. DOI: https://doi.org/10.4230/OASIcs.LDK.2019.19

Massari, A., Mariani, F., Heibi, I., Peroni, S., & **Shotton, D.** (2024). OpenCitations Meta. *Quantitative Science Studies*, 1–31. DOI: https://doi.org/10.1162/qss_a_00292

Moretti, A., & **Heibi, I.** (2023). Metadata crosswalk for citation data production in OpenCitations. *Social Sciences & Humanities Open Marketplace*. Available at: https://marketplace.sshopencloud.eu/workflow/MHwO4l

Nielsen, F. Å., Mietchen, D., & **Willighagen, E.** (2017). Scholia and scientometrics with Wikidata. Joint Proceedings of the 1st International Workshop on Scientometrics and 1st International Workshop on Enabling Decentralised Scholarly Communication. *1st International Workshop on Scientometrics and 1st International Workshop on Enabling Decentralised Scholarly Communication, co-located with 14th Extended Semantic Web Conference (ESWC 2017)*. Portorož, Slovenia. Available at: http://ceur-ws.org/Vol-1878/article-03.pdf.

Oberbichler, S., Boroş, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H., & **Tolonen, M.** (2021). Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology*, *73*(2), 225–239. DOI: https://doi.org/10.1002/asi.24565

Peels, R., & **Bouter, L.** (2018). The possibility and desirability of replication in the humanities. *Palgrave Communications*, 4, 95. DOI: https://doi.org/10.1057/s41599-018-0149-x

Peroni, S., & **Shotton, D.** (2020). OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, *1*(1), 428–444. DOI: https://doi.org/10.1162/qss_a_00023

Péter, R. (2023). Multilingual analysis and visualization of bibliographic metadata and texts with AVOBMAT. *Social Sciences & Humanities Open Marketplace*. Available at: https://marketplace.sshopencloud.eu/workflow/RzvCOP

Péter, R., Szántó, Zs., Biacsi, Z., Berend, G., & **Bilicki, V.** (2024). Multilingual analysis and visualization of bibliographic metadata and texts with the AVOBMAT research tool. *Journal of Open Humanities Data*, *10*(23), 1–10. DOI: https://doi.org/10.5334/johd.175

Péter, R., Szántó, Zs., Seres J., Bilicki, V., & **Berend, G.** (2020). AVOBMAT: a digital toolkit for analysing and visualizing bibliographic metadata and texts. In G. Berend, G. Gosztolya & V. Vincze (Eds.), XVI. *Magyar Számítógépes Nyelvészeti Konferencia Szeged: Szegedi Tudományegyetem, Informatikai Intézet*. pp 43–55. Available at: https://acta.bibl.u-szeged.hu/67682/1/msznykonf_016_043-055.pdf

Péter, R., Szántó, Zs., Seres, J., Bilicki, V., & **Berend, G.** (2022). Az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz bemutatása. *Digitális Bölcsészet*, *4*, 3–28. DOI: https://doi.org/10.31400/dh-hun.2021.4.3530

Petitfils, C., Dumouchel, S., Larrousse, N., Gray, E. J., Barbot, L., Roi, A., Ďurčo, M., Illmayer, K., Buddenbohm, S., & **Parkola, T.** (2021). D7.5 Marketplace – Governance. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.5608487

Possemato, T., Lionetti, A., & **Gazzarini, A.** (2021). SVDE 2.0 Linked Data Management System and Entity Discovery Portal: Progress Status of New Developments. *BIBFRAME Workshop 2021*. 21 September. Available at: https://www.casalini.it/bfwe2021/web_content/2021/presentations/possemato_lionetti_jakobsen_gazzarini.pdf

Riondet, Ch., & **Romary, L.** (2018). The Standardization Survival Kit: for a Wider Use of Metadata Standards within Arts and Humanities. In R. Depoortere, T. Gheldof, D. Styven & J. Van Der Eycken (Eds.), *Trust and Understanding: the value of metadata in a digitally joined-up world*. Archives et Bibliothèques de Belgique – Archief- en Bibliotheekwezen in België; *106*, 55–62. Available at: https://hal.science/hal-02124679⟩

Riva, P. (2022). The Multilingual Challenge in Bibliographic Description and Access. *JLIS.It*, *13*(1), 86–98. DOI: https://doi.org./10.4403/jlis.it-12737

Roig Sanz, D., & **Fólica, L.** (2021). Big translation history. Data science applied to translated literature in the Spanish-speaking world, 1898–1945. *Translation Spaces*, *10*(2), 231–259. DOI: https://doi.org/10.1075/ts.21012.roi

Shotton, D. (2017). The Initiative for Open Citations. *OpenCitations Blog*. Available at: https://opencitations.wordpress.com/2017/04/06/the-initiative-for-open-citations/ (last accessed: 6 December 2023). DOI: https://doi.org/10.59350/jdwj8-at997

Tolonen, M., Roivainen, H., Marjanen, J., & **Lahti, L.** (2019). Scaling up bibliographic data science. In C. Navarretta, M. Agirrezabal & B. Maegaard (Eds.), *Digital Humanities in the Nordic Countries: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*. pp. 450–456. Available at: https://ceur-ws.org/Vol-2364/41_paper.pdf

Tolonen, M., Vimr, O., Király, P., & **Panušková, Ch.** (2023). Bibliographical Data Science: from Catalogues to Research Data. *Social Sciences & Humanities Open Marketplace*. Available at: https://marketplace.sshopencloud.eu/workflow/tE2HiC

Tomczyńska, A., & **Korytkowski, P.** (2023). Studies on science and higher education system in Poland using the RAD-on platform. *Social Sciences & Humanities Open Marketplace*. Available at: https://marketplace.sshopencloud.eu/workflow/pVVLop

Tomczyńska, A., Knapińska, A., & **Ostrowska, S.** (Eds.) (2023). *RAD-on: Reports, analyses, data*. Warsaw: National Information Processing Institute. ISBN 978-83-63060-26-8.

Vimr, O., & **Rosiński, C.** (2022). Česká literatura ve světě: Možnosti mapování ve velkém rozsahu (1820–2020). *Česká literatura*, *70*(6), 711–734. DOI: https://doi.org/10.51305/cl.2022.06.03

Vrandečić, D., & **Krötzsch, M.** (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. DOI: https://doi.org/10.1145/2629489

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. DOI: https://doi.org/10.1038/sdata.2016.18

TO CITE THIS ARTICLE:
Malínek, V., Umerle, T., Gray, E., Heibi, I., Király, P., Klaes, C., Korytkowski, P., Lindemann, D., Moretti, A., Panušková, Ch., Péter, R., Tolonen, M., Tomczyńska, A., & Vimr, O. (2024). Open Bibliographical Data Workflows and the Multilinguality Challenge. *Journal of Open Humanities Data*, 10: 27, pp. 1–14. DOI: https://doi.org/10.5334/johd.190

Submitted: 12 December 2023
Accepted: 30 January 2024
Published: 18 March 2024