2. USING THE DATAVERSE PROJECT TO MOVE TOWARDS FAIR PRINCIPLES

[70]*Laura Vilela Rodrigues Rezende*
[71]*Sonia Barbosa*

2.1 INTRODUCTION

Over time, the scientific context has been changing with the increase in initiatives in favor of the opening of Science and consequently the strengthening of sharing and collaboration. In this opening scenario, there is contextualized research data, which makes it possible to confirm evidence from scientific studies. For this work, we will have as a basic premise the importance of sharing research data, which often have considerable potential for use, reuse, and reinterpretation in different studies beyond the possibilities of reproduction. However, there are several challenges faced by the actors involved in opening and sharing scientific data. Among them, it is possible to list: difficulty of data interoperability; difficulty in locating scattered and disorganized data; high degradation rate of data links (supplementary), among others.

> There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly concerned with long-term data stewardship; and a data science community mining, integrating and analyzing new and existing data to advance discovery (Wilkinson *et al.*, 2016, p. 1-2).

Faced with these challenges, several initiatives are underway to facilitate the opening and sharing of research data. In 2016, stakeholders representing academia, industry, funding agencies, and scholarly publishers designed and endorsed the FAIR Data Principles, which may act as a guideline for those wishing to enhance the reusability of their data holdings (Wilkinson *et al.*, 2016). For this purpose, the data must be Findable (metadata and data should be easy to find for both humans and computers), Accessible (users need to know how they can access the data, possibly including authentication and authorization), Interoperable (the data need to interoperate with applications or workflows) and Reusable (metadata and data should be well-described so that they can be replicated and/or combined in different settings) (GO FAIR, 2020).

This paper aims to discuss how managing data using the Dataverse tool facilitates moving data towards FAIR principles by presenting five examples of data shared in the Harvard Dataverse (HD) repository. First, we will present the conceptual approach of the research data repository and the Dataverse Project; In the following topic the cases are presented and finally some conclusive analysis.

70 PhD in Information Science. Federal University of Goiás. lauravil.rr@gmail.com.

71 BA Psychology & African American Studies/BSN Harvard University. sbarbosa@g.harvard.edu.

2.2 RESEARCH DATA REPOSITORY: THE DATAVERSE PROJECT

The data management process consists of a set of practices that benefit current research project stakeholders (researcher, funding agencies, research institutions, among others) once it makes it possible to recover and share data for future research ensuring their integrity, reproducibility, and replicability. The process of management occurs at all phases of the research cycle, since the planning for data management, before the project begins; the documenting, organizing and securing data during the project; and finally archiving data after the research is completed. Despite the discipline-specific set of knowledge, practices, and skills related to research data lifecycle activities (collecting, creating, manipulating, analyzing, and sharing data), it is important to consider the research data repositories aiming to provide the sustainable infrastructure for the long term storage and access to research data.

The Research Data Repository is a database infrastructure set up to manage, share, access, and file well-described and well-documented research data. These databases may be specialized to aggregating disciplinary or more general data, collecting over larger knowledge areas.

The research data repository may provide all these resources listed in Figure 1, improving the storing and sharing process. Besides general information and services, it must follow international standards related to technical aspects and metadata, aiming to guarantee findability and interoperability basically. It must offer clear terms and conditions that meet legal requirements related to data protection, allowing use and reuse without unnecessary licensing conditions. These aspects provide achieving quality standards in the management and preservation of data.

Figure 1 - The many planning aspects involved in the research data repositories

Source: Uzwyshyn (2016).

According to the Registry of Research Data Repositories (Re3data.org)[72], among some software available for data repositories, the most used are Dataverse[73], developed to store and share research data, DSpace[74], initially created for institutional repositories, CKAN[75], which was initially developed to promote the opening of government data, EPrints[76], developed to research data. This study is part of an investigation carried out by the digital curation team responsible for the development of Dataverse, at Harvard University, which is why this software was chosen.

