ADVANCED COMPUTING SESSION

CONTEMPORARY APPROACH TO DATA MANAGEMENT IN SOCIAL SCIENCES

Jelena Banović[1]*,
Aleksnadra Bradić-Martinović[2]

[1]Institute of Economic Sciences, Belgrade, Serbia
[2]Data Center Serbia for Social Sciences, Belgrade, Serbia

Abstract:
Data management is a set of different activities which ensure the sustainability of scientific work. Data archiving and publishing are the main parts of data life cycle. Good preparation of data before archiving and publishing will allow reuse of the data and provide better scientific engagement. In this paper, authors will be focused on processes that have to be taken before archiving and publishing. Also, some of the main activities will be highlighted, in order to raise awareness about the importance of this topic.

Keywords:
ICT, archiving, publishing, data, data management.

1. INTRODUCTION

The development of ICT significantly contributes to the improvement of scientific research. Quality research implies a long-term process of collecting and processing the data and presenting the results to the academic community. During the research process, however, data management is crucial, and it is a very long and demanding process. Data management involves defined procedures and methods for data collecting, processing, archiving, and publishing, with the possibility of reuse. Nowadays, many funding agencies require data to be made visible to the public after scientific research is completed, and this is a quite common practice among leading scientific journals. Also, one of the main challenges for the sciences that create the data is facilitating finding and accessing the data, so the need for better infrastructure to ensure better visibility is clear. The researchers from Serbia are still not familiar with the term data archiving and data publishing, or with the benefits of data archiving and publishing. In this paper, the authors present the process of archiving and publishing of the data, their benefits, digital archives that can be used for archiving, possibilities of access, levels of protection, possibilities for using data licenses, and potential for data reuse.

Correspondence:
Jelena Banović

e-mail:
jelena.banovic@ien.bg.ac.rs

DOI: 10.15308/Sinteza-2020-283-288

2. ARCHIVING AND PUBLISHING THE DATA

New requirements in science create the need for expanding and connecting research teams. Scientists can connect on conferences, while collaboration with researchers provides opportunities for them to improve their careers and strengthen their respective scientific fields. However, the development of ICT has made it possible for the researchers to connect because they now have access to the datasets of their colleagues, which were created during scientific research. The need for this kind of connection appeared before 2000, but software solutions were not developed as they are today. Growing consciousness of the importance of making the data public has an effect on the development of better software that can be used for these purposes. If we start from the hypothesis that science is a public good, then it is obvious why new achievements and findings have to be archived and published for future generations. Also, archiving and publishing can be helpful for some future research, and creators of public policies and procedures. Therefore, many theorists claim that scientists do not need to restrict access to their research results, but that they should present their data to the scientific community. According to Bradić-Martinović and Zdravković, researchers from Bosnia and Herzegovina, Croatia and Serbia (BiH 75%, Croatia 51%, Serbia 64%) have stated that sharing research data is very important in their discipline and only 2% (on average) believe that it is not very important [1].

Picture 1: Data life cycle [4]

The practice of data archiving and publishing appeared in the 1950s [2]. Data archiving is storing the data in digital storage. The main goal of archiving is the possibility of referencing and reusing the data in some future research. In data archiving, it is crucial to follow the data management procedures in order to ensure protection in the long term. The data management experts recommend placing the data on Cloud systems or some other digital repository. Also, data should be accompanied by storage documentation, so it can be comprehensible in the future.

