УДК 01:004.6
https://doi.org/10.20913/1815-3186-2021-2-103-107

Обзорная статья

Activities and projects of the Bibliographical Data Working Group (DARIAH-EU)[1]

Jakub M. Łubocki, **Elżbieta Herden**, **Dorota Siwecka**

Łubocki Jakub Maciej,

Department of Publishing Art, National Museum in Wrocław, Powstańców Warszawy str., 5, 50-153, Wrocław, Poland,

MA, Assistant

ORCID: *0000-0002-1957-0682*
e-mail: *jakub.lubocki@mnwr.pl*

Herden Elżbieta,

Institute of Library and Information Sciences, University of Wroclaw, Uniwersytecki str., 1, 50-137, Wrocław, Poland,

PhD, Assistant Professor

ORCID: *0000-0001-5981-3725*
e-mail: *elzbieta.herden@uwr.edu.pl*

Siwecka Dorota,

Institute of Library and Information Sciences, University of Wroclaw, Uniwersytecki str., 1, 50-137, Wrocław, Poland,

PhD, Assistant Professor

ORCID: *0000-0003-1649-7579*
e-mail: *dorota.siwecka@uwr.edu.pl*

Abstract. The material aims to introduce the Bibliographical Data Working Group – part of the Digital Research Infrastructure for the Arts and Humanities (DARIAH-EU) – and to feature projects fulfilled by this group. This working group since 2019 brings together specialists (as of today: 38 researchers) from a number of different countries and the main goals of the group are to foster the development of cooperation between bibliographies and serve as a platform for knowledge exchange aimed at bringing together creators of bibliographical data, scholars interested in using those resources in data-driven research, and theorists of bibliography and documentation. In the presentation two projects are described in detail: 1. the report "An analysis of the current bibliographic data landscape in the humanities: Bibliodata curation, research, and collaboration; 2. the project "Multilingual encyclopaedic dictionary of types of documents". The purpose of the article is not only to describe these projects but above all to invite Congress members to collaborate on them.

Keywords: Bibliographical Data Working Group, BiblioDataWG, "An analysis of the current bibliographic data landscape in the humanities", "Multilingual encyclopaedic dictionary of types of documents"

Citation: Łubocki J. M., Herden E., Siwecka D. Activities and projects of the Bibliographical Data Working Group (DARIAH-EU). Bibliosphere. 2021. № 2. P. 103–107. *https://doi.org/10.20913/1815-3186-2021-2-103-107*.

Received 20.05.2021
Revised 25.05.2021
Accepted 30.05.2021

[1] Информация подготовлена на основе доклада на III Международном библиографическом конгрессе, 27–30 апреля 2021 г.

Деятельность и проекты Рабочей группы по библиографическим данным

Я. М. Любоцкий, Э. Херден, Д. Сивецка

Любоцкий Якуб Мацей,

Отдел издательского искусства,
Национальный музей Вроцлава,
ул. Варшавских повстанцев, 5,
50-153, Вроцлав, Польша,
магистр, ассистент

ORCID: *0000-0002-1957-0682*
e-mail: *jakub.lubocki@mnwr.pl*

Херден Эльжбета,

Кафедра библиотечных
и информационных наук,
Вроцлавский университет,
ул. Университетская, 1, 50-137,
Вроцлав, Польша,
кандидат наук, доцент

ORCID: *0000-0001-5981-3725*
e-mail: *elzbieta.herden@uwr.edu.pl*

Сивецка Дорота,

Кафедра библиотечных
и информационных наук,
Вроцлавский университет,
ул. Университетская, 1, 50-137,
Вроцлав, Польша,
кандидат наук, доцент

ORCID: *0000-0003-1649-7579*
e-mail: *dorota.siwecka@uwr.edu.pl*

Аннотация. Цель статьи – представить Рабочую группу по библиографическим данным, которая является частью Цифровой исследовательской инфраструктуры для искусств и гуманитарных наук (DARIAH-EU), проинформировать о проектах, выполненных группой, и предложить участникам Конгресса сотрудничество по их реализации. С 2019 г. группа объединяет специалистов (на сегодняшний день – 38 исследователей) из ряда стран. Основные цели группы заключаются в содействии развитию сотрудничества между библиографиями, она служит платформой для обмена знаниями, направленной на объединение создателей библиографических данных, ученых, заинтересованных в использовании этих ресурсов в своих исследованиях, основанных на данных, и теоретиков библиографии и документации. В статье подробно описаны два проекта: «Анализ современного ландшафта библиографических данных в гуманитарных науках: кураторство библиоданных, исследования и сотрудничество» и «Многоязычный энциклопедический словарь типов документов».

