National Registry of Scientific Data Management Plans by IBICT

Fabiano Couto Corrêa da Silva[1],
Elizabete Cristina de Souza de Aguiar Monteiro[2(✉)],
Marcello Mundim Rodrigues[3], Miguel Ángel Márdero Arellano[4],
and Alexandre Faria de Oliveira[4]

[1] Federal University of Rio Grande do Sul, Av. Ramiro Barcelos, Porto Alegre 2275, Brazil
[2] São Paulo State University (UNESP), Av. Hygino Muzzi Filho, Marília 737, Brazil
ecsamonteiro@gmail.com
[3] Federal University of Minas Gerais, Av. Pres. Antônio Carlos, Belo Horizonte 6627, Brazil
[4] Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), SAUS Quadra 5, Lote 6, Bloco H, Brasília, Brazil
{miguel,alexandreoliveira}@ibict.br

Abstract. A Data Management Plan is a document where the main aspects of data management are recorded and sent to funding agencies. In addition, it is part of the data documentation and is filed together with the data in repositories or published by scientific journals. The National Registry of Scientific Data Management Plans (NRSDMP) by IBICT is a tool that seeks assisting researchers on their (meta) data planning and management. It was designed based on the FAIR Principles, which are guidelines to be followed in the search for findable, accessible, interoperable, and reusable data and metadata. The objectives were to: present the NRSDMP developed by IBICT; describe its implementation process; and demonstrate its compliance to the FAIR Principles. The study was designed in a qualitative research of exploratory and descriptive nature with use of the action-research method. The results contribute to the enrichment of theoretical discussions regarding factors that contribute to data management and its documentation, such as the data management plan, articulating the practices applied to the preparation of a FAIR DMP and with instructions that allow making data and metadata FAIR. The NRSDMP also serves as guidance to researchers for reaching FAIRness. In conclusion, the paper presents theoretical and practical contributions to: the development of FAIR compliant tools; as well as the creation of educational content (data literacy) for researchers in need of planning their data management, all in accordance with the international community.

Keywords: Data management · Data repository · FAIR Principles

1 Introduction

Data management integrates several professional and scientific systematized activities, with the application of procedures, protocols, standards or methodologies aiming to

A. L. Pinto and R. Arencibia-Jorge (Eds.): DIONE 2022, LNICST 452, pp. 29–38, 2022.
https://doi.org/10.1007/978-3-031-22324-2_3

collect, extract or generate, transform, store, treat, share, preserve and, in some cases, discard the data [1]. Such management is a prerequisite to the effectiveness in sharing data among the various scientific communities and contributes with communities that do not integrate the target group of the research, but that may have benefits with the data reuse [2].

A relevant element in the management of scientific data is to ensure that data can be understood and interpreted by other researchers over time and in divergent research from that which collected or generated the data. Data documentation serves this purpose. The Data Management Plan (DMP) - a document prepared by researchers with the initial intent of meeting the requirements of funding agencies for applying or granting funding - is part of such documentation and provides the contextualization of the research and its data, contributing to researchers who collect and manipulate datasets, to those professionals who work in scientific data repositories, and to data consumers.

Most researchers collect data with some sort of preconceived plan, but it is often incomplete or inadequately documented. In order to stimulate planning on the methods needed to preserve scientific data and make them available to other researchers, some agents in the process - such as scientific publishers, funding institutions, libraries, and repositories - have promoted the standardized use of Data Management Plans (DMPs), usually meeting specific institutional guidelines.

There are some open tools that enable the creation of DMPs, including free persistent identifiers (DMPTool, Pagoda DMPOnline and others). As a common characteristic among the available model forms for DMP description, they all constitute a formal document that describes the whole data life cycle, from its collection to the complete documentation of the research process. They record the decisions made regarding metadata standards, formats, databases, methods, security, and storage periods, as well as the costs associated with data management. Thus, a data management plan requires a documented sequence of actions designed to identify, secure resources, collect, maintain, protect, and use data archives. This includes securing funding and identifying technical and personnel resources for the complete data management cycle. Once the needs for data use are determined, the next step should be adopting a system to store and manipulate data that can be identified and developed.

