8. IMPLEMENTATION OF FAIR PRINCIPLES IN SCIENTIFIC DATA REPOSITORIES: a comparative analysis of DSpace and Dataverse software infrastructures

Fabiano Couto Corrêa da Silva[132]
Marcello Mundim Rodrigues[133]

8.1 INTRODUCTION

When planning institutional policies regarding management and curation of scientific data, Higher Education Institutions face several questions whose answers are complex, such as the identification of the most appropriate software infrastructure to preserve and organize heterogeneous scientific data. Likewise, graduate programs and organizations that fund, conduct, or support research in any other way are responsible for promoting and ensuring an adequate management of data and information coming from their activities. As obvious as it may be, it is worth stressing that an efficient data management is essential to guarantee that such assets are accessible now and in the future.

FAIR principles (findable, accessible, interoperable, and reusable data) were developed to assist surpassing common barriers to data finding and reusing, which has long been recognized as a problem in the scientific research, presenting guidelines that support the long-term retention and availability of these datasets. The FAIR aspects of findable and accessible data are mainly related to where data are deposited. Important points to be considered include the availability of the persistent identifier of the digital object, metadata, data reuse monitoring, licensing, access control, retention and long-term availability.

The FAIR aspects of interoperable and reusable data point out the need to reflex over issues that cover data format (holder x open), its updating or obsolescence, the interoperability (opening via Application Programming Interface [API]) of the repository selected to other international or disciplinary meta-repositories, or other improvement tools. The aspect of detailed documentation is also considered in the capacity of data reuse.

From this, it is noted that the research aimed at identifying the differences between two software infrastructures related to management and curation of scientific data in the light of the standard proposed by the FAIR principles.

---

132 Doutor em Información y Documentación pela Universitat de Barcelona, Professor no Programa de Pós-Graduação em Ciência da Informação, Universidade Federal do Rio Grande do Sul. E-mail: fabianocc@gmail.com.

133 Doutorando em Ciência da Informação pela Universidade Federal de Minas Gerais, Bibliotecário-Documentalista na Universidade Federal de Uberlândia. E-mail: marcellomundim@ufu.br.

8.2 THEORETICAL FRAMEWORK

Data infrastructure refers to more than only data archiving. It includes taking care of data from the moment it is created. Although open access fights for barrier-free access, it does not mean that all publications and scientific data are available. New licenses have been developed to offer alternatives to copyright. For instance, Creative Commons licenses provide a flexible variety of protections and freedom to authors, artists, and educators (Creative Commons *apud* Doorn; Tjalsma, 2007, p. 14).

In order to Reuters include data repositories in its Data Citation Index, these need to meet certain criteria, such as show stability of data objects and of the repository that supervises its curation, as well as curation standards and data release, and links established to academic research (Force; Auld, 2014, p. 97).

Lee and Stvilia (2014) understand that the definition of identifier should mention characteristics of identification systems, types of entities assigned, and purpose of identifiers. They define a data identifier as a sequence of symbols drawn to identify, cite, note, and/or link scientific data to its metadata. Different systems of identification can be used to refer to distinct types of entities (Lee; Stvilia, 2014, p. 3).

Schopfel and others (2014) explore scientific data related to electronic theses and dissertations as a specific part of the emergent infrastructure of research. Computer systems, data, informational resources, networking, digitally activated sensors, instruments, virtual organizations, observatories, services, and tools interoperable through software – these are the technological components of cyberinfrastructure, which was defined by the US National Science Foundation Cyberinfrastructure Council in 2007 (Schopfel *et al*., 2014, p. 613).

In the past, hard copies of thesis and dissertations were submitted with supplementary materials in several formats and different supports (print attached, punch card, floppy disk, audio tape, slide, CD-ROM, among others), which made its processing (file location) and reuse difficult. In the new infrastructure of electronic thesis and dissertation, these materials can be submitted and processed with text files. If they are disseminated via open repositories, these research results could become a rich source of scientific datasets to be reused and explored. These complementary materials are generally small data or little science, hidden and unexplored data, from public funding and personal production. Its large variety affects its accessibility, openness and reusability (Schopfel *et al*., 2014, p. 616).