Good archiving and publishing of the data have big contributions to the scientific community:

- Encourage scientific research
- Promote innovation
- Provide opportunities for new collaborations
- Transparency and accountability are maximized
- Improves research methods
- The cost of duplicating data collection is reduced
- The impact and visibility of research is increasing
- Provides important resources for education and learning. [2]

The awareness of the importance of data archiving and publishing, however, is not at enviable level, especially in Serbia. Because of that, data creators need to be aware of all the benefits of sharing and archiving of the data. On the other hand, according to Bradić-Martinović and Zdravković, the majority of researchers in Bosnia and Herzegovina, Croatia, and Serbia would archive research data if the data would be kept safe with regulated access because 45% of them (on average) answered with *Yes, certainly* and 40% answered with *Yes, probably* [1]. This research shows that the researchers were hesitant because they were not familiar with the process of data sharing and managing. Corti et al. state that researchers often think their results will not be interesting to other researchers; they are biased when it comes to paying for shared data in digital archives [2]. On the other hand, according to Data management expert guide (DMEG), the research conducted by Van den Eynden & Bishop, 2014; Hahnel et al., 2017, shows that some of the main motives for archiving and the publishing are [3]:

- career benefits – numerous studies show that archiving and publishing lead to increased visibility of researchers and recognition of scholarly work.
- scientific progress – data archiving and publishing have direct benefits for the research itself

(more robust), as well as benefits for the discipline and science in general by enabling new collaborations, new data usages, and establishing connections with the next generation of researchers.

- norms – norms of the project, research group, and/or discipline may determine whether a researcher is prone to publishing data. Overall, the availability of research data is at the heart of scientific ethics, as illustrated by the quote below.
- external drivers – funding agencies and publishers of scientific journals. Many funding agencies require that the data, following the research process, be made available so that users can access and reuse it. On the other hand, scientific journals are increasingly adopting data availability policies that advise or even request authors of manuscripts to make the research data, on which the manuscript is based, available.

According to Fienberg, the benefits of data archiving [4]:

- Reinforces open scientific inquiry. When data are widely available, the self-correcting features of science work most efectively.
- Encourages diversity of analysis and opinions.
- Promotes new research and allows for the testing of new or alternative methods.
- Improves methods of data collection and measurement through the scrutiny of others.
- Reduces costs by avoiding duplicate data collection efforts.
- Provides an important resources for training in research.

Also, many researchers are concerned about the issue of whether their data is good enough, and if it is useful to the research community. The truth is that some data have greater potential for reuse than others. The main questions to pose in order to understand if the data has value are [3]:

- Does the data have reuse potential?
- Does the data have national or international importance?
- Is the data unique and authentic?
- Is the data original, does it fit into the Big Data concept?
- Does the data come from some innovative research?
- Is the data set reusable? Is descriptive metadata available?
- It is important to point out that even if the data is not sufficiently good at one point, it can always be properly documented later, and shared retrospectively.

Data management experts declare that the most important thing in data archiving and publishing is time, because if the data is archived when the project is over, the project team's knowledge is higher, and the data can be properly described, selected and accompanied by timely documentation. As such, it will take a shorter amount of time to prepare the data for a deposit while simultaneously guaranteeing the highest possible data quality for future users [3].

Data publishing is providing access to data. Also, data publishing is the public distribution of collected data, by which data becomes visible, searchable, accessible, and ready for reuse. Many researchers refuse to make the data public because they do not recognize its benefits or they believe that the data is not of interest or importance to the scientific community, that others may misinterpret it, or that potential misuse may occur.

3. HOW TO ARCHIVE AND PUBLIC THE DATA?

The data may be archived and published on the online platforms designed for it. Depending on the scientific conditions, these online platforms are either national repositories that comply with standards for data storage or institutional repositories. Researchers can choose between self-archiving and archiving with the help of experts, but even if self-archiving is a fast way to publish the data, with the experts' help, the data can be properly treated, protected, and prepared for reuse. According to the researchers from the Institute of Edinburg [4], good publishing practice is followed with metadata, documentation, control, and review of the experts, furthermore, the data is findable. Although the publishing of the data that does not follow these criteria is not necessarily bad, there is no guarantee that the data will be stored on the same platform after a certain period, that the files will not be damaged, or that access to the data will remain the same.

For good publishing, a digital resource is essential, in the form of a digital repository or digital archive. Although there are many platforms built for this purpose today, it is always advisable to find a qualified archive that follows all standards of data archiving and publishing.