Based on the 2018 report – *Acesso Aberto a Dados de Pesquisa no Brasil: Soluções Tecnológicas* (Rocha, 2018) that the main reasons that make Dataverse the most used research data storage and sharing software, among other features, is that it has easily configurable resources for defining various types of environments and different characteristics for repositories, including different organizational hierarchies and management of policies for units or groups, various metadata and license schemes.

The Dataverse software was the brainchild of Dr. Gary King, Faculty Director of Institute for Quantitative Social Science (IQSS) at Harvard University, to bring research data to the community and to make data FAIR, especially data that comes with a scientific claim in related publications (King, 2007).

> A Dataverse repository is the software installation, which then hosts multiple virtual archives called Dataverses collections. Each Dataverse collection contains datasets [and may also contain other dataverses], and each dataset contains descriptive metadata and data files (including documentation and code that accompany the data) (The Dataverse Project, 2020).

The Dataverse software is now in release 5.0 and continues to improve on its support of the FAIR principles, particularly in providing support for persistent identifiers (at the dataset and file level), with Uniform Resource Locator (URL), and metadata registered to DataCite, customizable metadata (including support for multiple standards) exportable in numerous formats, versioning for datasets and files, deaccessioning of datasets (and versions of datasets), linked data support, data access and use terms, and file conversion to reusable formats[77]. It is important to note that while the Dataverse software helps to move data towards FAIR, data authors and collection managers must contribute to this goal by using appropriate community metadata and vocabulary standards.

In the next section, we present five collections of the HD repository to represent resources related to the FAIR principles served by the Dataverse software and their respective collection curation team.

72 For more details, see: https://www.re3data.org/metrics/software. Access on: Oct. 3, 2024.

73 For more details, see: https://dataverse.org/. Access on: Oct. 3, 2024.

74 For more details, see: https://duraspace.org/dspace/. Access on: Oct. 3, 2024.

75 For more details, see: https://ckan.org/. Access on: Oct. 3, 2024.

76 For more details, see: https://www.eprints.org/uk/. Access on: Oct. 3, 2024.

77 For more details, there is a Dataverse metadata standard page available from: https://guides.dataverse.org/en/latest/user/appendix.html. Access on: Oct. 3, 2024.

2.3 DATA SHARED IN HD REPOSITORY

Since the FAIR principles do not prioritize orienting issues related to data quality, but rather enhance their sharing, it must first be understood that making data aligned with these principles is a continuous process that requires, in addition to aligned technological aspects, considerable time, energy, and expertise of those involved. The work of managing the collections´ data is essential in the process of alignment with the FAIR principles. With this in mind, the examples that will be presented below bring not only the technological resources implemented by default in the Dataverse software, but also some additional resources, policies, and workflows adopted that also increase and favor the data sharing process guided by the FAIR principles.

2.3.1 Methodological description

To carry out an analysis of the characteristics of some HD collections related to functions implemented by the software and best practices in data management actions, we sought to choose different segments that could represent different institutional and data generation contexts. One collection of: an organization/institution, a scientific journal, a University Department, an individual researcher, and a research group.

Regarding the analysis of the resources offered by the Dataverse software and the curation work carried out by the collection managers aligned with FAIR principles, the reference study chosen that presents the necessary interpretations and considerations was that of Jacobsen *et al*. (2020). The authors presented the opinions of the original creators of the principles, supported by discussions of the experiences of pioneering FAIR implementers. They also pointed out the importance of presenting a common understanding around the original intentions of the guiding principles, aiming to avoid divergence into non-interoperability.

2.3.2 The principle of "Findability"

The principle of findability, with its sub categories, are related to supporting users in their discovery process. It is considered the most fundamental of the FAIR principles, as globally unique and persistent identifiers are essential elements providing unambiguous identification of resources. In addition, this principle also contemplates facets of search, keywords, and templates from the communities that facilitate capturing uniform and harmonized metadata.

The Dataverse citation resource, metadata tab of the dataset and files contain registered Digital Object Identifier (DOI) and Message-Digest Algorithm (MD5) (UNF for tabular files) code. In addition, the software provides search facets, keywords, and templates as resources related to discoverability. These are considered good practice in FAIR once the resource and its metadata are persistently linked, and these identifiers may then successfully be used as the search term to discover its metadata record. Dataverse is also committed to using standard-compliant metadata to ensure that collections´ metadata can be easily mapped to standards' schemas and exported in format for preservation and interoperability[78].