The scope and amount of detail in a data management plan depends on the project itself and the audience for which it is being developed. In general, whether a researcher seeks scholarship or research funding from an agency, or seeks to organize the methods for data preservation, one might look for forms external to his institution, usually foreigner ones, making the national DMPs dispersed in random repositories. These plans have high value to national research, as they require a description of the project and the data generated, where and how they are stored, both in the short and long term, also the access provisions and legal requirements that adhere to them. Several tools are available in order to assist the elaboration of a DMP. In Brazil, the project called Cadastro Nacional de Planos de Gestão de Dados Científicos or National Registry of Scientific Data Management Plans (NRSDMP) has the intention of developing and making publicly available a tool that meets the needs of researchers. The tool has as its main characteristic meeting the FAIR Principles, guidelines to be followed in the search for findable, accessible, interoperable, and reusable data and metadata.

The FAIR Principles are intended to define a minimum set of principles and practices (related, but independent and separable), which allow machines and humans to find, access, interoperate and reuse scientific data and metadata [3] (RESEARCH DATA ALLIANCE, 2020, p. 4). This research communication presents a holistic view of the factors underlying the development of a data management plan tool that is compliant to the FAIR Principles.

The general objective of this paper was to present the nrsDMP project. The specific objective was to describe its implementation and its suitability to FAIR Principles.

The results of this study contribute to the enrichment of theoretical discussions regarding factors that contribute to data management and its documentation, such as the data management plan, articulating the practices applied to the preparation of a FAIR DMP and with instructions that allow making data and metadata FAIR. In possession of such information, researchers have an overview related to the topic and attention in preparing and publishing data on digital platforms.

Starting from the lack of a national registry of DMPs, as well as the growing demand from funding agencies that demand data management planning from sponsored researchers, the general objective of the NRSDMP by IBICT is to collaborate reducing the efforts on adapting demands for Open Science and increasing the visibility of the research carried out by Brazilian researchers. As for the specific objectives, we seek to: a) Develop of a tool a national DMP model by means of a form that gathers registries of Universities, Research Centers and Funding Agencies at a national level; b) Implement all the indicators of the FAIR Principles to the form (the DMP will receive an extract on the fulfillment of the FAIR Principles); and c) Develop a repository for the digital preservation of the DMPs, and assign individual persistent identifiers to its digital objects.

The present work also deals with the preservation of Data Management Plans due to their importance for the scientific communication process involving the necessary steps for digital preservation. During the NRSDMP development, we discovered the need for a DMP repository, where it could to preserve DMP, publish some descriptive, administrative, technical, preservation, use and reuse metadata, assign persistent identifiers to metadata, link metadata to related digital objects in the web, metadata standards, Knowledge Organization Systems (KOS), also trace data provenance, etc.

2 Methodological Procedures

The project adopted a triangulation of methods, consisting of a theoretical, bibliographic, and empirical research. We applied the content analysis method when investigating the DMPs deposited in research funding institutions' websites and repositories. There's also a continuous consultation of the scientific literature available. The empirical phase intends to directly question researchers on their demands and doubts facing the need of filling out their DMPs. The dialogue with researchers will occur after nrsDMP is launched publicly, and the suggestions will be incorporated gradually. The development is being done in partnership with the Brazilian Institute for Information in Science and Technology (IBICT), two librarians and a web developer. The goal is to make available a public tool so that anyone can register DMPs nationwide, and deposit them into a

repository for their management and preservation, guaranteeing long-term access for research institutions (laboratories, universities, etc.), funding agencies (development agencies, government, etc.), and the society.

For the development of the tool, qualitative research of exploratory and descriptive nature was used, based on a bibliographic and documental search that contemplates theoretical and applied recommendations for the development, implementation, and guidance on its use.

So, the method adopted was action research, of empirical basis, applied to any process where a cycle is performed in which it is improved with systematic action between acting on the practice and investigating it. [4]. Therefore, the steps for the development of the tool were followed: diagnosis, planning, implementation, description and evaluation.

1) Diagnosis: we performed a diagnosis of the characteristics and features of the DMPTool, DMPOnline, and Data Stewardship Wizard; the application of the FAIR Principles, possible improvements and implementation of features.
2) Planning: the actions and implementations for the diagnosed framework were defined.
3) Action: application of action research with the implementation of the planned actions.
4) Evaluation: analysis of the implementation of the actions against the theoretical and practical support used as a starting point for the definition of the actions.
5) Learning: learning from the tests in the elaboration of the Data Management Plan by the group.

After the application of the action research steps, a pilot will be applied to researchers from different fields, which aims to identify their demands on the NRSDMP filling out and to indicate adjustments.

3 FAIR Data Management Plan

The Data Management Plan is a document in which the different activities and processes related to the data life cycle are described and involves [...] data design and creation, storage, security, preservation, retrieval, sharing, and reuse, all taking into account technical capabilities, ethical considerations, legal issues, and governance structures [5].

