Folia Forestalia Polonica, Series A – Forestry, 2018, Vol. 60 (3), 192–198

METHODOLOGICAL ARTICLE

DOI: 10.2478/ffp-2018-0019

Open data in scientific communication

Dorota Grygoruk

Forest Research Institute, Department of Forest Ecology, Sękocin Stary, Braci Leśnej 3, 05-090 Raszyn, Poland, e-mail: farfald@ibles.waw.pl

Abstract

The development of information technology makes it possible to collect and analyse more and more data resources. The results of research, regardless of the discipline, constitute one of main sources of data. Currently, the research results are increasingly being published in the Open Access model. The Open Access concept has been accepted and recommended worldwide by many institutions financing and implementing research. Initially, the idea of openness concerned only the results of research and scientific publications; at present, more attention is paid to the problem of sharing scientific data, including raw data. Proceedings towards open data are intricate, as data specificity requires the development of an appropriate legal, technical and organizational model, followed by the implementation of data management policies at both the institutional and national levels.

The aim of this publication was to present the development of the open data concept in the context of open access idea and problems related to defining data in the process of data sharing and data management.

Key words

open access, open data, research data, data management

Introduction

Modern information technology allows collection and analysis more and more data resources. At the beginning of our century, it was estimated that new stored information grew about 30% a year between 1999 and 2002 (Berkeley Report 2003). The scientific studies are one of the main sources of data, and their results are increasingly available in the form of scientific publications in the open access model. The beginning of Open Access (OA) dates back to the 1960s, when the first centres of scientific information were established in the USA. Publishing in prestigious scientific journals has become the guarantee of the professional advancement of authors and promotion of research centres (Hofmokl et al. 2009). In opinion Nielsen (2008), the growth of the scientific journal system has created a body of shared knowledge and a collective long-term memory that is the basis for progress in science. New possibilities of dissemination of research findings emerged along with the development of the Internet and digital technology. The first journals exclusively published on the Internet were launched in the late 1980s (Hofmokl et al. 2009). The first open scientific repository in the fields of physics, astronomy, mathematics and computer science was established in 1991, actually (as on 30/08/2018) contains 1,433,214 documents (https://arxiv.org/). Currently, no library in the world subscribes to all the printed scien-

Received 2 August 2018 / Accepted 20 August 2018

tific journals, as their prices and the number of studies published grow faster than the libraries' budgets. The essence of Open Access is both free of charge access to research results and the possibility of their re-use for scientific purposes – by reading, saving to a computer disk, copying, printing, looking up, linking as well as correct quoting through verifying the work authorship. The OA model ensures the process of publication reviewing, does not violate copyrights and adheres to anti-plagiarism regulations (Suber 2014). The goals of Open Access have been defined in three declarations, that is, the Budapest Open Access Initiative (2002), the Bethesda Statement on Open Access Publishing (2003) and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003). According to the Registry of Open Access Repository Mandates and Policies (ROARMAP), Open Access policies were adopted in a total of 83 funders, 56 funders and research organisations, 11 research organisations, 716 universities and research institutions, 75 sub-units of research organisations (as of March 2018). From Poland, 6 scientific units have been entered in the register: the Adam Mickiewicz University, Institute of Nuclear Physics Polish Academy of Sciences, the Interdisciplinary Centre of Mathematical and Computational Modelling (ICM), the University of Warsaw, the Medical University of Lodz, the Nofer Institute of Occupational Medicine in Lodz and the Polish Academy of Science Institute of Biochemistry and Biophysics (http://roarmap.eprints.org).