78 For more details, there is a Dataverse Metadata Crosswalk available at: https://docs.google.com/spreadsheets/d/10Luz-ti7svVTVKTA-px27oq3RxCUM-QbiTkm8iMd5C54/edit#gid=222839033. Access on: Oct. 3, 2024.

In the HD repository, the "Citation Metadata" element is the only required metadata block. This metadata element has five required fields for all datasets: "title and author name" are used to build the citation, and "e-mail contact, description, and subject" are used to enrich the dataset metadata and lend to the discoverability of content on the HD. Harvard Dataverse also supports DOIs at the file level. Collections created within the HD can utilize customization by selecting additional metadata elements to support their data. The Table 1 details the additional steps taken by the five examples in this study to enhance "findability" of their data, beyond the default software features.

Table 1 - "FINDABLE" FAIR Principle in selected HD Collections[79]

	F1: unique and persistent identifier	F2: data are described with rich metadata	F3: metadata clearly and explicitly include the identifier of the data it describes	F4: (meta) data are registered or indexed in searchable resource
Organization: The International Food Policy Research Institute (IFPRI)[80]	YES (software implemented at dataset and file level)	Uses multiple metadata blocks; Uses optional metadata fields; Links to keyword and topic classification standard vocabulary; "widget" feature; "file tags"; additional metadata "terms".	When available - "Related publication" metadata field to connect to journal articles (bidirectional link)	YES (Software implemented)
Journal: American Journal of Political Science (AJPS)[81]		Uses multiple metadata blocks, including "journal metadata block" and "related publication" metadata field; uses optional metadata fields; "file tags"; "widget" feature;	Always - "Related publication" metadata field to connect to journal article (bidirectional link)	
Department: Harvard University Department of Government[82]		Uses multiple metadata blocks; uses optional metadata fields; "file tags"; "widget" feature; uses additional metadata terms;	When available - "Related publication" metadata field to connect to journal articles (bidirectional link)	
Researcher: Gary King[83]		Uses multiple metadata blocks; uses optional metadata fields; "widget" feature; "file tags".	When available - "Related publication" metadata field to connect to journal articles (bidirectional link). Links to replication software utilized by the dataset.	

79 The features implemented by the collections listed in the table are described in: https://dataverse.org/software-features. Access on: Oct. 3, 2024.

80 The IFPRI dataverse is available from: https://dataverse.harvard.edu/dataverse/IFPRI. Access on: Oct. 3, 2024.

81 The AJPS dataverse is available from: https://dataverse.harvard.edu/dataverse/ajps. Access on: Oct. 3, 2024.

82 The Harvard University Department of Government dataverse is available from: https://dataverse.harvard.edu/dataverse/GovDept. Access on: Oct. 3, 2024.

83 Gary King dataverse is available from: https://gking.harvard.edu/data. Access on: Oct. 3, 2024.

	F1: unique and persistent identifier	F2: data are described with rich metadata	F3: metadata clearly and explicitly include the identifier of the data it describes	F4: (meta) data are registered or indexed in searchable resource
Research Group: Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL)[84]	YES (software implemented at dataset and file level)	Uses multiple metadata blocks; uses optional metadata fields; uses additional metadata "terms"; "file tags" .	When available - "Related publication" metadata field to connect to journal articles (bidirectional link)	YES (Software implemented)

Source: self-elaboration (2020).

2.3.3 The principle of "Accessibility"

The Dataverse software supports the principle of Accessibility with support for full dataset citations, DOIs with URLs, and metadata registered to Data Cite[85]. Citation and discoverable metadata are available using several standards, including Schema.org, Dublin Core, and Data Documentation Initiative (DDI). Data files can be restricted (authentication/authorization required) or open for access. There is a Terms landing page with metadata for usage information. There is a citation for each data file, with a DOI and URL for each file. Downloads of the metadata include machine-actionable dataset landing pages with meta-tags for citation metadata. The Deaccession feature allows removal of a dataset and leaves a "tombstone" citation page which is findable and citable; metadata includes reason for "deaccessioning" / "Versioning." There is also Support for the web Hypertext Transfer Protocol (HTTP) (W3C); the data transfer protocol with mirroring, incremental backups, and file copies between systems: Rsync over SSH (GNU GPL); REpresentational State Transfer (RESTful) API; Authentication API Tokens; Authorization service.