Ключевые слова: Рабочая группа по библиографическим данным, BiblioDataWG, «Анализ современного ландшафта библиографических данных в гуманитарных науках», «Многоязычный энциклопедический словарь типов документов»

Для цитирования: Любоцкий Я. М., Херден Э., Сивецка Д. Деятельность и проекты Рабочей группы по библиографическим данным // Библиосфера. 2021. № 2. С. 103–107. *https://doi.org/10.20913/1815-3186-2021-2-103-107*.

Статья поступила в редакцию 20.05.2021
Получена после доработки 25.05.2021
Принята для публикации 30.05.2021

Introduction

DARIAH-EU is an acronym of Digital Research Infrastructure for the Arts and Humanities[2]. It is a network of people, expertise, information, knowledge, content, methods, tools, and technologies useful to enhance and support digitally-enabled research and teaching across the Arts and Humanities. This network integrates researchers and projects from across Europe, enabling transnational and transdisciplinary approaches. The Bibliographical Data Working Group (BiblioData WG) is the best example of it. The idea of the BiblioData WG results from the ongoing cooperation between Czech Literary Bibliography[3] (Institute of Czech Literature, Czech Academy of Sciences), and Polish Literary Bibliography[4] (Institute of Literary Research of the Polish Academy of Sciences), that have been working closely on joint projects in recent years. We believe that there is a need for knowledge exchange in the field of bibliographic data in the humanities and thus a DARIAH-EU Working Group should be the best as a platform for data providers and researchers working with bibliographical data.

1. Mission of Bibliographical Data Working Group

For decades bibliographies have been among the most widely-used sources of scientific information for researchers in the humanities. There is a variety of bibliographical projects in the humanities all over Europe, which have been transforming bibliographies into modern digital scholarly content

[2] DARIAH-EU. URL: dariah.eu/ (accessed 01.05.2021).

[3] Česká literární bibliografie. *Ústav pro českou literaturu.* URL: ucl.cas.cz/cs/oddeleni/clb (accessed 01.05.2021).

[4] Pracownia Bibliografii Bieżącej. *Instytut Badań Literackich Polskiej Akademii Nauk.* URL: ibl.waw.pl/pl/o-instytucie/pracownie-i-zespoly/pracownia-bibliografii-biezacej (accessed 01.05.2021).

services. That projects were accomplish to provide bibliographic information for researchers in a faster and easier way. By the way, they started to function as sources of datasets for data-driven research in Arts and Humanities. Marking the shift from traditional bibliographic reference books and simple databases to research datasets, the digital humanities have increasingly influenced the bibliographical work – from the digitisation of information, through software tools for creating and processing metadata, to the data-based research on bibliographical data. The focus of the Working Group would be to foster the development of cooperation between bibliographies and serve as a platform for knowledge exchange aimed at bringing together creators of bibliographical data, scholars interested in using those resources in data-driven research, and theorists of bibliography and documentation.

This is a tentative list[5] of topics that could be covered by the group:

1. Data preparation for advanced research: managing the adaptation of bibliographical data to data-driven research (what types of datasets are useful in contemporary research, and how data producers might cooperate with data researchers to build datasets that are well-suited for their research, etc.).

2. Facilitating international data-based cooperation: analysis of European bibliographical datasets; working towards adopting standards and introducing data processing procedures that would allow for future data exchange and combining datasets from different countries (issues of multilingualism, mapping of the subject headings, authority files, descriptive thesauri, etc.).

3. Methodological issues: how bibliographical data is created nowadays (metadata standards and formats, subject headings vocabularies, issues of defining the scope of subject bibliographies, selection of processed sources, processing of manuscripts, online or non-textual documents, etc.).

4. Publishing bibliographies: how bibliographical data is published (online services and interfaces, Linked Open Data, challenges of data publishing, and data documentation, including i.e. data papers, and the issues of open data standards, etc.).

5. Processing bibliographic metadata: tools for data quality improvement (data reconciliation, parsing, conversion, database development, software for data analysis, etc.).