In Poland, the Ministry of Science and Higher Education (MNiSW) is responsible for science policy. In 2004, Poland signed the OECD Declaration on Access to Research Data From Public Funding. According to the provisions of the Declaration, open access to research data is a prerequisite for innovation, improvement of scientific staff qualifications, as well as international scientific and technological cooperation. The document does not provide guidelines for the implementation of the open science model (OEDC 2004). Since 2010, MNiSW has financed the Springer Open Choice/Open Access Program, under which the employees and students affiliated with all the Polish academic, educational and scientific institutions can publish their research in the scientific journals published in open access by Springer (SpringerOpen 2010). In 2011, an expert opinion on the implementation and promotion of open access to scientific and educational contents was commissioned by MNiSW. The results of the analysis, carried out with reference to 12 countries and selected international organizations, were used to develop a model for the implementation of the OA model in the Poland's science system. The most important recommendations in the Report regarded incorporation of the open access policy in the parametric evaluation of research centres and introduction of the OA mandate in the Polish institutions financing the research. At the same time, the need for OA trainings and modernization of IT infrastructure was emphasized (Niezgódka 2011).

A range of international organizations, and the European Union (EU) as well, have a great influence on shaping the science system in Poland. For example, the EU's documents, such as 2012/417/EU: the Commission Recommendation of 17 July 2012 on access to and preservation of scientific information and Regulation (EU) No1290/2013 of the European Parliament and of the Council of 11 December 2013 laying down the rules for participation and dissemination in 'Horizon 2020 - the Framework Programme for Research and Innovation (2014–2020)' and repealing Regulation (EC) No 1906/2006, recommend open access to research results financed by the EU; under the Horizon 2020 projects, open access to research results is obligatory. In 2015, the Minister of Science and Higher Education adopted the 'Directions for the development of open access to publications and the results of scientific research in Poland.' The document emphasizes that dissemination of open access to research results is a global trend, largely related to the development of information and communication technologies (MNiSW 2015). In 2018, the 'Report on the implementation of the policy of open access to scientific publications in 2015–2017' was published, which discusses the basic problems of the process of introducing open access to scientific content in Poland and provides recommendations for future activities. According to the authors of the Report, only 20 research centres and universities in Poland have the institutional policy of OA, and only 18% of all scientific publications are published in the OA system. In Poland, there are no systemic solutions and adequate OA infrastructure. In addition, OA activities are not rewarded in the evaluation of scientific units or in the assessment of the academic staff (MNiSW 2018).

Open Data

The development of the idea of open access to data is closely related to the activity of CODATA – the Data Committee of the International Council for Science (ICSU), which was established in 1966. The mission of the organization is to promote global cooperation in order to improve the availability and usability of data for all areas of research and to support international science for the benefit of society. CODATA performs its tasks both on an international scale and in the scale of individual member states, which also include Poland. CODATA also runs publishing activities, collaborates in the organization of large data conferences such as SciDataCon and International Data Week (http://www.codata.org). The Data Science Journal as a peer-reviewed, open electronic journal publishes articles on the management, dissemination, use and reuse of research data and databases in all areas of research. The scope of the journal includes descriptions of data systems, their implementation and publication, applications, infrastructure, software, legal issues, reproducibility and transparency, accessibility and usability of complex data sets, with particular emphasis on principles, policies and practices for open data (http://datascience.codata.org).

Open access to data increases transparency of the research process as well as promotes scientific cooperation and the implementation of interdisciplinary scientific research. The development of some scientific disciplines (e.g., bioinformatics) is based on access to data, while other fields (e.g. astronomy, physics, climatology) are strongly associated with collecting and sharing data at a global level. The growing interest in the availability of research data is to a large extent related to the rapid development of digital technologies. Modern IT solutions enable generating, storing, processing and transmitting ever-larger data sets (Hofmokl et al. 2009).