Harvard Dataverse uses CC0[86] license by default, and allows depositors to opt out and use their license of choice. Depositors can choose whether their data are open or restricted for access, but in the latter case they must enable the "request access" feature for data requestors, or provide terms describing how users can request access to restricted content, or if content is embargoed for a period of time. Following DataCite standards, all metadata for datasets are visible and discoverable, even if files are not immediately downloadable for access. The examples in the table below include open data, embargoed content, and content that requires additional contact with the data owners for access. The Table 2 details the additional steps taken by the five examples in this study to enhance "accessibility" of their data, beyond the default software features.

84 The SIIL dataverse is available from: https://dataverse.harvard.edu/dataverse/SIIL. Access on: Oct. 3, 2024.

85 For more details about Data Cite: https://datacite.org/. Access on: Oct. 3, 2024.

86 The HD licenses and terms of use are described in: https://dataverse.org/best-practices/harvard-dataverse-general-terms-use.

Table 2 - "ACCESSIBLE" FAIR Principle in selected HD Collections[87]

	A1: (meta)data are retrievable by their identifier using standardized communications protocol	sub-principle A1.1: the protocol is open, free and universally implementable	sub-principle 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
Organization: The International Food Policy Research Institute (IFPRI)	YES (software implemented)		"File restriction" feature; "request access" feature.	YES (software implemented)
Journal: American Journal of Political Science (AJPS)			Restricted content provides copyright info and access information provided.	
Department: Harvard University Department of Government			"File restriction" feature, with embargo.	
Researcher: Gary King			Restricted content provides copyright info and access information provided.	
Research Group: Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL)			"File restriction" feature; "request access" feature.	

Source: self-elaboration (2020).

2.3.4 The principle of "Interoperability"

The Dataverse software supports the principle of "Interoperable" by supporting variable metadata for tabular data files, using DDI standard, Machine-actionable Variable description from DDI, and summary statistics in DDI automatically calculated upon data upload.

87 The features implemented by the collections listed in the table are described in: https://dataverse.org/software-features. Access on: Oct. 3, 2024.

Harvard Dataverse integrated the Data Explorer[88] tool developed by Scholars Portal. Data Explorer is a Graphical User Interface (GUI) which lists the variables in a tabular data file allowing searching, charting and cross tabulation analysis. Every example in our table below utilizes the tabular data functionality where possible. The HD also uses the File Previewer tool, a set of tools that display the content of files - including audio, html, annotations, images, Portable Document Format (PDF), text, video, tabular data, and spreadsheets - allowing them to be viewed without downloading. The Table 3 details the additional steps taken by the five examples in this study to enhance "interoperability" of their data, beyond the default software features.

Table 3 - "INTEROPERABLE" FAIR Principle in selected HD Collections[89]

	I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation		I2: (meta) data use vocabularies that follow fair principles *complete	I3: (meta)data include qualified references to other (meta) data
Organization: The International Food Policy Research Institute (IFPRI)	YES (software implemented) and integrated tools (Data explorer, File Previewer)	Of 10k files, 7k are tabular files	"keywords," & "Topic Classification" controlled vocabulary w/links to standards http://aims.fao.org/	When available - "Related publication" metadata field to connect to journal articles (bidirectional)
Journal: American Journal of Political Science (AJPS)		Of 8500k files, 2200 are tabular files *note this is a replication data journal so each dataset normally contains one data file, and one code file, and one readme file	Uses multiple metadata blocks, including "journal metadata block" and "related publication" metadata field; uses optional metadata fields; "file tags" ; "widget" feature; uses Center for Open Science "Open Materials and Open Data" badges.[90]	Always - "Related publication" metadata field to connect to journal articles (bidirectional)
Department: Harvard University Department of Government		Of 1350 files, 478 are tabular files	Uses multiple metadata blocks; uses optional metadata fields; "file tags"; "widget" feature; uses additional metadata terms;	When available - "Related publication" metadata field to connect to journal articles (bidirectional)
Researcher: Gary King		Of 1870 files, 563 are tabular files	Uses multiple metadata blocks; uses optional metadata fields; "widget" feature; "file tags".	When available - "Related publication" metadata field to connect to journal articles (bidirectional)

