FIRST POLISH DATA STEWARD SCHOOL

MARIA M. PAWŁOWSKA
AND MARTA E. WACHOWICZ

Visnea sp. z o.o.
Ząbkowska 27/31,
03-736 Warszawa, Poland

(received: 20.12.2021; revised: 21.12.2021;
accepted: 29.12.2021; published online: 30.12.2021)

Abstract: The paper describes the genesis and the teaching process of the Data Steward School, Edition 2020, the first Polish school for data stewards. The initiative was implemented by Visnea sp. z o.o. ("Visnea") in cooperation with GO-FAIR in the period from September 2020 to April 2021. The participants of the Training Programme, future data stewards, gained knowledge of the role of correct data management in achieving institutional objectives. The need to protect the legal and financial interests arising from the possession and archiving of data and data management plan design guidelines, along with the methodology for data collection, metadata, the existence of data repositories, data security and the means of data sharing and storage were also presented. The paper presents the evaluation of the Training Programme and discusses proposals related to the role and importance of the new profession, i.e. the data steward in a scientific institution.

Keywords: ethics in scientific research, open science, Open Access, researcher, research data management, scientific disciplines

DOI: https://doi.org/10.34808/tq2021/25.4/t

1. Genesis of the Data Steward School Training Programme - Edition 2020 in Poland

The original idea to prepare and carry out the first Data Steward School in Poland – Edition 2020 resulted from the following circumstances: analysis of global trends in the functioning of science and the role of scientists in the socio-economic life and the diagnosed needs of the Polish scientific community. The teaching programme reflected both these considerations in order to take into account the consequences of changes in the European scientific policy and trends in the science development in Europe, as well as the domestic needs of scientific institutions.

The management of scientific data is a new challenge for the scientific community, as it is necessary to understand the role of correct data management in achieving the scientific objectives of both individual researchers and entire scientific institutions. Ensuring the sustainability, security and widest possible access to publicly funded research results becomes a priority in the management of science throughout Europe. The management of scientific data raises a number of complex issues, such as the data and metadata collection methodology and data security. Issues relating to legal responsibility and liability for the entire process of production, sharing, collection and archiving of data, as well as ethical problems related thereto, are also crucial. Visnea has experience in the RD sector, and in the monitoring of large and complex scientific projects. This experience has led to the conviction that Polish science needs, above all, to improve the management processes, including the management of scientific data. Laboratories of Polish scientific institutions are often fitted with excellent equipment and tools, and internationalisation of research and teaching processes is increasing. Unfortunately, there is generally little competence in the scientific community with regard to managing research and its results. Visnea's experts have ambitions to facilitate the management of science in Poland and, above all, the scientific data aspect.

At the same time, global and European trends in science management indicate the need for new initiatives to be implemented in connection with a change in the paradigm of scientific publications. A lot of papers have already been published on the potential benefits and many challenges of open science and the need to share and document data. Open access and open science are pathways to more effective cooperation in science and faster information processing [3]. The paradigm of open science radically changes the publishing landscape, as new requirements for data and source codes and an appropriate description of metadata are emerging. Many research funding agencies currently require (or at least strongly encourage) publication in open access journals [6].

Universalism and the possibility of verifying hypotheses and sources is a fundamental principle of science; however, the only results that can be verifiable, questioned or tested and reproduced by others qualify as scientific. Science can therefore function properly only if the results of research are made available to the scientific community for reuse or verification by other researchers. In addition, new studies are based on established results from previous studies. Limited access to data hinders scientific activity on its own ground. It is believed that cashing in on access to new and existing research results is deeply at odds with the ethos of science. There is no longer any justification for this, and therefore some call for the end of a subscription model for scientific publications, and the need for scientific publishers to provide services to help researchers disseminate their results [9]. Therefore, one of the most important challenges currently faced by the Polish science is to adapt to the principles of the virtual and digital world and to develop competences of the managerial staff in data management. Data Management Plans carry the necessary components of an application for research work funding, both at national and European levels.

One of the objectives of the data steward school was also to establish the best practices related to the professional role of the data steward in Poland, taking into account the specificities of scientific units [1,4]. Therefore, the Data Steward School Training Programme was planned as a comprehensive, 7-month programme implemented by Visnea in cooperation with the GO–FAIR Initiative [12].

2. The Training Programme

2.1. Substantive aspect of the Training Programme

Participation in the Data Steward School allowed participants to acquire new professional qualifications and unique competences in the field of scientific data management. The aim was to provide comprehensive education in Data Stewardship, as well as support, during the many months of learning issues related to scientific data management.