Activities for Open Access and Open Data are complementary; however, data specificity requires the development of a legal, technical and organizational model and the implementation of appropriate data management procedures. The first key problem in the field of access to data is the lack of an agreed definition of 'research data' (Szprot et al. 2014; Strzelczyk 2017). The diversity and specifics of scientific fields cause that research data is defined in various ways, for example:

- data as registered factual materials, necessary to evaluate the results of scientific research and widely recognized by the scientific community
- data as information, in particular – collected facts and figures that can be used for research and be treated as a basis for further conclusions, discussions or calculations
- data as records of facts (expressed as numbers, text, graphics or sounds) that are the result of study (e.g., observations, measurements, experiences, experiments, etc.), used as a base for scientific conclusions
- data as raw data, which was obtained directly as a result of the use of a research tool (e.g., computer program, measuring equipment, survey, questionnaire) – organized but not processed, for example, by means of statistical analyses
- data as descriptions and information on data origin – metadata

Similar problem arise when defining 'open research data'. According to James (2013), open data can be freely used, distributed by anyone and anywhere for any purpose. The authors of other definitions introduce certain limitations by, for example, licenses specifying the conditions for data sharing and the information on data source. In contrast to the publications made available in the Open Access model, the essential feature of open data is the possibility of its reuse in new analyses and re-dissemination. For this reason, data may be subject to exclusive rights, that is, copyrights, database rights and regulations on the subject of access to public data or the protection of personal data (Szprot et al. 2014).

Consistent with the European Commission (EC), open access to research data from the projects financed from public funds should be a standard practice (Amsterdam Call for Action on Open Science 2016). The EC recommends the FAIR (Findable, Accessible, Interoperable, Re-usable) Principles for research data stewardship to make data findable, accessible, interoperable and reusable, therefore – easy to find in open repositories or on the Internet, for example, by linking to a scientific publication, available to one and all (also on license rights) and stored in standard formats that are easy to open, read and reuse (Guidelines on FAIR Data Management in Horizon 2020). In 2013, the European Commission, the United States National Science Foundation, the National Institute of Standards and Technology and

the Australian Government's Department of Innovation launched the Research Data Alliance (RDA) as a community-driven organization. The goal of this organization was to create the social and technical infrastructure to enable open sharing of data. As of September 2018, the RDA has over 7251 individual members from 137 countries (representatives of the Interdisciplinary Centre of Mathematical and Computational Modelling [ICM], the University of Warsaw are members of the RDA and represent the Polish scientific community). The Research Data Alliance enables data to be shared without barriers through Working Groups and Interest Groups (a total of 93 working groups), formed of experts from all around the world – from academia, industry and government (https://www.rd-alliance.org/).

The diversity of data collected in research processes, recording formats and storage standards require implementing system solutions in the area of open access policy at the institutional, national and international levels. Providing open access to research data will enable the reuse of data for analyses, surveys and tests, as well as publication of new results (Bednarek-Michalska 2012; Strzelczyk 2017; MNiSW 2018).

Research Data Management

Each scientific process as the data life cycle includes the stages of collecting, processing, analysing, using and data sharing. The life cycle of scientific data can be extended ('given a second life') by appropriate management procedures, which enable data re-sharing and using in other scientific projects. Specific activities in the field data management are required by some of the scientific journals (e.g., Nature, PLoS, etc.) and are also included in the grant agreements, for example those signed with the EC or Poland's government agency the National Science Centre (NCN) (Sommer 2015). Along with the EC recommendations, research data management should be carried out both during and after the implementation of the scientific project (the Guidelines on FAIR Data Management in Horizon 2020). These activities include: defining data, selecting formats, describing metadata, determining current storage location, which is followed by selecting and preparing data for long-term storage as well as choosing measures to secure data and to ensure data sharing. The selection of data for archiving should be carried out based on scientific and historical criteria, as well as the assessment of data documentation quality and the possibilities of future use of data and replication. An important issue is also the regulation of legal status at the stage of data sharing; for instance, data can be made available without a license on any terms of use, or with a license FAIR Data Management Creative Commons, or with a statement of surrender (Sommer 2015).

According to Görögh (2014), regardless of the methodology used, the long-term protection of scientific data including raw data at an institutional level has numerous advantages. It contributes to the comprehensive gathering of knowledge, increases the transparency of research, builds the prestige of a given scientific institution, as well as enhances the development of international cooperation and encourages participation in research consortia. Furthermore, open access to raw data at an institutional level ensures legal data protection, in particular, protection against the risk of copyright.

