Open Research Data in Institutional Repositories

Approach and Practice of Polish Universities

Anna Wałek
University Library
Gdańsk University of Technology
Gdańsk, Poland
e-mail: anna.walek@pg.edu.pl

Dorota Siwecka
Institute of Information and Library Science
University of Wrocław
Wrocław, Poland
e-mail: dorota.siwecka@uwr.edu.pl

***Abstract* — Research data is a type of data that is gaining importance in terms of both scientific value and commercialization potential. Research data is one of the essential parts of the entire data universe. It can be big data, small data, raw or processed data. In order to improve functionality, searchability and indexation, the data is associated with metadata and linking data, allowing it to be linked with other resources, databases, and reference data. Complex calculations are carried out on large amounts of research data generated as part of scientific experiments. An increasing number of universities prepare and implement an Open Access policy, following the latest international guidelines. To meet the conditions of these arrangements, scientific institutions in Poland implement projects to ensure open access to their employees' publications and Open Research Data. This paper presents the results of an analysis of four university repositories in Poland. The study covers, inter alia, the type of published data, database scope, institutions participating in the project implementation, and the possibility of re-use of open data (data sharing format, metadata format, applied ontologies and data schemas, information on linked data, legal licenses).**

Keywords-Open Research Data; Linked Data; institutional repository; data repository.

I. INTRODUCTION

Research data comprises all data produced as a result of scientific research, regardless of whether we are dealing with sciences, natural sciences, social sciences, or humanities. Data can be the results of analyses, measurements, surveys, historical sources, laboratory logs, images or samples, among others. It can take both digital and analog forms. The significance and quality of research data are widely discussed topics in the scientific community and, nowadays, researchers have become increasingly open in sharing and disseminating research findings. The amount of data is continuously growing, and opening research data is not only a good practice, but an obligation. Leading research funding agencies, including the European Commission, implement policies of sharing scientific data associated with publicly financed projects and place increasing emphasis on proper management of the research results they fund. Many guidelines and recommendations in this area have been published (including guidelines from the European Commission, research funding institutions, international organizations and associations). In the big data era, society, business, and research need quality data to evolve. One of the best ways to ensure data quality is managing research data under FAIR principles. FAIR data is data which meets principles of Findability, Accessibility, Interoperability and Reusability [1] (hence the acronym).

Horizon 2020 introduced a policy that obliges grantees to publish research results in the Open Access (OA) model. A pilot of Open Research Data has also been introduced. The announced changes in Horizon Europe will include the need to ensure OA by depositing work in a repository immediately after publication, without any embargo period, leaving copyright with the author, and using open licenses, such as Creative Commons licenses. In the field of research data, the principle of 'as open as possible, as closed as necessary' will be maintained, according to which, by default, data should be made available in an open manner. Data can only remain closed in justified situations, with considerable attention being paid to research data management.

On 4 April 2019, the European Parliament adopted a Directive on open data and re-use of public sector information [2]. The document was updated in June 2019. EU member states were given two years to adapt their laws to the new requirements. This means, among other things, that regulations for the sharing of research data by scientific institutions must be in place by 2021. The adoption of the Directive will also affect the Polish legal order, as its implementation will require amendments to the Act of 25th February 2016 on the re-use of public sector information [3]. It is worth paying attention to the fact that easy integration, linking, and re-use of data across silos is enabled by, among others, Linked Open Data (LOD) technology [4].