The programme consisted of the Foundation Level module – i.e. a week of intensive learning, followed by three specialist training sessions to deepen the knowledge about specific issues as well as mentorship. The entire substantive concept of the Training Programme was developed with the participation of the GO-FAIR Initiative. GO-FAIR is a bottom-up and self-governing initiative aimed at implementing the FAIR rules on data, thanks to which data is to be Findable, Accessible, Interoperable, Reusable. GO-FAIR creates an open ecosystem for persons, institutions and organisations cooperating within the Implementation Networks, enabling the implementation of established scientific data management standards. GO-FAIR implements a bottom-up open implementation strategy for the European Open Science Cloud (EOSC) as part of the wider global Internet of FAIR Data Services (IFDS). This approach is largely based on the recommendations of the Commission's High Level Expert Group on EOSC [12]. Visnea is an official representative of the GO–FAIR Initiative in Poland and is responsible for implementing data management standards in accordance with the guidelines of this organisation, and Visnea experts are involved in an international discussion on setting standards and challenges related to the implementation of good practices.

The Data Steward School was run in the train the trainers formula, which means that graduates of the programme have become Data Stewardship experts, and obtained the appropriate competencies to train further persons in their parent units. Such a model, developed in cooperation with the GO FAIR Initiative, was designed to enable the most effective implementation of international best practices in Poland.

The training offer was addressed to university administration staff and staff members of scientific institutes handling international scientific projects, librarians and persons responsible for selecting management control processes at universities or scientific institutions, as well as members of research groups responsible for the management of research results.

Participants in the training programme acquired knowledge of the role of correct data management in achieving the scientific objectives of individual researchers and of the entire scientific institutions. The issue of the need to protect legal and financial interests arising from the possession and archiving of data was also presented. After the completion of the Training Programme, participants will be able to prepare a data management plan, including a methodology for data collection, metadata collection, existence of data repositories, data security and ways of sharing and storing data.

The teaching process course was as follows: each participant would gain qualifications in the basics of data management during the intensive training week and then acquire expertise during two selected specialist training courses on database and data security issues, intellectual property issues or strategic management. The trainers and lecturers from the Data Steward School - Edition 2020 were European experts in scientific data management, as well as specialists in Big Data, databases, security and scientific data. The classes were in the form of lectures, practical classes, case studies and mentoring conducted by GO-FAIR experts. The total number of theoretical training hours was 65 and practical classes 30.

2.2. Foundation Level

All Programme participants took part in the Foundation Level training module (approx. 45 hours of training) and learned issues related to the processing, collection and archiving of scientific data. The key issues covered include:

- introduction to scientific data management topics (justification of the need for data management; European and Polish documents stipulating scientific data management requirements);
- global trends in science management – necessity of reproducibility, reuse, sharing of data resources;
- the role of the data steward in a scientific institute;
- Open access and Creative Commons;
- replicability/scientific verification – Metadata;
- data protection (physical protection – repositories, documentation, descriptions; databases, collections, quotes);
- financial aspects of data management;
- Data Management Plan – an important document in research - how to prepare and amend it (discussion of key elements of DMP, analysis of data sharing standards and issues; long-term strategy for data storage and protection; need to amend and update a DMP).

The Data Management Body of Knowledge defines data management as [2]:

”The development, implementation and supervision of plans, policies, programmes and practices which provide, control, protect and enhance the value of data and information resources throughout their life cycle”. On the other hand, the data management specialist is: ”Any person who deals with any aspect of data management... [and] has a number of roles, from highly technical (e.g. database administrators, network administrators, programmers) to strategic (e.g. data stewards or main data inspectors)".

The Data Management Body of Knowledge distinguishes 11 substantive areas related to data management [2]. The training programme included 6 out of 11 of the following substantive areas in the Foundation Level basic module:

- planning, supervision and control of data management and the use of data and data resources;
- data architecture – general data and data resources structure as an integral part of the system architecture of the given organisation;
- modelling – data analysis, testing and retention;
- data storage and operations – implementation and management of structured physical data resources; creating documents – storing, protecting, indexing and enabling access to data found in unstructured sources and making such data available for integration and interoperability with structural data (databases);
- analytical data processing management and enabling access to data supporting decision making for reporting and analyses;

The remaining 5 areas have been included in specialist trainings. These are issues covering:

- data security – ensuring protection, confidentiality and adequate access to data;
- data integration and interoperability – acquisition, extraction, transformation, replication, operational support;
- reference and basic data – managing common data in order to reduce excessiveness and ensure better data quality through standardisation;
- metadata – collection, categorisation, maintenance, integration, control and delivery of metadata;
- data quality – defining, monitoring, maintaining data integrity and improving data quality.