It is recommended that researchers archive their data in national archives, but if they do not exist, they have to find a suitable repository to archive their data. It is always a good practice to follow OpenAire guidelines for digital archives [3]:

- Choose trusted archive
- Choose institutional archive
- If neither of the previous two is available, choose one of public platforms online that has good politics and procedures.

4. HOW TO MAKE YOUR DATA VISIBLE?

This is a major issue that comes after the process of archiving and publishing. Data can be either open access or restricted and publishing the data in the digital archives does not necessarily mean that the data will be open access. For example, sensitive data, in some cases, needs to be restricted, secured from public access. It is always advisable to choose open access where it is possible, but with quality licensing. If open access is chosen, researchers have access to the data, and it is more likely that some will reuse it, and that the data will have an impact on someone else's work.

Access to data:

- Open Access – data can be accessed by users, whether they are registered users or not. Data in this category do not contain any personal information.
- Access for registered users only – data in this category do not contain direct identifiers, but there is a possibility of revealing the identities of respondents by linking indirect identifiers, and this represents a high risk.
- Limited access – access to the data in this category is possible only on request. In most cases, it is very sensitive data that must be restricted due to the delicacy of its nature. Sometimes, these data can have an embargo, which means that they are unavailable for a certain period, and only basic metadata is available during that period; when the embargo expires, the data is available to the users.

For example, Social Science Data Archives from Slovenia (ADP) has a very detailed and rigorous data access policy. ADP data can be open access data - where anyone can use the data, data with standard access - where registration is required, and data access under special conditions - where the users ask for permission from the data center to use the data [5]. Also, there is special access, where the user signs a contract to access the data and commits that he will use the data only for certain purposes and access it from secure rooms. An example of ADP shows that each archive regulates access to its data following its policies and procedures.

On the other hand, Data Archiving and Networked Services (DANS) also have defined data access policies. Although DANS supports the open access initiative and encourages the movement of public data without any restrictions, that often is not possible, because of the nature of the data. Because of this, DANS also has access categories – open access, access for registered users, and completely restricted access, when the user must ask for written permission to access the data [6].

5. FAIR DATA

Data archiving and publishing should follow a set of specific principles that facilitate data finding. FAIR is a set of crucial principles that make data findable, accessible, interoperable, and reusable [7]. The term FAIR was launched at a Lorentz workshop in 2014, and the resulting FAIR principles were published in 2016 [7]. The European Union, as well as many financiers, faculties, and institutes involved in mass production of data, advocate the use of the FAIR concept. These principles can serve as guidelines for creating data management tools and infrastructures, but also for defining policies and procedures for archiving and publishing scientific data. Following the FAIR principles, a framework of protection and easier recognition is provided [3]:

- Findable – this principle advocates that the data, produced during a project, has a unique digital object identifier (DOI). Also, in order for this principle to be fulfilled, it is recommended that data is described in metadata.
- Accessible – this principle is related to the availability of the data. Is the data open access or restricted? Documentation on the software for accessing the data needs to be included. Priority should be given to certified archives.
- Interoperable – this principle is related to enabling sharing and reuse of the data among researchers, institutions, and organizations. The main question is what kind of data, standards, or metadata methodologies will be followed to make the data interoperable..

- Reusable – this principle is related to how the data will be licensed and protected to allow reuse. Also, when will the data be available to the public? If there is an embargo, it has to be explained why the embargo exists and how long it will apply. Also, if the reuse of some data is restricted, the reason for it has to be defined.

6. DATA CITING

Data citing is a very important process in enabling easy data identification and connecting data with its creators. According to DMEG, there are two different types of identifiers [3]:

- A persistent identifier (PID) to your dataset – ensures that the data fulfills FAIR principles.
- A persistent author identifier – ORCID is the most common. With ORCID, it is easier to identify an individual researcher in the academic community and to connect a researcher to the dataset.