88 This feature is described here: https://guides.dataverse.org/en/latest/admin/external-tools.html

89 The features implemented by the collections listed in the table are described in: https://dataverse.org/software-features

90 This feature is to acknowledge open practice of the dataset: https://osf.io/tvyxz/wiki/home/. Access on: Oct. 3, 2024.

	I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation		I2: (meta) data use vocabularies that follow fair principles *complete	I3: (meta)data include qualified references to other (meta) data
Research Group: Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL)	YES (software implemented) and integrated tools (Data explorer, File Previewer)	Of 1280 files, 287 are tabular files	Uses multiple metadata blocks; uses optional metadata fields; uses additional metadata "terms"; "file tags".	Links to associated manuscripts where possible; uses additional metadata "file tags"

Source: self-elaboration (2020).

2.3.5 The principle of "Reusability"

The Dataverse software supports the principle of Reusability by supporting the integration of Make Data Count[91]. Citation and discoverable metadata using DataCite, Schema.org, Dublin Core, DDI standards. Additional metadata support, including domain specific. Terms with license usage or data use agreement. PROV metadata (provenance). Domain relevant file download standards. Variable metadata for tabular data files using DDI standards, machine actionable variable descriptions from DDI, summary statistics in DDI, automatically calculated upon data upload.

Harvard Dataverse makes use of the Make Data Count integration. Provenance information is requested at the dataset level. The use of the Data Explorer tool allows for analysis and visualization of tabular data files. The Table 4 details the additional steps taken by the five examples in this study to enhance Reusability of their data, beyond the default software features.

91 For more details see: https://makedatacount.org/. Access on: Oct. 3, 2024.

Table 4 - "REUSABLE" FAIR Principle in selected HD Collections

	(meta)data are richly described with a plurality of accurate and relevant attributes	metadata are released with a clear and accessible data usage license	(meta)data are associated with detailed provenance	r1.3: (meta) data meet domain relevant community standards
Organization: The International Food Policy Research Institute (IFPRI)	Metadata verification via established workflows; Use of dataverse templates to ensure consistency of metadata; lengthy dataset descriptions, use of additional citation metadata fields, including: more than one "keyword," more than one "topic classifications," and "social science" and "geospatial" metadata, " in addition, linking to standard agricultural standard vocabulary.	Public facing additional License and term of access / Data sharing agreement / Open Access and Open Data Policy / Donors policy.[92]	**Citation metadata block,** in addition: "grant information; distributor information; dates metadata, "contributors", "software," "series," "related publication and datasets," "data collectors," "data source." **Geospatial metadata block:** coverage country/nation; coverage state/province; coverage city; unit. **Social Science and Humanities metadata block:** "universe;" "unit of analysis;" "sampling procedure;" "collection mode;" "type of research instrument." **Use of templates** for consistency in required metadata fields and formatting of information in such fields.	Metadata Blocks used: Citation; Geospatial; Social Science; Additional metadata: Dataverse and Datasets description; Summary of collection content; Links to relevant documentation; search facets; templates

92 These licenses are available from: http://ebrary.ifpri.org/utils/getfile/collection/p15738coll2/id/133521/filename/133732.pdf. Access on: Oct. 3, 2024.