The Data Management Plan deals with the data that will be generated or collected during a given research study. Since it is a plan, it is based on future actions. Therefore, the researcher who answers the plan questionnaire is making conjectures, planning, proposing what will be done in the near future. In the data management plan, the researcher is contextualizing the research, saying what will be done before, during and after the data collection, where the data will be preserved and how it will be made publicly available. As a proposal, it is not possible to state precisely what information about the data will be collected, but it is reasonable to point out ways.

The FAIR Principles describe resources necessary to achieve archiving, preservation, maintenance, organization, retrieval, use and reuse of scientific data. Among such resources are those related to the software infrastructure that manages the platform on

which the data will be available during or after the completion of a particular research project.

The FAIR Principles seek reasonableness between humans and machines, for common (universal) communication and understanding between data and metadata, systems and vocabularies. FAIR data and metadata are contained in the fourth scientific paradigm (e-Science), where the volume of scientific data expands exponentially. The FAIR transformation is linked to digital preservation, knowledge organization, and information retrieval.

Central to the realization of FAIR are FAIR Digital Objects. These objects could represent data, software, protocols or other research resources. They need to be accompanied by Persistent Identifiers (PIDs) and metadata rich enough to enable them to be reliably found, used and cited. Data should, in addition, be represented in common – and ideally open – formats, and be richly documented using metadata standards and vocabularies adopted by the related research community to enable interoperability and reuse [6].

It is important to emphasize that FAIR data is not necessarily open data, because 'fairification' provides licenses and copyright to the use and reuse of its assets. Some data can and should be embargoed or restricted from access and reuse, for various reasons.

The FAIR Principles are divided into four acronyms. Each acronym is subdivided into 3 or 4 statements that serve as valuable guidelines for managers, technicians, and experts in digital preservation, information systems, and knowledge organization. Figure 1 presents each of the 15 FAIR sub-principles. The FAIR goal is hardly ever achieved, because data and metadata are not always one hundred percent FAIR. This means that the control that one must have involves many variables, between humans and machines, with one of them being vocabulary control, since we do not have a universal language between people and systems. In this way, the FAIR ideal becomes a great challenge to contemporary society.

The import and use of metadata that describe scientific data amongst systems is a commitment of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

Still, in order to meet some of the FAIR Principles, one must have indexer expertise, in the case where one must use resources/instruments when indexing representative terms of domain content. Therefore, in order to align with the FAIR Principles, a dataset needs to be deposited in a FAIR repository, such as Dataverse; it needs to be well described; it needs to be indexed; etc.

During the NRSDMP development, we discovered the need for a DMP repository, where it could publish some descriptive, administrative, technical, preservation, use and reuse metadata, assign persistent identifiers to metadata, link metadata to related digital objects in the web, metadata standards, Knowledge Organization Systems (KOS), also trace data provenance, etc.

In that way, the DMPs will be stored in a Dataverse repository called Cariniana: Brazilian Network for Digital Preservation Services by the Brazilian Institute of Information in Science and Technology (IBICT), which aims to ensure continuous access and long-term preservation of the content stored therein [7]. The Dataverse platform is a free software architecture for the archiving, publishing, preservation, reuse, analysis and

The FAIR guiding principles: https://doi.org/10.1038/sdata.2016.18

To be Findable:

F1. (meta)data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata (defined by R1 below)

F3. metadata clearly and explicitly include the identifier of the data it describes

F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1. the protocol is free, open and universally implementable

A1.2. the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2. (meta)data uses vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

To be reusable:

R1. (meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with data provenance

R1.3. (meta)data meet domain relevant community standards

Fig. 1. The FAIR principles

citation, analysis of digital objects [7]. Repositories implemented through Dataverse are in greater compliance with the FAIR Principles [8].

4 Results and Discussions

This section presents and discusses the results and analyses related to the phases of action research:

1) Diagnosis: the initial diagnosis highlighted the characteristics and requirements in the implementation of the tool and the steps, questions and guidelines on the preparation of the DMP available in the tools and templates according to Table 1. The functionalities, guidelines on the FAIR Principles, improvements and implementations of items not yet applied by the analyzed tools were identified. The tools evaluated are in column 1 and the templates consulted in column 2.
2) Planning: the implementations and actions of each group member were defined. With the diagnosis, the following were defined: categories with the questions to be answered, which information will be available for completion, and FAIR adaptation. The functionalities were also defined: certification and application of the institutional logo to each institution represented by its researchers, and the logo of the funding agencies. Furthermore, courses will be available on the NRSDMP website for orientation on its use, and on the management of scientific data.
3) Action: application of the action research with the starting actions of the planning stage.
4) Evaluation: evaluations are made at each action stage with verification of successes and mistakes, making the necessary adjustments.