In the last decade, the role of data management (RDM) in scientific communication has grown. The research data management policy was implemented at the University of Oxford in 2012. The main policy objectives refer to the evaluation of data collected under the University projects, the determination of the minimum period of data storage after the publication of research results and the scope of responsibility of scientists and the university. It is the responsibility of the researchers to develop and document procedures as regards collecting, retaining, using, reusing and sharing scientific data. The University provides access to appropriate services and devices, including support by IT staff and organizes trainings in the research methods and data management (http://researchdata.ox.ac.uk/). The policy implementation was preceded by a survey related to the research data and sharing principles. About 300 academic employees responded to the survey questions. The answers confirmed diversity of data (text, numeric, spatial, statistical, multimedia, audio, bibliographic) collected during the implementation of scientific projects. About 75% of the respondents said that sharing data is not necessary, but at the same time, the majority of respondents acknowledged that data management is necessary and important for research process as well as considered the access to the scientific data from completed projects as an inspiration for new research ideas (Wilson 2015).

The data management policy has also been implemented at the University of Cambridge. Here, regardless of the funder, each project starts with the preparation of a Data Management Plan. The Plan includes choosing the data format, software type and the method for storing data. There are recommended formats less vulnerable to obsolescence and easy to describe by metadata, which facilitates data interpretation and reuse in the future (http://www.data.cam.ac.uk/).

The open access and data management policy has also been gradually implemented in the Polish research centres. The Interdisciplinary Centre for Mathematical and Computer Modelling at the University of Warsaw (ICM UW) has established the first repository of accessible publications by Polish scientists. Since 2011, the Centre of Open Science (CeON) has collected and made available to anyone, free of charge, scientific articles, books, post-conference materials, scientific monographs and doctoral dissertations (in compliance with the CC-BY or CC-BY-SA copyright license). The data repository functions compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH); therefore, the publications are easily accessible via the websites providing information on digital scientific resources (Grodecka 2013).

Due to the intensive data increase in many fields of science, the demand for technical infrastructure with high capacity, durability and performance has also increased. In Poland, as part of OCEAN projects (Open Data Centre and Analysis) and RepOD (Open Data Repository), a modern infrastructure for storing and sharing data was implemented in order to manage research data responsibly and provide services to other scientific institutions and public sphere institutions (http://ocean.icm.edu.pl/, https://repod.pon.edu.pl/pl/group/icm-uw).

According to Kędzierska et al. (2014), although Polish scientific community declares the need to sharing raw data together with scientific publications and appreciates data re-processing, it still however has more data stored in personal computers, and not in the institutional data repositories. The surveys with regard to the rules of sharing raw data and research results have been carried out in more than 200 research centres in Poland. Respondents indicated direct measurements and experimentation as the main sources of raw data and personal computers as the main tool for data storage. The idea of establishing a central data repository in research centres was approved by about 70% of the respondents, who, at the same time, confirmed greater acceptance of open access to scientific publications than to raw data (Stępniak 2014).

In Poland, a survey was also carried out regarding the collection, storage and sharing of scientific data at the Forestry Research Institute in 2015. 64% of the Institute's academic staff took part in the research. The results of the survey confirm the diversity of data collected in the research on forest ecosystems. Most of the data is generated during field measurements, where modern measuring equipment is increasingly used, for example, terrestrial laser scanners, telemetry devices. The size of database resources of the Institute has clearly increased in the recent years, which is the result of increased processing and analysis of spatial data. The respondents most often indicated personal computers as a tool for archiving their databases (raw and processed data), thus – a means not assuring storage quality and security. It is worth noting that modern IT tools are available in IBL because in the period of 2010–2014, the Institute's IT system was modernized as a part of the infrastructure project (POIG.02.03.00-00-052/10). The scope of the project included, among others, new technological solutions in the field of data archiving. Most of the survey respondents (82%) considered it useful to use archival data at the stage of drawing scientific conclusions and planning new research. The Open Access concept was accepted by 74% of respondents –above all, in the context of access to scientific publications. Open access to databases raises many controversies and fears among the scientific staff of the institute (e.g., in the context of copyright) (Grygoruk 2017).