	(meta)data are richly described with a plurality of accurate and relevant attributes	metadata are released with a clear and accessible data usage license	(meta)data are associated with detailed provenance	r1.3: (meta) data meet domain relevant community standards
Journal: American Journal of Political Science (AJPS)	Metadata verification via established workflows; Use of dataverse templates to ensure consistency of metadata; lengthy dataset descriptions, use of additional citation metadata fields, including: more than one "keyword;" reproducibility verification workflow	Public facing verification police document, CC0 by default[93]; open to allow authors to use other licenses as needed; restricted content is clearly labeled with copyright and access information	**Citation metadata block,** in addition: "dates" metadata, "related publication and datasets; **Social Science metadata block; Geospatial metadata block: Journal Metadata Block;** Use of templates for consistency in required metadata fields and formatting of information in such fields.	Metadata Blocks used: Citation; Journal; Geospatial; Social Science; Additional metadata: Dataverse and Datasets description; Summary of collection content; Links to relevant documentation; search facets; templates
Department: Harvard University Department of Government	Metadata verification via established workflows; Use of dataverse templates to ensure consistency of metadata; lengthy dataset descriptions, use of additional citation metadata fields, including: more than one "keyword," "topic classifications, "kind of data","related materials"; "related dataset".	Files embargoed with date of release; resolves to CC0 once released; Data PASS Terms standard 1.0[94]	Citation metadata block; in addition: producer and distributor information; "dates" metadata, "related publication and datasets, Geospatial metadata block: "geographic coverage"; Social Science and Humanities metadata block: "unit of analysis;" templates	Metadata Blocks used: Citation; Social Science; Geospatial; Additional metadata: Dataverse and Datasets description; Summary of collection content; search facets; templates; links to relevant documentation

93 For details see: https://ajps.org/ajps-verification-policy/ ; https://creativecommons.org/publicdomain/zero/1.0/ ; https://guides.dataverse.org/en/5.1.1/user/dataset-management.html?highlight=cc0. Access on: Oct. 3, 2024.

94 For details see: http://www.data-pass.org/sites/default/files/metadata.pdf. Access on: Oct. 3, 2024.

	(meta)data are richly described with a plurality of accurate and relevant attributes	metadata are released with a clear and accessible data usage license	(meta)data are associated with detailed provenance	r1.3: (meta) data meet domain relevant community standards
Researcher: Gary King	Use of dataverse templates to ensure consistency of metadata; lengthy dataset description, use of additional citation metadata fields, including: more than one "keyword," more than one "topic classification."	CC0 by default; restricted content is clearly labeled with copyright and access information	Citation metadata block; Use of templates	Metadata block used: Citation. Additional metadata: Dataverse and Datasets descriptions; Summary of collection content; search facets; templates
Research Group: Feed the Future Innovation Lab for Collaborative Research on Sustainable Intensification (SIIL)	Metadata verification via established workflows; use of dataverse templates to ensure consistency of included metadata; lengthy dataset descriptions, use of additional citation metadata fields, including: "keyword," and Social Science and Humanities, Geospatial,Life Sciences, and Journal metadata blocks	Default CC0 waived with SIIL terms of use clearly defined	Citation metadata block; Use of templates	Metadata block used: Citation, Social Science and Humanities, Geospatial, Life Sciences, and Journal. Additional metadata: Dataverse and Datasets descriptions; summary of collection content; search facets; templates

Source: self-elaboration (2020).

2.4 FINAL CONSIDERATIONS

It is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects — from data to analytical pipelines — benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability (Wilkinson *et al*., 2016). Important to note is that while the software provides functionality to help move data towards FAIR, data authors, collection managers and curators may contribute to this effort by utilizing the features provided and using best practices when sharing their data. The five examples in this paper detail the use of features to move their collections towards FAIR, and demonstrate not only the differences between the collections' utilization of workflows and tools, but the impact of such use on FAIR.

The International Food Policy Research Institute (IFPRI) and Feed the Future dataverses demonstrate the effectiveness of planned workflows and use of templates to guide large teams of curators, and organization level terms of access. IFPRI, in linking to their standards, demonstrates the use of optional, but encouraged, standards and features. Both utilize a team of curators and data managers, and make use of the Featured Dataverses option to highlight their collections, and use additional metadata blocks offered by the software to describe their data, and

extensive searchable metadata facets to improve the Findability of their datasets. They also make use of multiple Dataverse templates prepopulated with the appropriate terms of use for each collection, saving curators and data managers the time needed to complete this information for each dataset. Of the five collections, IFPRI and Feed the Future utilize the request access feature for restricted files, which allows them access control. The request access works in alignment with the terms of access to fulfill one sub principle of "accessibility".