5) Learning: learning by tests in the elaboration of the Data Management Plan. Afterwards, step 4 will be applied to researchers as a pilot stage.

Table 1. Templates evaluated for the indication of the questions in the tool

Tool	Templates
DMPTool	DMPTool Harvard University Ohio State University (OSU) Princeton University Stanford University Washington University
DMPOnline	Amsterdam University of Applied Sciences (Hogeschool van Amsterdam) Digital Curation Center (DCC) FAIRsFAIR - Fostering Fair Data Practices in Europe Imperial College London Lancaster University London School of Hygiene and Tropical Medicine (LSHTM)
Data Stewardship Wizard	Horizon 2020 Science Europe

It is recommended that the DMP be drafted early in the research workflow, as it can provide important insights into how to gather, curate, and disseminate data, building a common understanding in the data life cycle and providing the opportunity to reflect on decisions that will affect the FAIRness of the data [6]. For the DMP completion, guidance is given on how the data management planning could meet the FAIR Principles in ways users need not know their statements. In addition, it is indicated how users could meet Berners-Lee's Five Stars whilst planning (by answering the form) and/or managing their research data. Thus, upon completion of the DMP, the data and metadata must have:

- F1 A corresponding persistent identifier to ensure long-term access (The DMP will receive a Digital Object Identifier (DOI) with an IBICT prefix);
- F2 Data described by rich metadata (the questionnaire is exhaustive, and the DataCite Metadata Schema and Dublin Core metadata standards have been defined;
- F3 The persistent identifier associated with the DMP needs to be published on the landing page where the metadata is (the questionnaire has the field for inclusion);
- F4 The metadata need to be published on an accessible web page, that is, it needs to be indexed in a searchable resource, so that it is retrievable by Google search box or other search engines;
- A1 DMP metadata need to be retrievable by its identifier, using a standardized communication protocol. It has been requested to use OAI-PMH for the identifier syntax;
- A1.1 The OAI-PMH protocol is open and universally implementable;

- A1.2 The protocol must allow authentication and authorization, through user account creation and system access via login and password (CAFe/Keycloack was requested). The protocol must be processed by HMAC (hash-based message authentication code);
- A2 Metadata must be accessible, and to ensure this, persistent identifiers and a trusted server will be applied. The Cariniana federated network will be used for backup and long-term preservation (via CLOCKSS);
- I1 The metadata must be filled out in vernacular language, according to the current grammatical rules, it must be public and accessible, i.e., the record in the database must be compiled into a verifiable record at a persistent electronic address. This would be a referential base, a channel that addresses the metadata to the data (that will be published during or after the end of a given research);
- I2 The vocabularies cited are controlled vocabularies, generally responsible for the organization of information and knowledge in a specific scientific field. These vocabularies should comply with the FAIR Principles, i.e., they should be verifiable and accessible, ideally having persistent identification that leads the user to the address of the KOS in use. The data owner should choose to describe/contextualize their research formally, indexing subjects with the use of authorized terms, always citing their source;
- I3 External (meta)data sources or relationships amongst research are indicated with DOIs or electronic addresses of papers related to the research during plan completion;
- R1 Structured metadata fields, as well as well-described data, especially when using standards;
- R1.1 License options for the user, such as Creative Commons and others available;
- R1.2 Metadata fields indicating relationships between people and institutions, all involved in the research, the collection, and the composition of the datasets. That is, indication of relationships in the context of the research;
- R1.3 A list of controlled domain vocabularies. Users must choose from options available in the tool. Users must indicate metadata standards, standardized language by field of study.

Based on the understanding of the needs of each sub-principle to be met, we designed the following questions in order to guide users to the desired actions in achieving FAIR data and metadata:

- F1 When shared, will your dataset(s) have a persistent identifier (DOI, handle, etc.) assigned to it by the system?
- F2 What metadata will be provided to help others identify and discover the data? / What metadata standard will be used?
- F3 Will the deposit of your dataset(s) have its persistent identifier (DOI, handle, etc.) published along with the metadata at an accessible and retrievable electronic address (e.g. via Google)?
- F4 Will your dataset(s) be published in an accessible and retrievable electronic address (e.g., via Google)?
- A1 Does the system in which you will deposit, share and preserve your dataset(s) allow for retrieval via a persistent identifier?