7. PROTECTION OF ARCHIVED AND PUBLISHED DATA WITH LICENSES

Despite the ability to access the data, it is very important to protect archived data with licenses, usually Creative Commons licenses. The choice of license most often depends on the nature and structure of the data. The benefits of Creative Commons licenses - they are very easy to use, they are widely accepted in the scientific community, they are very flexible and their readability and recognition enables finding data [5]. Data management experts point out that the most important thing is determining who owns the copyright. DMEG highlights that one open-access license for data is the Creative Commons license CC0. The copyright owner waives all his rights, including the database right and the right to be identified as the creator [3].

8. THE CURRENT SITUATION IN EUROPE AND SERBIA

The Consortium of European Social Science Data Archives – CESSDA ERIC is one of the most important organizations in this field in Europe. CESSDA provides large-scale, integrated, and sustainable data services to the social sciences. It brings together social science data archives across Europe, with the aim of promoting the results of social science research and supporting national and international research and cooperation [8]. CESSDA is a good example of a reliable archive, and their experts are always available to help researchers understand all the steps in the process of archiving and publishing the data. On the other hand, by following the CESSDA procedures, research data will be safe, visibility of the data will increase, and it will be easier to find it, while the potential for long term preservation is at a high level. Data is accessible and safe from unauthorized use.

Serbia has a national digital center for data archiving. Data Center Serbia for Social Sciences is an organizational part of the Institute of Economic Sciences, established in 2014. The Center is part of the national infrastructure and provides long-term preservation of scientific data to all researchers in the field of social sciences in Serbia. Also, the Center is a national service provider in this field, supported by the Ministry of Education, Science, and Technological Development of the Republic of Serbia. Data Center is a part of the Consortium of European Social Science Data Archives. Data Center Serbia uses the *Eprints* platform for data archiving and publishing and follows main CESSDA procedures in the process of data management.

9. CONCLUSION

Every research team during a science project should keep in mind two sides of work – first, how the project will be implemented, and second, how to manage data. Data management is a very demanding job and involves a set of skills and knowledge about processes concerning data processing. Also, it is important to improve researchers' awareness of the potentials of data, and of all benefits that can come from properly processing, archiving, and publishing data. In Serbia, researchers are not that familiar with data archiving and publishing, as are the researchers in Europe. The Ministry of Education, Science, and Technological Development has adopted the Open Science Platform in the middle of 2018. This document is based on two postulates – archiving and publishing of scientific papers of researchers from Serbia, which is required; and archiving and publishing of the data, which is still at the recommendation level. It is important to inform the researchers about all processes that they can perform on the data, with the aim of improving and empowering science, and connecting the researchers in Serbia and Europe.

As a national service provider, Data Center Serbia for Social Sciences is available to provide help to the scientific community in Serbia.

ACKNOWLEDGMENT

This papers is supported by Ministry of education, science and technological development of the Republic of Serbia.

REFERENCES

[1] A. Bradić-Matinović, A. Zdravković, „Researchers' interest in data service in Bosnia and Herzegovina, Croatia, and Serbia. IASSIST quarterly, vol. 38, no 2. pp. 22-28. 2014.

[2] L. Corti, V. Van den Eynden, L.Bishop, B. Morgan-Brett. Managing and Sharing Data. Colchester: UK Data Archive. 2011.

[3] Data Management Expert Guide, CESSDA. https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide accessed: 10.3.2020

[4] S. E. Fienberg. „Sharing Statistical Data in the Biomedical and Health Science: Ethical, Institutional, Legal and Professional Dimensions". In Annual Review of Public Healt, 15, 1994.

[5] Arhiv družboslovnih podatkov: ADP. https://www.adp.fdv.uni-lj.si/ accessed: 11.3.2020.

[6] Data Archiving and Networked Services: DANS. https://www.adp.fdv.uni-lj.si/ accessed: 11.3.2020

[7] https://www.force11.org/group/fairgroup/fairprinciples accessed: 9.3.2020.

[8] CESSDA. https://www.cessda.eu/ accessed: 8.3.2020.