Making access to scientific data related to digital thesis and dissertations available is a challenge to academic libraries, and due to that, Schopfel and others (2014) make three questions: "Which information system best meets such needs? How to facilitate the retrieval of these datasets? What are the legal conditions for their dissemination, access and reuse? (Schopfel *et al*., 2014, p. 618).

Data repositories can be institutional, as the majority of thesis and dissertation repositories, however, they are also managed by third-party providers such as Dryad, Zenodo or Figshare. Furthermore, heterogeneous scientific datasets cannot be compared to Big Data produced by CERN and others because they are similar to personal data[134]. The ideal architecture should combine characteristics of personal data warehouses (small data) with those of institutional information (big data). Due to the specific nature of data and supplementary files, it seems appropriate not to store text and data files in the same repository, but distinguish between document servers and data repositories, depositing texts and data in different platforms (Schopfel *et al*., 2014, p. 618).

Amorim and others (2016) observe that repositories such DSpace are widely used among institutions to manage publications, and that these institutions can support the platform and expand, meeting additional requirements. They point out that some repositories do not implement interfaces with repository indexers, which could affect the statistical updating of indexing databases (Amorim *et al*., 2016, p. 853). For the authors, the access to the source code can be a valuable criterion for selecting a platform, thus avoiding problems of discontinuity of certain service. The availability of the source code also allows additional modifications (personalized workflow).

Furthermore, it is understood that the existence of an API enables maintenance and future development of the repository. It is noted that some platforms fail when they do not offer unique identifiers to resources deposited, which makes data citation in publications more difficult (Amorim *et al*., 2016, p. 853). The authors point out that an institution can both outsource an external service, install and personalize its repository (assisting maintenance expenses). They state that DSpace, ePrints, CKAN or any Fedora solution[135] can be installed and run under the research institution's control (better control over stored data) (Amorim *et al*., 2016, p. 855).

ePrints and DSpace are not developed to assist collaborative environments in real time, where researchers can incrementally produce and describe their data. Adopting dynamic approaches to data management can encourage researchers to use management platforms as part of their daily activities, while they work on data (Amorim *et al*., 2016, p. 856-857). DSpace, known by its capacity of dealing with research publications, has also been recognized by handling scientific data (Amorim *et al*., 2016, p. 858).

Some institutions may want the servers where data are stored under their control, as well as directly manage their datasets. Platforms as DSpace or CKAN are appropriate for such actions, for they can be installed in an institutional server (Amorim *et al*., 2016, p. 860-861).

DOI can be used as a reference to the current location of data. It is also persistent, which means that once assigned, it can never be deleted nor reassigned (Beaujardière, 2016, p. 21).

Garnett and others (2018) point out that in the light of FAIR principles, scientific data should be structured to enable their finding by humans and machines. Without FAIR data, finding and reusing become difficult, for a single researcher may have to go to several places to find and access data (Garnett *et al*., 2018, p. 201-202).

134 Personal data is understood to be data generated in small volumes, but in different formats.

135 "Fedora creates an innovative, free and open source platform for hardware, clouds, and containers that enables software developers and community members to build customized solutions for their users." (Red Hat, [2020?]). Fedora has operating systems as specific solutions for specialized niches.

FAIR data must be machine-readable and actionable; also, it is not equivalent to open data; it is an aspiration, it is never entirely FAIR; when releasing restrict data, licenses, and agreements of data use must be clearly defined by authors or data providers; a repository platform such as Dataverse can make the creation of FAIR scientific data much easier; however, the data authors must contribute using metadata standards and appropriate community vocabulary (Crosas, [2019?]).

Between the DSpace and Dataverse platforms, Rocha (2018) conclude that:

> Dataverse has resources for configuration of several types of repository environments, including organizational hierarchies and distinct management policies for unities or groups, including metadata and licenses schemes. It is possible in DSpace; however, it demands adaptations or configurations, with some limitations in the control of versions (Rocha, 2018, p. 74).

Brownlee (2009) states that Dublin Core (DC) is appropriate to the bibliographic description of the majority of items, in cases in which the collection comprises traditional publication formats such as research articles and conference proceedings (Brownlee, 2009, p. 4).