The EU Directive in points 27 and 28 speaks explicitly about the motives and ways to make research data open:

> 'The volume of research data generated is growing exponentially and has potential for re-use beyond the scientific community. In order to be able to address mounting societal challenges efficiently and in a holistic manner, it has become crucial and urgent to be able to access, blend and re-use data from different sources, as well as across sectors and disciplines. Research data

 ISBN: 978-1-61208-842-6

includes statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. It also includes meta-data, specifications, and other digital objects. (…) Beside open access, commendable efforts are being made to ensure that data management planning becomes a standard scientific practice and to support the dissemination of research data that is findable, accessible, interoperable, and reusable (FAIR principle). (28) For the reasons explained above, it is appropriate to set an obligation on Member States to adopt open access policies with respect to publicly funded research data and ensure that such policies are implemented by all research performing organizations and research funding organizations. (…) certain obligations stemming from this Directive should be extended to research data resulting from scientific research activities subsidized by public funding or co-funded by public and private-sector entities. (…) However, in this context, concerns in relation to privacy, protection of personal data, confidentiality, national security, legitimate commercial interests, such as trade secrets, and to intellectual property rights of third parties should be duly taken into account, according to the principle "as open as possible, as closed as necessary"' [2].

The second section of this paper describes the legal and organizational situation of research data in Poland (based on European Union regulations). It describes the method of calculating the number of repositories in Poland and indicates the reason for selecting those repositories described in the study. The third section describes the method adopted in the study and the individual elements analyzed. This section is divided into the following parts: providers and cooperation; types of provided information; publications and Research Data licenses; repositories and FAIR principles; research Data format; Linked data and semantization; access points; metadata license; used schemas and ontologies. The last section of the paper describes the preliminary conclusions of the research and, above all, the indicated areas that require in-depth analysis and further research. As this publication is an introduction to the research and indicates problems that require analysis, topics have been identified that have not been researched yet, but will be analyzed in the future.

II. Open Research Data at Polish Universities

Based on the Directive mentioned above, the Ministry of Digitalisation of the Republic of Poland prepared in 2020 a draft law on open data and re-use of public sector information [5], which will be introduced in 2021. As a consequence of introducing the Act, guidelines and recommendations for universities will also be presented.

The National Science Centre – NSC (the Polish national funder) also supports the idea of making research data accessible. From 2019, grant applications submitted in NSC competitions must be accompanied by a Data Management Plan (DMP), which specifies, among other things, what research data will be used, produced and made available within the project, as well as how they will be stored.

The academic year 2019/2020 in Poland began with 133 public universities, including 19 universities, 16 technical universities and 13 universities of natural science, economics and pedagogy, 9 medical universities, and approx. 240 non-public universities [6]. According to the statistical data of the Polish Central Statistical Office for 2019, 93,100 academic teachers were employed in various research and teaching positions [7].

Considering the large number of universities in Poland, the number of repositories where their employees can share data is low. The Directory of Open Access Repositories (Open DOAR) database shows 119 open repositories in Poland [8]. After analyzing the list of repositories, it can be concluded that most of the platforms registered there are digital libraries, which are not typical repositories. However, they often contain, in addition to cultural heritage objects, scientific publications and sometimes research data. The content of these databases would require additional separate research. However, considering only the research data repositories registered in the Re3Data.org database, we can see only 8 instances [9]. This is a much better source of information than the Open DOAR, because the repository must meet specific criteria to be registered in that database. Registered repositories are: Copernicus' Heliograph, Open Forest Data, Kujawsko-Pomorska Digital Library (as the only one of the digital libraries), CLARIN-PL, MOST Wiedzy Open Research Data Catalog (Bridge of Data), GEOMIND (no longer available since 2018), RepOD, Polish Platform of Medical Research.

III. The Basis of the Analysis

Repositories chosen for this study had to meet two main criteria: they should provide access to Open Research Data and implement LOD technology to increase accessibility, improve quality and increase the possibility of reusing science resources according to international recommendations. According to a survey concerning Polish LOD projects conducted in January 2021, only four institutional repositories functioning in Polish universities meet the LOD criterion [10]: AZON 2.0 (the Atlas of Open Science Resources) [11], the Bridge of Data (MOST Wiedzy Open Research Data Catalog – Gdańsk University of Technology) [12], RUJ (Repository of Jagiellonian University, Cracow) [13] and PPM (Polish Platform of Medical Research)[14]. Those examples show a diverse approach to providing access to information about publications, research, and Open Data conducted by Polish researchers. Information on the projects comes from the data posted on the official websites of the projects and from questionnaires sent to suppliers between January 21 and 27, 2021.