FAIR DATA aspects and challenges of open science and open access were also discussed [5,11]. Challenges related to the new data steward profession were presented. Separate training hours were allocated to discussing scientific data management in the context of the broader research unit in order to draw attention to scientific data management as part of the research process and the importance of scientific data management at the level of the research unit as a whole.

The Data Management Plan was of particular importance in the teaching process. The European Commission requires the preparation of a Data Management Plan for projects funded under the Framework Programme for Research and Innovation Horizon Europe. In Polish National Science Centre (NCN) competitions, applicants are already obliged to present a plan for the management of research data which will be created as part of the project implementation.

Pursuant to the Ordinance of the Director of the NCN No. 38/2020 of 27 May 2020, in all newly announced NCN competitions an open science policy will be implemented, which will require the management of research data. One of the objectives of the training programme was to thoroughly familiarise participants with data management plans in order to be able to transfer the acquired knowledge in a cascade-like manner in the parent scientific units. The practical classes in this regard referred to the institutional strategy and model plan developed by Science Europe [10] (in the Polish version, the template was made available by Pawłowska, Wachowicz, 2020 [7]).

A significant number of hours were devoted to the presentation of international data management achievements. GO-FAIR experts shared reflections on problems in implementing data management strategies and ensuring data quality. It was particularly important that data stewards from all over the world (from Europe, North America, Asia and Africa) highlighted the global context and the global need to accurately collect and share scientific data [12].

2.3. Specialist training courses

The participants had to choose 2 out of 3 specialist training courses. 85% of the participants decided to participate in all 3 specialist training courses:

a) Data management strategy in a scientific unit; (approx. 15 training hours);
b) Security and storage of research data (approx. 15 training hours);
c) Scientific data and intellectual property (approx. 15 training hours).

Specialist trainings covered Big Data issues – management of large data sets in a scientific institution. Intellectual property and data issues were discussed extensively with regard to cooperation with industry and the role of copyright in data management. Experts in industrial property rights familiarized the participants with the question of the confidentiality of scientific data and means of protecting scientific data in the context of obtaining exclusive rights.

2.4. Mentoring

An additional teaching element was the 7-month mentoring programme conducted by GO-FAIR experts. The mentoring process was aimed at supporting the participants in planning their careers as data stewards, developing professional competencies and preparing for strategic management of scientific data in scientific units. In some cases the mentoring process was carried out in pairs, provided that the participants reaped the benefits of this form of mentoring.

Mentees could choose the mentoring theme conducted by GO-FAIR mentors from among the following issues:

- How to prepare a DMP in the parent scientific unit?
- How to prepare a strategy for the management of scientific data in the parent scientific unit?
- Scientific repositories
- Use of IT tools for data management

- Data storage
- Data collection
- Data sharing
- Data archiving
- Metadata
- Data stewards training
- Cooperation with scientists

Programme Organisers paired mentors and mentees in a manner which ensured the best match in terms of competences and needs specified in questionnaires. Cooperation with the mentor was carried out through at least 5 individual meetings in selected pairs: mentor- mentee(s)t and obtaining a certificate was also conditional on the completion of the mentoring process.

3. Training Programme Evaluation

13 participants took part in the first pilot edition of the school of data stewards. Librarians accounted for 75% of participants. The results of the evaluation indicate that the participants (mainly belonging to a professional group of librarians) have broad knowledge of data collection and sharing, quotation and reference to digital resources, as well as sensitive data issues. Attention is drawn to the fact that the issue of data quality assessment is problematic. Since data quality control is the role of a data steward, in future training programmes particular attention should be paid to the fundamental issue of data quality assessment. The understanding of six key data characteristics should be verified in future data stewards [8], i.e.:

- Accuracy: it is important that the data be collected after a clear definition of the purpose for which it is to be used;
- Validity: Certain guidelines should always be laid down so that the data can be used as closely as possible;
- Reliability: data must be collected from sources that can be trusted to be reliable;
- Timeliness: it is important to collect data as close as possible to the analysis time;
- Rightness: data should relate to the achievement of the measured research objectives;
- Completeness: data should in no case be complete in the light of predefined guidelines.