Garnett and others (2018) go beyond and define DC as a descriptive metadata standard used by digital repositories such as DSpace, which describes digital objects including scientific data. DataCite assists the finding of scientific data on the Web when focusing on elements that define location, identification, and unique citation of these data. DataCite requires the creation of DOIs, which allows easy identification and data citation, and provide persistent metadata, whether data are open or not (Garnett *et al*., 2018, p. 205-206).

8.3 METHODOLOGY

The study is an applied, qualitative, exploratory, descriptive and document research (Creswell, 2014). The objects of study of this investigation were DSpace and Dataverse, platforms that aim at the conception of virtual environments known as digital repositories. It was not intended to investigate other software infrastructures, since recent investments in institutional repositories have focused on these two options. Therefore, we sought to evaluate the compliance of the standard infrastructure of DSpace and Dataverse with the FAIR principles. Thus, it was possible to identify if both platforms meet the minimal requirements to serve as natural tool to the FAIRification (FAIR adaptation process) of the investigation, when configuring data finding, its accessibility, interoperability, and reuse.

FAIR principles are divided in subprinciples, each one corresponding to a requirement suggested as being excellent to scientific data management and curation. It was identified that the subprinciples referring to the software infrastructure are F.1, F.3, F.4, A.1, A.1.2, A.2, and R.1.1.

The information found that served as results were retrieved from scientific literature and official documents hosted on their respective websites. The technique applied was the content analysis. For the results' presentation, a check-list was created referring to the FAIR subprinciples. In this check-list, it was sought to answer the following question: Is the software infrastructure (DSpace and/or Dataverse) in accordance with the FAIR x, y, z [...] sub-principle?

8.4 RESULTS

FAIR principles suggest standardization of techniques and environments referring to scientific datasets management and curation. Management involves knowledge organization processes; curation, long-term preservation. Preserving means guaranteeing safety throughout time; organizing is making something findable and accessible. Therefore, making data and metadata become FAIR demands efforts in different levels of a same workflow.

It can be said that the proposition of FAIR workflow is divided in three layers: a) standard layer of software infrastructure chosen for storing and preserving data and metadata; b) layer of technical knowledge held by managers, curators, and analysts; c) layer of domain knowledge held by data and metadata depositor/holder.

The first layer has objective characteristic, for even with open-source code, the default settings of a software are available for its users/clients, thus ensuring minimal standardization of its functions.

The other layers noticed are subject to the tacit knowledge (subjectivity) of the agents involved in the same scientific data and curating project, even with guidance based on a well-established institutional policy. For instance, when subprinciples F.2 and R.1 suggest that: F.2. data are described with valuable metadata; and R.1. (Meta)data are richly described with a plurality of relevant and accurate attributes (Go Fair, [2016?]); they are referring to the quality of data description through metadata and metadata standards or schemes that are inherent to human functions.

Fair data can be conceived as a spectrum or continuum ranging from partial to entirely FAIR digital objects. Similar to the five-star open data, different FAIR levels can be conceived to articulate minimal conditions to find and reuse richly documented and functionally linked FAIR data. It will vary according to the community. Some principles will be trivial to certain domains of research and problematic to others; therefore, each field of research needs to define what it means to be FAIR and decide the appropriate measures to evaluate it (European Commission, 2018, p. 51).

Therefore, it is noticed that some FAIR subprinciples can be responsible for the institutionalization of scientific data management and curation policies that are discrepant among research institutions, whether by the divergence among their objectives, needs, or teams and users/clients. It was not intended to investigate FAIR subprinciples aimed to the guidance of subjective practices. Thus, the research identified the subprinciples F.1, F.3, F.4, A.1, A.1.2, A.2 and R1.1 as pertaining to the first layer aforementioned.

These subprinciples attest that: F.1. (meta)data are assigned a unique and globally persistent identifier, F.3. Metadata clearly and explicitly include the data identifier that they describe; F.4. Metadata are recorded and indexed in a researchable resource; A.1. (Meta)data are retrievable through their identifiers using a standardized communication protocol; A.1.2. The protocol allows an authentication and authorization procedure when necessary; A.2. Metadata are accessible, even when data are not available anymore; and R.1.1. (Meta)data are released with a clear and accessible data use license (Go Fair, [2016?]).

Check-list

a. Is the software infrastructure in accordance with the FAIR F.1 subprinciple?