A. Providers and cooperation

The selected examples show that, in the Polish landscape, there are different approaches to creating repositories. We can

list here two main types: 1) projects implemented by a single university and 2) projects created cooperatively by more than one institution. In the first group, we can indicate a) projects designed for the university's employees only, e.g. RUJ; b) projects implemented by a single university but open to all research communities in the country, e.g., Bridge of Data, which the Gdańsk University of Technology introduced, but where any interested researcher can create a profile and share information about research conducted and publications. In the second group, we can identify a) projects implemented by "homogeneous" universities, e.g. PPM, which is a result of the cooperation of 8 medical institutions (7 universities with Wrocław Medical University as a leader and 1 Institute of Occupational Health); and b) projects implemented by different kinds of institutions, like AZON 2.0 created by Wrocław University of Science and Technology (leader), the Wrocław University of Environmental and Life Science, Wroclaw Medical University and Systems Research Institutes of Polish Academy of Science. Further analysis of remaining Polish repositories should lead to the creation of a typology that includes them.

B. Types of provided information

Analyzed projects show that these are not only typical institutional repositories. In addition to data on employees' publications and Ph.D. theses, they also provide information about researchers' profiles, conferences, research infrastructure, research projects, research data, and sometimes about inventions (detailed information on the types of provided data are summarized in Table I).

TABLE I. TYPES OF PROVIDED DATA

	AZON 2.0	Bridge of Data	PPM	RUJ
Researchers' profile	+	+	+	–
Publications' metadata	+	+	+	+
Full text of publications	+	+	+*	+
Research data	+	+	+	+
Conferences	conference recordings, news, etc.	conference ranking according to the list of the Polish Minister of Science and Higher Education; conference proceedings are listed in researchers' profiles	+	conference proceedings and conferences organized by Jagiellonian University
Projects/ Grants	–	+	+	+
Institutional publishing output	?	+	+ (journals)	Publications of Jagiellonian Library; series publications of Jagiellonian University
Ph.D. thesis	+	+	+	+
Inventions	?	+	+ (patents)	–
Research infrastructure	+	+	+	–

* Only those with Open Access and deposited in the repository (ca. 17 % of all publications records in the database).

C. Publications and Research Data licenses

Because all of the analyzed repositories provide access to publications and research data, they must define the licenses under which they are available online. In every case, information about the type of license is given in addition to the metadata. Moreover, the information about the license usually links to the full text of the license on the Internet. The exception is the PPM platform, where clicking the license info directs us to the list with results of all publications deposited in the PPM with the same license. The most common are Creative Commons licenses. If the full text of the publication is available outside of the repository, an appriate link is also given. Most of the suppliers also provide special tutorials for researchers about licenses.

D. Repositories and FAIR principles

As mentioned in the introduction, one of the best ways to ensure data quality is managing research data under FAIR principles. The analyzed repositories in their tutorials for researchers write about FAIR principles [15; 16; 17]. The only repository that doesn't provide such information on the project's website is AZON 2.0. The actual check of whether a given repository's resources meet the FAIR rules is the so-called FAIRification, which consists of the technical review of data using dedicated IT tools [18]. This is an issue that requires a separate study.

E. Research Data format

Research Data are mostly available in many formats, due to the multidisciplinary nature of the repositories. The most common are CSV and PDF. But, for example, the set of formats of the AZON 2.0 repository also includes DOC/X, PPT/X, TIFF, JPG, PNG, GIF, PSD, WRP, STL, OBJ, PTS, PTX, TXT, JSON, MP4, MOV, STR, VTT, STL, AVI, MP3, MD, and many technical formats for datasets. Depositaries are committed to sharing files in open formats and/or providing alternative formats to closed formats.