The presented survey results (Stępniak 2014; Wilson 2015; Grygoruk 2017) characterize various scientific environments both in Poland and other countries. The results obtained also show similarities in the work of the researcher, irrespective of the field of science. The modern measurement and analytical technology available today allows you to generate and process a variety of data resources with growing volume. However, the routine and habit of the scientific community are still a mental barrier to the dissemination of new forms of sharing knowledge in many scientific institutions, even though the concept of Open Access in science is no longer a niche initiative (Nielsen 2008).

Conclusion

Contemporary science is closely related to the development of information technology. In the Internet age, open access to scientific publications as well as research data influences the development of communication/scientific cooperation. Until recently, the achievements of scientific centres were mainly evaluated on the basis of completed projects, scientific publications and professional achievements of the scientific staff. Today, the evaluation criteria are organizational and technological solutions that enable analysis. In this situation, it becomes necessary to implement data management policy by research institutions. Securing data against loss and guaranteeing access to them for future generations is a challenge for science centres, not only in Poland.

References

Amsterdam Call for Action on Open Science. Conference Amsterdam 4–5 April 2016. Available at http://www.openaccess.nl/en/events/amsterdam-call-for-action-on-open-science.

Bednarek-Michalska, B. 2012. Repozytoria surowych danych – dlaczego biblioteki powinny je znać? *Biuletyn EBIB*, 135, 1–8. Available at http://repozytorium.umk.pl/bitstream/handle/item/207/135_michalska_.pdf.

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. 2003. Available at http://oad.simmons.edu/oadwiki/Declarations_in_support_of_OA.

Berkeley Report. How Much Information? 2003. Available at http://www.sims.berkeley.edu/research/projects/how-much-info-2003/.

Bethesda Statement on Open Access Publishing. 2003. Available at http://oad.simmons.edu/oadwiki/Declarations_in_support_of_OA.

Budapest Open Access Initiative. 2002. Available at http://oad.simmons.edu/oadwiki/Declarations_in_support_of_OA.

Commission Recommendation of 17 July 2012 on access to and preservation of scientific information (2012/417/UE). Available at https://eur-lex.europa.eu/legal-content/PL/TXT/?uri=CELEX%3A32012H0417.

Görögh, E. 2014. Conference on Grey Literature and Repositories. Proceedings 2014: The Value of Grey Literature in Repositories. National Library of Technology, Prague, 47–52. Available at http://invenio.nusl.cz/record/180589/files/idr-879_1.pdf.

Grodecka, K. 2013. Udane projekty open access w Polsce. EBIB, Toruń. Available at https://www.ifj.edu.pl/library/open-access/materials/Grodecka.pdf.

Grygoruk, D. 2017. Open access to research data on forest ecosystems in Poland. *Task Quartely*, 21 (4), 415–421. DOI: https://doi.org/10.17466/tq2017/21.4/w.

Guidelines on FAIR Data Management in Horizon 2020. Available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Hoffman-Sommer, M. 2015. Workshop "Research data management", Warsaw 9 and 11 December 2015. Platforma Otwartej Nauki, ICM UW, Warsaw. Available at https://www.slideshare.net/OpenSciencePlatform/zarzdzanie-danymi-badawczymi.

Hofmokl, J., Tarkowski, A., Bednarek-Michalska, B., Siewicz, K., Szprot, J. 2009. Przewodnik po otwartej nauce. ICM UW, Warsaw. Available at https://depot.ceon.pl/bitstream/handle/123456789/65/przewodnik-po-otwartej-nauce.pdf.