DSpace: Yes. Handle is a standard persistent identification system in DSpace (Unesco, 2014).

Dataverse: Yes. Dataverse network is an open-code application that provides guidelines and tools for data citation. Dataverse specifies the global record handle as its persistent identification system. The DOI can also be used as a Dataverse standard identifier system (Lee; Stvilia, 2014, p. 18-19).

b. Is the software infrastructure in accordance with the FAIR F.3 subprinciple?

DSpace: Yes. DSpace offers DC as predefined descriptive metadata scheme (Brownlee, 2009, p. 4). DC has in its standards the *dc:identifier* tag to the description of the persistent identifier assigned to data and metadata.

Dataverse: Yes. Dataverse allows the citation for whole dataset. DOI, with URL and metadata registered in DataCite. In addition, the citation for data file, with DOI and URL for each file (Crosas, [2019?]).

c. Is the software infrastructure in accordance with the FAIR F.4 subprinciple?

DSpace: Yes. It has an integrated search engine: DSpace comes with Apache Solr, an open-code corporatize search platform that allows the faceted research and the navigation in all the objects. The complete text of common file formats is researchable, with all metadata fields. The navigation interfaces are also configurable (Duraspace, 2020). DSpace is indexed in Google Scholar (Unesco, 2014).

Dataverse: Yes. Dataverse allows citation and detectable metadata using DataCite, schema.org, DC, DDI, e Schema.org JSON-LD standards (findable in Google Dataset Search) (Crosas, [2019?]).

d. Is the software infrastructure in accordance with the FAIR A.1 subprinciple?

DSpace: Yes. A single Handle server normally opens three network listeners in ports 2641 UDP, 2641 TCP and 8000 TCP. Port 2641 (UDP and TCP) is the port number assigned by Internet Assigned Numbers Authority (IANA) for the Handle cable protocol. The Handle service model and the connection protocol are described in RFC 3650, RFC 3651 and RFC 3652. TCP is generally necessary for administrative requests and is used as a substitute for whenever UDP is slow or unavailable. Port 8000 offers HTTP and HTTPS interface. Handle servers use "port unification" so that HTTP and HTTPS are available in the same port. If the standard Handle protocol ports are not available, clients can resort to wired tunneling protocol over HTTP. For any HTTP request that matches the proxy domain name with a Handle, for example: http://hdl.handle.net/20.1000/5555, one of the proxy servers will consult the Handle, obtain a URL in the Handle registration (or if there are several URLs in the Handle registration, one will be selected, and this selection is not in a specific order) and a HTTP redirection will be sent to this URL to the user's browser. If there is no URL value, the proxy will show the Handle registration (Corporation for National Research Initiatives, 2018).

Dataverse: Yes. In the maintenance of Handle as standard persistent identifier, the process will be the same as the aforementioned.

e. Is the software infrastructure in accordance with the FAIR A1.2 subprinciple?

DSpace: Yes. Safety: DSpace provides its own integrated system of authentication/authorization; however, it can also be integrated into existing authentication systems such as LDAP or Shibboleth (Duraspace, 2020). The current distribution of Handle.Net software uses standard Java cryptography libraries for low-level cryptography routines. The Handle system provides two ways of authentication: public key and secret key. In the current implementation, authentication of the public key is performed using DSA or RSA algorithm. Authentication of secret key depends on a safe MAC algorithm. In general, the authentication of secret key uses three parts: (1) the authentication client; (2) the server where the client is performing an operation; and (3) another server able to verify the client authentication (Corporation for National Research Initiatives, 2018).

Dataverse: Yes. In case the files are restricted data files, authentication and authorization are necessary (Crosas, [2019?]). In the attempt to access restricted data via its Handle, the communication protocol will be identical to the one previously described.

f. Is the software infrastructure in accordance with the FAIR A.2 subprinciple?

DSpace: Yes. DSpace is a set of Web applications in Java and utility programs in cooperation that keep assets and associate metadata storage. Web applications provide interfaces for administration, deposit, ingest, search and access. Asset's storage is kept in a file system or similar storage system. The metadata (including access and configuration information) are stored in a relational database. In addition, DSpace enables the temporary data embargo via author/creator's request. However, it maintains access to the metadata (DURASPACE).