F. Linked data and semantization

Three of the four analyzed repositories provide information about LOD or semantization of their collection on their websites (AZON 2.0 on the main page [11] after choosing "How to further explore resources"; Bridge of Data

 ISBN: 978-1-61208-842-6

[19]; PPM [20]). Information about LOD implementation in the RUJ project was received in the questionnaire.

All of the repositories implemented linking data within the projects. In AZON 2.0, metadata are connected also to external datasets such as Geonames, Wikidata, plWordnet, General Multilingual Environmental Thesaurus (GEMET), Medical Subject Headings (MeSH), AgroVOC, International Plant Names Index (IPNI), The Plant List, World Flora Online (WFO), Ośrodek Przetwarzania Informacji (OPI), ORCID, Polska Bibliografia Naukowa (PBN), ResearcherID, Scopus. Information about linking data to external datasets in the PPM was not provided.

G. Access points

All of the projects provide API for data re-use, and in most cases, this requires filling an access form to receive an access code or token. As to API's data format, there are many variations: the most common are JSON, JSON-LD, RDF, XML, and Turtle. The exception is the PPM platform, which lists all available formats of data and protocols on its website [21].

H. Metadata license

Almost none of the projects implemented by Polish institutions provide information about metadata licenses. The Bridge of Data uses the Creative Commons CC-BY license for metadata, but that information is not visible on the project's website. At RUJ, the suppliers didn't apply any license for metadata because, in their opinion, "it was not necessary".

Of course, the issue of a metadata license is strongly related to the applicable law in the country, and this should be the basis for further research.

I. Used schemas and ontologies

Creating an extensive database that provides access to diverse content is a big challenge. It requires, for instance, the development of a data entry schema. The analyzed projects mainly created schemas based on other schemas and ontologies already available on the Web. The most common are: Schema.org, datacite, Friend of a Friend (FOAF), Bibliographic Ontology (BIBO), Dublin Core (DC), Bibliographic Framework (BIBFRAME), Metadata Authority Description Schema (MADS), Gemeinsame Normdatei (GND), Simple Knowledge Organization System (SKOS), Functional Requirements for Bibliographic Records (FRBR).

IV. CONCLUSIONS

Research Data Management is a topic that has been discussed for many years and is gaining in importance. The basic premise for appropriate data management is the proper and efficient use of public funds. Considering the changes in scientific communication and the guidelines of research funding agencies, as well as legislative changes, Polish universities will have to create data repositories or use existing platforms for sharing research data. Both the repositories analyzed and the remaining repositories must be carefully analyzed according to specific qualitative criteria, defined at a later stage of the research. The conducted study showed that, among similar initiatives, there are significant differences in the functionality and standards used.

Further research should also cover other issues related to the quality of Polish repositories. These include indexing the repository content in dataset indexing databases, such as Google Dataset Search, DataCite, and Data Citation Index (Web of Science). Another criterion against which repositories will be assessed will be their certification (e.g., Core Trust Seal) and presence on the lists of trusted repositories published by scientific journals and publishers (e.g., Nature Scientific Data Recommended Repositories [22]). Further research should also show in which other data repositories researchers from Poland share their data. Preliminary analyses indicate that these may be repositories with an international scope, such as Zenodo or Figshare, domain repositories and repositories of scientific journal publishers.

REFERENCES

[1] M. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". Sci Data 3, 160018 (2016), doi: 10.1038/sdata.2016.18.

[2] European Union, "Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast)", Official Journal of the European Union, L 172/56, 26.6.2019. [Online]. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L1024&from=EN, 2021.02.24.

[3] The Act of February 25, 2016 on the re-use of public sector information, Dziennik Ustaw 2016 poz. 352. Available from: http://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU20160000352/U/D20160352Lj.pdf, 2021.02.24.