6. Remediation of bibliographical information: how the robust resources of hand-written and printed bibliographies can be digitized and turned into structured data (methodologies of retro-conversion: digitization, NLP methods, machine learning, manual work based on crowdsourcing, etc.).

7. Development of user-oriented services: assessment and implementation of services responding to the users' needs (links to the full text or digital libraries, tools for export of the data in various formats, tools for data analyses, visualisation, network analysis, and statistical interpretation, bibliographical tutorials, etc.).

[5] Dariah-EU Bibliographical Data Working Group: poster. URL: bit.ly/2LdsK8G (accessed 01.05.2021).

2. Projects of Bibliographical Data Working Group in its history

Our group began to form on May 15, 2019. On September 13, 2019 the group was officially approved. The main goals of the group are to foster the development of cooperation between bibliographies and serve as a platform for knowledge exchange aimed at bringing together creators of bibliographical data, scholars interested in using those resources in data-driven research, and theorists of bibliography and documentation. Now this group brings together specialists (as of today: 38 researchers) from a number of different countries across Europe (Belgium, Bulgaria, Czech Republic, Finland, Germany, Great Britain, Hungary, Ireland, Italy, Netherlands, Norway, Poland, Spain, Switzerland). We have two heads of the BiblioDataWG: Tomasz Umerle from the Polish Academy of Sciences and Vojtěch Malínek from the Czech Academy of Sciences. November 12, 2019 started the project about the multilingual dictionary of types of documents; February 25, 2020 started the report about the bibliographical data landscape; and the newest one – funding application in Collaboration of Humanities and Social Sciences in Europe programme – started at March 16, 2021. Because the last one is at the very beginning, we would like to say something more only about the two first projects.

2.1. "Multilingual encyclopaedic dictionary of types of documents" project

This project is for publishing the multilingual encyclopaedic dictionary facilitating exchange metadata describing one of the features of document – its form of perpetuation, i.e. type of document. Two of the barriers (or source of noise) are the language and/or terminological barrier. Words labeling type of document are most often terms and their translation from one language into the second language has to be strict. Therefore it has to use exact equivalent (not relative equivalent nor neologism). The issue is not the only choice of proper equivalent, but also a danger of false equivalency made by anisomorphism. This happens when on the lexical level equivalent is correct, but is in fact incorrect due to their semantic, grammatical, or cultural differences. Calques also make difficulties with appropriate data exchanging. The projected source of information – a multilingual encyclopaedic dictionary,

so not only with equivalents but also with definitions and annotations for individual languages – will remove these obstacles in the field of document types, because this way it will describe correspondence of designations for identical concepts. So far we established the list of entries for Polish. The list contains 278 terms in 12 main classes[6]. This dictionary aims to facilitate the exchange of bibliographical data between many languages in the scope of naming types of documents.

The project involves creating a normative source of information in the form of a dictionary informing about the meaning of individual terms identifying particular types of documents. The features determining the nature of this source of information require an instance. It is assumed that the forthcoming source of information will be a multilingual, controlled, annotated, encyclopaedic dictionary in the scope of types of documents[7]:

1. Multilingual means that the scope of our dictionary should cover as many languages as possible, especially those less common. The assumed electronic form of presentation of the dictionary will support the development of the project in time and will not limit the possibility of adding new resources and languages.

2. Controlled means that our dictionary will not newly (in an original way) define its vocabulary but will mediate in the transmission of already established lexicographic knowledge. "Original" definitions we are stating as those which make a new, proprietary spin of an object, preceded by proper research and reflection. Meanwhile, in this dictionary, the definitions will be taken from other, authoritative, existing sources.

3. Annotated means that the microstructure of our dictionary will contain, except the classic lexicographic triad – a *definiendum* (a term which has to be defined) + conjunction + a *definiens* (term's definition), also annotations containing a commentary – wherever the proper meaning of the term is not clear from the simple definition.

4. Encyclopaedic means that the dictionary's entries will not present linguistic facts (grammar, syntactic or stylistic features – as linguistic dictionaries do), but will describe the meaning of the entry words.

5. Dictionary means that the created source will only describe lexicographic units, not the objective meaning of words.

6. In the scope of types of documents means, that the dictionary will capture terms that are connected with the form of document perpetuation. Terms connected with publishing form or genre basically will be not captured in our dictionary, although some of the more common genres will be included because they are important as a search or selection criterion for documents in bibliographic databases.