Dataverse: An inactive destination page with the basic citation metadata will always be accessible to the public if it uses a persistent (Handle or DOI) provided in the citation for this dataset. Users will not be able to see any of the files or additional metadata that were available before the deactivation (Dataverse project, 2024). Dataverse stores information of the package structure (dataset) in relational database, that is, it stores packages in a software-dependent way. However, Dataverse allows exporting metadata from a dataset (dataset files not included) in DDI Codebook format, which results in an XML file that describes the whole package, including structural metadata (physical and logical structures of the documents, in addition to variables in tabular documents). (Rocha, 2018, p. 47).

g. Is the software infrastructure in accordance with the FAIR R1.1 subprinciple?

DSpace: Yes. Creative Commons is the standard license in DSpace (Unesco, 2014,).

Dataverse: Yes. By standard, all the new datasets created by Dataverse's User Web Interface receive a Creative Commons CC0 Public Domain Dedication (Dataverse Project, 2024).

8.5 FINAL CONSIDERATION

Data organization and manipulation are big challenges for the beginning of the 2020s. According to the International Data Corporation (2024), in every two years we double the quantity of produced data. In science, it is not entirely unique.

This concern in sharing is caused at the end of the research, what makes the available data not always meet the FAIR principles, tending to be low, or with no semantic, with a diversity of standards and formats. In order for data to be better used, they should have a rich semantic, and, according to Tim Berners-Lee, be classified as five-star data.

Clearly, the infrastructure is important, although, for scientific data, the access, and reuse do not only depend on the repository performance, but on formal characteristics associated with datasets and processes related to their production. In this manner, it was shown that both platforms analyzed (DSpace and Dataverse) are in accordance with the FAIR subprinciples investigated; therefore, they are appropriate to scientific data management and curation. However, it is necessary to pay attention to the objective and institutional policy during a scientific data repository implementation. An organization that has an implemented institutional (bibliographic) repository and that holds few resources to invest in a new project (as it would be if Dataverse were chosen) can opt to adapt DSpace to scientific data management and curation, with no great losses. DSpace would allow greater control of documents and data together, besides making it easier to link them. On the other hand, the Dataverse option would bring a platform aimed exclusively at data management and curation, besides enabling a wider visibility of institutional deposits once its infrastructure allows scientific data sharing among higher education and research institutions worldwide.

REFERENCES

AMORIM, R. C. *et al*. A comparison of research data management platforms: architecture, flexible metadata and interoperability. **Universal Access in the Information Society**, Berlin, v. 16, p. 851-862, June 2016. DOI: 10.1007/s10209-016-0475-y. Available from: https://link.springer.com/article/10.1007/s10209-016-0475-y. Access on: 2 July 2020.

BEAUJARDIÈRE, J. de la. NOAA Environmental Data Management. **Journal of Map & Geography Libraries**, [*S. l.*], v. 12, n. 1, p. 5-27, Mar. 2016. DOI: 10.1080/15420353.2015.1087446. Available from: https://www.tandfonline.com/doi/abs/10.1080/15420353.2015.1087446?tab=permissions&scroll=top. Access on: 13 July 2020.

BROWNLEE, R. Research data and repository metadata: policy and technical issues at the University of Sydney Library. **Cataloging & Classification Quarterly**, [*S. l.*], v. 47, n. 3-4, p. 370-379, Apr. 2009. DOI: https://doi.org/10.1080/01639370802714182. Available from: https://www.tandfonline.com/doi/abs/10.1080/01639370802714182. Access on: 18 June 2020.

CORPORATION FOR NATIONAL RESEARCH INITIATIVES. **HANDLE.NET (version 9) Technical Manual**. Reston, Virginia, 2018. Available from: http://www.handle.net/tech_manual/HN_Tech_Manual_9.pdf. Access on: 18 Oct. 2020.

CRESWELL, J. W. **Research design**: qualitative, quantitative, and mixed methods approaches. 4. ed. Los Angeles: Sage, 2014. 340 p.

CROSAS, M. **The FAIR Guiding Principles**: implementation in Dataverse. [*S. l.*], [2019?]. Available from: https://scholar.harvard.edu/files/mercecrosas/files/fairdata-dataverse-mercecrosas.pdf. Access on: 16 Oct. 2020.