Evaluation of the course and the thematic scope of the Data Steward School - Edition 2020 and the satisfaction surveys of participants of the Training Programme indicate the need to extend the programme to include elements related to database management and broadly understood formal and legal protection related to the collection, processing and disclosure of data in a public unit. The results of surveys, participants' opinions and test results also indicate the need to broaden the knowledge of future data stewards in principle in three aspects:

- strategic management at the level of the scientific unit - there is a need to regulate the collection and storage of data, in particular the system for sharing and citing research data through clear and transparent procedures or formal and legal regulations. Data stewards should be able to enforce the quality of data and care for the life cycle of data at the level of a division or other organisational unit, which requires the introduction of regulations on the management of scientific data in the existing regulations for the performance of research works or the management of intellectual property of universities or scientific institutes;
- protection of intellectual property – there is a need to broaden knowledge in this regard, in particular of the means of data protection, especially the data collected in the form of databases and knowledge of the possibilities offered by copyright and related rights provisions;
- knowledge of scientific repositories and issues related to the technical aspect of data collection and sharing.

One of the objectives of the data steward school was to establish the best practices related to the professional tasks and role of the data steward in Polish conditions, taking into account the specificities of scientific units. The conclusions of the completed training programme indicate that data stewards should be able to tackle challenges related to the management of scientific data in a specific organizational unit of the research institution, while knowing the specificities of a scientific discipline or a research group. Data stewards, in line with the European trends, should be responsible for:

- data quality and veracity;
- establishing guidelines that data must fulfil in order to be included in a database or publication, etc., and subsequently developing and complying with procedures for data collection or processing;
- care for data documentation and metadata;
- care for the usefulness of the data made available;
- broadly understood data security.

The school curriculum for data stewards should therefore continue to cover these aspects in order to best prepare for the role of the data steward in a scientific unit.

Acknowledgements

The authors express their sincere thanks to Ms Anna Wałek, PhD, Director of the Library of the Gdańsk University of Technology, for a valuable substantive contribution to the preparation of the Data Steward School – Edition 2020, sharing very specialist knowledge and commitment to promoting open access to scientific resources in the academic community.

References

[1] Chatfield T and Selbach R 2011 *Data Management for Data Stewards, Data Management Training Workshop, Bureau of Land Management (BLM)*

[2] Cupoli P 2014 *Data Management Body of Knowledge DAMA-DMBOK2*

[3] *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 PE/28/2019/REV/1*

[4] Earley S 2011 *The DAMA dictionary of data management*, Technics Publications LLC., New Jersey

[5] Hodson J 2018 *FAIR Data Action Plan. Interim recommendations and actions from the European Commission Expert Group on FAIR data* doi: 10.5281/zenodo.1285290

[6] Martyn P C 2021 *Open Science: Open Data, Open Models, ...and Open Publications?*, *Water Resources Research* **57** (4) doi: 10.1029/2020WR029480

[7] Pawłowska M and Wachowicz M 2020 *Wprowadzenie do zarządzania danymi naukowymi*, Difin Publishing House, Warsaw

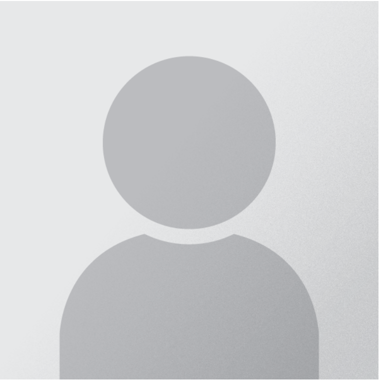
[8] Peng P 2018 *A Conceptual Enterprise Framework for Managing Scientific Data Stewardship*, *Data Science Journal* **17** (15) doi: 10.5334/dsj-2018-015

[9] Schiltz M 2018 *Science without publication paywalls: Coalition S for the realisation of full and immediate Open Access*, *PLoS Medicine* **15** (9) 1002663

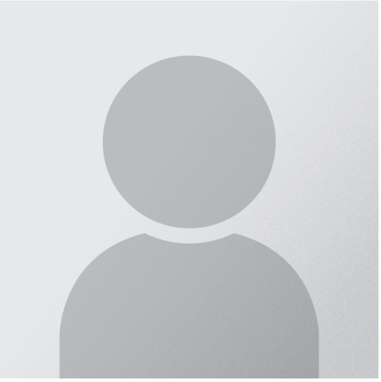
[10] Science Europe 2020 *Implementing Research Data Management Policies Across Europe experiences From Science Europe Member Organisations* doi: 10.5281/zenodo.4915951

[11] Wilkinson M D, Dumontier M and Aalbersberg I J 2016 *The FAIR Guiding Principles for scientific data management and stewardship*, *Scientific Data* **3** 160018 doi: 10.1038/sdata.2016.18

[12] *www.go-fair.org*



Maria M. Pawłowska



Marta E. Wachowicz