- A1.1 Is the persistent identifier configured as a link, with syntax starting with HTTP or HTTPS (e.g.: https://doi.org/10.5072/FK2/O4HLLA)?
- A1.2 Does the system where you will deposit, share and preserve your dataset(s) allow access via login (username and password)?
- A2 Will the system in which you intend to deposit, share and preserve your dataset(s) keep your metadata accessible, even after the data becomes obsolete?
- I1/I2 What controlled vocabulary/thesaurus will be used to describe the research and data collected?
- I1/I2 Does the controlled vocabulary/thesaurus in use have a persistent identifier?
- R1 Which metadata standard will be used?
- R1.1 Inform the license that will be applied to the data.
- I.3/R1.2 Point out the electronic address or identifier (DOI, handle, etc.) that connects the source to the data or metadata.
- R1.2 Who owns the rights to the data?
- R1.2 Under which license? If not, who has granted the right of reuse?
- R1.3 Is the metadata standard in use specific to the research field? E.g., the Social Sciences field uses DDI (native in Dataverse).
- R1.3 Is the controlled vocabulary/thesaurus in use specific to the research field? E.g.: The field of Agricultural Sciences uses AGROVOC.

On the homepage of the NRSDMP by IBICT, there will be an educational course on data management and on the filling out of a Data Management Plan prepared/instructed by the team. It will contain videos, texts and exercises. In addition, there will be a tutorial on the use of the NRSDMP by IBICT.

5 Final Considerations

This paper presents theoretical and practical contributions for the development of a tool to elaborate a Data Management Plan that meets the FAIR Principles. The NRSDMP by IBICT presents guidelines that serve as instructions for the management of the data life cycle and orient those who are involved in it in any way. It is recommended that the DMP be prepared at the beginning of the research project and, because it is a dynamic document, it can be updated throughout the development of the project.

Open science represents a paradigm shift in the way science is conducted, which involves conducting all stages of scientific research (design, data collection, review, publication, etc.) with an "open" vision. Thus, the change is not in what is done nor in the means available to do it, but in how it is done. Therefore, it does not involve a scientific or technological breakthrough, but has a social and cultural character; it represents a new way of conducting and understanding scientific research, and also of disseminating and evaluating it.

These changes in the habits and behavior of researchers are being made official and will gradually be put on the agenda of Brazilian scientists through NRSDMP by IBICT, especially after the requirements made by the funding agencies are expanded. This will lead to the implementation and drive of a change that intends to revolutionize the way research is conducted throughout Brazil.

References

1. Monteiro, E.C.S.: A gestão de dados em seu ciclo de vida: principais aspectos. In: Sena, P., Pinheiro, M. (eds.) Profissionais da informação no contexto de inovações tecnológicas. No prelo (2022)
2. Monteiro, E.C.D.S.A., Sant'ana, R.C.G.: Plano de gerenciamento de dados em repositórios de dados de universidades. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis **23**(53), 160–173 (2018). https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n53p160. Accessed 21 Sep 2020
3. Research Data Alliance. FAIR data maturity model: specification and guidelines. RDA FAIR data maturity model Working Group (2020). https://doi.org/10.15497/rda00045. https://www.rd-alliance.org/system/files/FAIR%20Data%20Maturity%20Model_%20specification%20and%20guidelines_v0.90.pdf. Accessed 21 Sep 2020
4. Tripp, D.: Pesquisa-ação: uma introdução metodológica. Educ. Pesqui. **31**(3), 443–466 (2005)
5. Cox, A.M., Pinfield, S.: Research data management and libraries: current activities and future priorities. J. Librariansh. Inf. Sci. **46**(4), 299–316 (2014). http://lis.sagepub.com/content/46/4/299.full.pdf+html. Accessed 25 Oct 2020
6. European Commission. Turning FAIR into reality. Luxemburgo: Final Report and Action Plan from the European Commission Expert Group on FAIR Data. Publications Office of the European Union (2018). https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_1.pdf. Accessed 25 Oct 2020
7. Araújo, L.M.S., Márdero Arellano, M.A., Ferrer, I.D.: Guia para os usuários do repositório Dataverse do Ibict. Ibict, Brasília (2018). https://cariniana.ibict.br/images/artigos/Dataverse/Guia_Usuarios_Dataverse_Ibict.pdf. Accessed 11 Feb 2022
8. Rodrigues, M.M., Dias, G.A., de Lourenço, C.A.: Repositórios de dados científicos na América do Sul: uma análise da conformidade com os Princípios FAIR. Em Questão, Porto Alegre, Online First, e-113057 (2022)