James, L. 2013. Defining Open Data. Open Knowledge International Blog. Available at http://blog.okfn.org/2013/10/03/defining-open-data/.

Kędzierska, E., Kavalchuk, N., Stepniak, J. 2014. Conference on Grey Literature and Repositories. Proceedings 2014: The Value of Grey Literature in Repositories. The report from the Survey of Polish Scientific and Research-Development Units. National Library of Technology, Prague, 56–60. Available at http://invenio.nusl.cz/record/180589/files/idr-879_1.pdf.

Kierunki rozwoju otwartego dostępu do publikacji i wyników badań naukowych w Polsce. 2015. MNiSW, Warsaw. Available at https://www.nauka.gov.pl/g2/oryginal/2018_04/f168c48611e66a5507ca5391e4b7e8e1.pdf.

Nielsen, M. 2008. The future of science. Available at http://michaelnielsen.org/blog/the-future-of-science-2/.

Niezgódka, M. et al. 2011 Wdrożenie i promocja otwartego dostępu do treści naukowych i edu-

kacyjnych. Praktyki światowe a specyfika polska. Przewidywane koszty, narzędzia, zalety i wady. ICM UW, Warsaw. Report for MNiSW. Available at https://depot.ceon.pl/bitstream/handle/123456789/1545/20120208_EKSPERTYZA_OA ICM.pdf?sequence=1&isAllowed=y.

OECD. 2014. Declaration on Access to Research Data From Public Funding. Available at https://legalinstruments.oecd.org/public/doc/157/157.en.pdf.

Raport nt. realizacji polityki otwartego dostępu do publikacji naukowych w latach 2015–2017. 2018. MNiSW, Warsaw. Available at https://www.nauka.gov.pl/g2/oryginal/2018_04/7ed78f459cb760b267b19f8f38f8bb22.pdf.

Regulation (EU) No 1290/2013 of the European Parliament and of the Council of 11 December 2013 laying down the rules for participation and dissemination in "Horizon 2020 – the Framework Programme for Research and Innovation (2014–2020)" and repealing Regulation (EC) No 1906/2006 Text with EEA relevance. Available at http://ec.europa.eu/research/participants/data/ref/h2020/legal_basis/rules_participation/h2020-rules-participation_pl.pdf.

SpringerOpen. 2010. Available at https://www.springer.com/gp/open-access/springer-open-choice/springer-compact/springer-open-choice-for-polish-institutions/11027898.

Stępniak, J. 2014. Otwarte surowe dane i wyniki badań. Raport z badań w krajach grupy wyszehradzkiej. Seminar „Open science and open model of scientific communication", 20.10.2014. Warsaw University of Technology, Warsaw. Available at htpp://repo.pw.edu.pl.

Strzelczyk, E. 2017. Otwarte dane badawcze – kolejny krok do otwierania nauki. In: Bibliographic databases: perspectives and development problems (eds.: I. Sójkowska, L. Derfert Wolf). III Scientific Conference Consortium BazTech. Cracow, June 26–27 2017. EBIB. 25. Available at http://open.ebib.pl/ojs/index.php/Mat_konf/article/view/599.

Suber P. 2014. Otwarty Dostęp. ICM UW. Available at http://www.bm.cm.uj.edu.pl/documents/21651741/56d5843c-2e99-43a2-a4bf-f34324a1b048.

Szprot, J., Leśniak, A., Morys-Twarowski, M., Sieniewicz, K., Starczewski, M., Stępniewska-Ustasiak, L. 2014. Open Science in Poland 2014. A Diagnosis. ICM, Warsaw. Available at http://pon.edu.pl/index.php/nasze-publikacje?pubid=16.

Wilson, J.A.J. 2015. Good Practice in enabling the reuse of Research Data: The University of Oxford. Conference „Open Access to Research Data as a Driver for Open Science", 15–16.01.2015, Athens. Available at http://helios-eie.ekt.gr/EIE/bitstream/10442/14579/1/Wilson-RECODE.pdf.