DATAVERSE PROJECT. **Dataset + File Management**. [*S. l.*], 3 July 2024. Available from: https://guides.dataverse.org/en/latest/user/dataset-management.html. Access on: 19 Oct. 2020.

DOORN, P.; TJALSMA, H. Introduction: archiving research data. **Archival Science**, Netherlands, v. 7, p. 1-20, Sept. 2007. DOI: 10.1007/s10502-007-9054-6. Available from: https://link.springer.com/content/pdf/10.1007/s10502-007-9054-6.pdf. Access on: 21 Apr. 2019.

DURASPACE. **Technical Specifications**: DSpace. Beaverton, OR, 2020. Available from: https://duraspace.org/wp-content/uploads/dspace-files/specsh_dspace.pdf. Access on: 18 Oct. 2020.

EUROPEAN COMMISSION. **Turning FAIR into reality**: final report and action plan from the European Commission Expert Group on FAIR Data. Luxembourg: European Commission, 2018. Available from: https://op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1. Access on: 19 Oct. 2020.

FORCE, M. M.; AULD, D. M. Data Citation Index: promoting attribution, use and discovery of research data. **Information Services & Use**, [*S. l.*], v. 34, n. 1-2, p. 97-98, Jan. 2014. DOI: 10.3233/ISU-140737. Available from: https://content.iospress.com/download/information-services-and-use/isu737?id=information-services-and-use%2Fisu737. Access on: 19 June 2020.

GARNETT, A. *et al*. Open metadata for research data discovery in Canada. **Journal of Library Metadata**, [*S. l.*], v. 17, n. 3-4, p. 201-217, Mar. 2018. DOI: https://doi.org/10.1080/19386389.2018.1443698. Available from: https://www.tandfonline.com/doi/full/10.1080/19386389.2018.1443698. Access on: 15 July 2020.

GO FAIR. **FAIR principles**. [*S. l.*], [2016?]. Available from: https://www.go-fair.org/fair-principles/. Access on: 16 Oct. 2020.

INTERNATIONAL DATA CORPORATION. [*S. l.*], 2024. Available from: https://www.idc.com. Access on: 4 Oct. 2020.

LEE, D. J.; STVILIA, B. Developing data identifier taxonomy. **Cataloging & Classification Quarterly**, London, v. 52, n. 3, p. 303-336, Mar. 2014. DOI: https://doi.org/10.1080/01639374.2014.880166. Available from: https://www.tandfonline.com/doi/abs/10.1080/01639374.2014.880166. Access on: 19 June 2020.

RED HAT. Fedora. [*S. l.*], [2020?]. Available from: https://getfedora.org/. Access on: 3 Mar. 2021.

ROCHA, R. P. da (coord.). **Acesso aberto a dados de pesquisa no Brasil**: soluções tecnológicas: relatório 2018. Porto Alegre, RS: UFRGS, 2018. Available from: https://lume.ufrgs.br/handle/10183/185126. Access on: 19 Oct. 2020.

SCHOPFEL, J. *et al*. Open access to research data in electronic theses and dissertations: an overview. **Library Hi Tech**, [*S. l.*], v. 32, n. 4, p. 612-627, Nov. 2014. DOI: 10.1108/LHT-06-2014-0058. Available from: https://www.emerald.com/insight/content/doi/10.1108/LHT-06-2014-0058/full/html. Access on: 25 June 2020.

UNESCO. **The role of libraries in the promotion of open access**. Paris: UNESCO, 2014. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000231556. Acess on: 8 Out. 2024.

How to cite this chapter: SILVA, Fabiano Couto Corrêa da; RODRIGUES, Marcello Mundim. Implementation of FAIR principles in scientific data repositories: a comparative analysis of DSpace and Dataverse software infrastructures. *In*: SALES, Luana Faria; VEIGA, Viviane Santos de Oliveira; VIDOTTI, Silvana Aparecida Borsetti Gregório; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **FAIR Principles Applied To Research Data Management**: Brazilian Experiences. Brasília, DF: Editora Ibict, 2024. cap. 8, p. 103-112. DOI: 10.22477/9788570131959.cap8.