This kind of dictionary, by preventing the incorrect equivalence of terms in the field of document types, will help in formulating correct oral and written presentations, descriptions of documents used in information resources, their correct translation, and building search tools on their basis – in particular, bibliographic databases and digital libraries, where they are used as a search or selection criteria for materials. This project has an independent website[8] you can visit to know more about the scientific basis of it. More details can be found at an article[9] too.

2.2. "An analysis of the current bibliographic data landscape in the humanities" report

The second project is the report "An analysis of the current bibliographic data landscape in the humanities: Bibliodata curation, research, and collaboration" is still a growing paper describing bibliographical data in the contemporary world from many sides (sometimes for the first time and/or in the unique perspective). This report is aimed at all the actors active in the humanistic bibliodata field, especially at the public sector stakeholders who could benefit from the report's insights and engage in collaborations to reach common goals. The report consists of three main parts:

1. Introduction defining bibliographical data (with a description of bibliodata producers, usage, and users, putting a few words about standardization and semantization of bibliodata).

2. The analysis of the current bibliodata landscape (with a description of dimensions of the contemporary bibliodata landscape and main bibliodata stakeholders; there is also defined influx of data, automation, and data-driven research).

3. The discussion of the main challenges and opportunities that the current bibliodata landscape poses for public stakeholders (infrastructure, open science, data management/scope and documentation of data).

In this report, the bibliodata landscape is discussed as a dynamic ecosystem of multiple categories of stakeholders who produce, process, and use

[6] Wielojęzyczny słownik encyklopedyczny typów dokumentów. Siatka haseł. Problem ekwiwalencji = Multilingual encyclopaedic dictionary of types of documents. The list of terms. The problem of equivalence. URL: cutt.ly/SnhgY6U (accessed 01.05.2021).

[7] Wielojęzyczny słownik encyklopedyczny typów dokumentów. Specyfikacja wstępna projektu – zaproszenie do współpracy = Multilingual encyclopaedic dictionary of types of documents. Initial specification of the project – call for participation. URL: cutt.ly/QnhgsfD (accessed 01.05.2021).

[8] Dictionary of types of documents. Encyclopaedic, multilingual, annotated and controlled source of information about terms describing types of documents. URL: typesofdocuments.wordpress.com (accessed 01.05.2021).

[9] Herden E and Łubocki JM (2021) Problem ekwiwalencji terminów w międzynarodowej komunikacji naukowej w kontekście projektu wielojęzycznego słownika encyklopedycznego typów dokumentów. *Academic Journal of Modern Philology* 11: (in print).

diverse bibliodata datasets and services. The authors of the report believe that the public bibliodata stakeholders in the humanities need to engage in a long-term effort to build a joint agenda, which would support the systematic solutions to the extensive work that is being done in the bibliographic data field by the public actors. Because of this in the report are described two dimensions to organize the contemporary bibliodata landscape: The Public/Private dimension and The Produce/Use dimension, which appeared as a very useful axis for talking about changes in bibliodata. The Public/Private dimension organizes stakeholders and initiatives mainly according to their legal status, whether private or public. Furthermore, it captures the availability of funding, in full or part, and the type of business model, if any (for-profit, not-for-profit, foundation, etc.). Lastly, it considers the data policy: whether bibliodata is proprietary or open, and with which licenses. The Produce/Use dimension organizes stakeholders and initiatives according to what they actually do with bibliodata. On the one hand, we have the production of new bibliodata, moving towards backend processing and enrichment of bibliodata and other services, towards frontend processing and enrichment of bibliodata to its direct use.

Conclusion

We hope that this short article encouraged some of you to take part in our work. It is very important for us to connect different members of the bibliographic community: from different institutions, from different countries, with different experiences. Only this will allow us to study bibliodata from a sufficiently broad perspective. If you will feel, that our WG is suitable for you, please – contact us: tomasz.umerle@ibl.waw.pl or malinek@ucl.cas.cz. You can visit our Twitter[10] or WG DARIAH-EU subpage[11]. If you are not sure – let us know: we will send you all details and if you are not interested – let your friends know about our ideas. We hope that we could collaborate soon with you together during current and future Bibliographical Data Working Group projects!

[10] Bibliographical Data DARIAH-EU Working Group. URL: twitter.com/bibliodataWG (accessed 01.05.2021).

[11] Bibliographical Data (BiblioData). URL: dariah.eu/activities/working-groups/bibliographical-data-bibliodata/ (accessed 01.05.2021).