[4] D. Reynolds, „Linked (Open) Data" in: B.C. Dalsgaard, Data reuse and how metadata can stimulate reuse. A workshop held on Monday 5th of December 2011 at the 7th IDCC conference in Bristol. [Online]. Available from: https://libereurope.eu/wp-content/uploads/2020/11/WGSC_20111205.pdf.

[5] Ministerstwo Cyfryzacji, „Draft act on open data and re-use of public sector information, RCL, 24.08.2020.[Online]. Available from: https://legislacja.rcl.gov.pl/projekt/12337400, 2021.02.24.

[6] S. Zdziebłowski, „On Tuesday, the beginning of the 2019/2020 academic year", Nauka w Polsce, 01.10.2019. [Online].Available from: https://naukawpolsce.pap.pl/aktualnosci/news%2C78787%2Cwe-wtorek-poczatek-roku-akademickiego-20192020.html, 2021.02.24.

[7] Statistic Poland, Higher education and its finances in 2019, Warszawa, Gdańsk 2020. Available from: https://stat.gov.pl/obszary-

tematyczne/edukacja/edukacja/szkolnictwo-wyzsze-i-jego-finanse-w-2019-roku,2,16.html. 2021.02.24.

[8] OpenDOAR, “Browse by country and region: Poland”. [Online]. Available from: https://v2.sherpa.ac.uk/view/repository_by_country/Poland.html, 2021.02.24.

[9] Re3data.org. [Online]. Available from: https://www.re3data.org/search?query=&countries[]=POL, 2021.02.24.

[10] The results of the survey were presented during study day *The semantic web and cultural heritage from data convergence to knowledge crossing*, 3 February 2021, GERiiCO & Universitté de Lille [online]. D. Siwecka, “Linked Open Data in the Polish (librarian) landscape”.

[11] AZON 2.0, https://zasobynauki.pl/, 2021.02.24.

[12] Bridge of Data, https://mostwiedzy.pl/en/, 2021.02.24.

[13] Repository of Jagiellonian University, https://ruj.uj.edu.pl/xmlui/, 2021.02.24.

[14] Polish Platform of Medical Research, https://ppm.edu.pl/index.seam?lang=en&cid=44836, 2021.02.24.

[15] PPM, Research data in a nutshell. Guide. [Online]. Available from: https://biblioteka.wum.edu.pl/sites/biblioteka.wum.edu.pl/files/poradnik_dane_badawcze_ppm_v7-1.pdf, 2021.02.24.

[16] M. Szufita-Żurawska, Data Management Plan, 2020. [Online]. Available from: https://cdn.mostwiedzy.pl/c0/a3/7d/50/0_202002141036301505451_FME/plan-zarzadzania-danymi-2020.pdf, 2021.02.24.

[17] RUJ, Research Data. [Online]. Available form: https://cawp.uj.edu.pl/documents/102715934/141758336/NCN_zarzadzanie-danymi.pdf/266e134e-418e-4ca0-842f-1ebb47227235, 2021.02.24.

[18] H. Koers, P. Herterich, R. Hooft, M. Gruenpeter, and T. Aalto, “M2.10 Report on basic framework on FAIRness of services” 30th November 2020, Zenodo, doi: 10.5281/zenodo.4292598.

[19] Bridge of Data, 5-star Open Data. [Online]. Available form: https://mostwiedzy.pl/pl/open-data?pageId=1_50711040454643_SPA, 2021.02.24.

[20] PPM: 5* Open Data. [Online]. Available from: https://ppm.edu.pl/fiveStarData.seam, 2021.02.24.

[21] PPM, API. [Online]. Available from: https://ppm.edu.pl/api.seam, 2021.02.24.

[22] “Recommended data repositories”, Scientific Data. [Online]. Available from: https://www.nature.com/sdata/policies/repositories, 2021.02.24.

 ISBN: 978-1-61208-842-6