The *American Journal of Political Science* (AJPS) and Gary King dataverses are cases that support replication verification in data sharing. The AJPS dataverse utilizes the "submit for review" workflow to verify reproducibility of data before data publishing. This process is supported by the National Academies Press (NAP) (2019, p. 3) statement that, "journals should consider ways to ensure computational reproducibility for publications that make claims based on computations, to the extent ethically and legally possible". In addition, AJPS demonstrates the desired level of curation and workflow to achieve and allow others to reproduce or replicate their results, as recommended by NAP that, "journalists should report on scientific results with as much context and nuance as the medium allows" (NAP, 2019, p. 4). The Harvard Dataverse is a medium that allows rich reporting of scientific results via the Dataverse features and best practices guidelines. The journal's use of Open Badges is an additional acknowledgment to the authors for depositing content that is verified reusable. Gary King's dataverse, self-curated with numerous replication data supported by author confirmed reproducibility verification, are designated "replication" datasets that include data, code, documentation, and links to software[95] that allow maximum reuse of the data. Replicability of data, as demonstrated by the two collections, is aligned with the sub principle of Reusability associated with detailed provenance.

The Harvard Department of Government Dissertation Dataverse was selected as a unique case supporting the early engagement of graduate students in data sharing. All students in this department are required to share their dissertation data within Harvard Dataverse to fulfill their graduation requirement. The space was designed by the Harvard Curation team utilizing templates with prefilled metadata fields, and instructions for data deposits. The terms of use section of the template is prefilled with a 5-year embargo period to give graduates sufficient time to publish on their dissertation research, before the data becoming open access. This case supports the introduction of early data sharing incentives and guidelines for graduate students, demonstrating the different levels of open access. The embargo allows the support of Findability, Accessibility, and Reusability because the dataset metadata remains visible to the research community and the Terms of Access detail when data will become available for public consumption.

This paper richly demonstrates how the Dataverse Software, individual installation workflows, and additional curation and data management by data depositors, can enhance the FAIR principles in Dataverse repositories. We demonstrate the vast diversity in Dataverse data sharing options that support FAIR, and provide examples of opportunities to educate researchers in the FAIR data sharing process. While the software provides functionality to move data towards FAIR, the researcher, data manager, and curator's use of the software features is what allows maximum Findability, Accessibility, Interoperability, and Reusability of the shared research content.

95 For details see: https://gking.harvard.edu/software. Access on: Oct. 3, 2024.

APPENDIX - Glossary (terms definitions used in this paper)

Bidirectional linking (via related publications' metadata field): The dataset metadata field used to link to the related article that supports the data.

Dataset: a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.

Dataverse: Dataverse is an open-source web application to share, preserve, cite, explore, and analyze research data.

Deaccessioning (tombstone page): the process of removing a published dataset for legal or valid reasons. This always results in a tombstone landing page with the basic citation metadata always accessible to the public if they use the persistent URL (Handle or DOI) provided in the citation for that dataset. Users will not be able to see any of the files or additional metadata that were previously available prior to deaccession.

Facets: metadata fields to use as facets for browsing datasets and dataverses in this dataverse.

File: A data file is a computer file which stores data to be used by a computer application or system, including input and output data (Wikipedia).

Guestbook: GuestBooks allow you to collect data about who is downloading the files from your datasets.

Make Data Count: is a project to collect and standardize metrics on data use, especially views, downloads, and citations. Dataverse can integrate Make Data Count to collect and display usage metrics including counts of dataset views, file downloads, and dataset citations.

Metadata blocks: metadata based on standards , shipped with Dataverse (e.g., DDI for social science) and you can learn more about these standards in the Appendix section of the User Guide.

Publishing on Dataverse: When you publish a dataset, you make it available to the public so that other users can browse or search for it using the DOI/Handle or metadata.

Tabular Data: dataverse software extracts the data content from the user's tab files and archive it in an application-neutral, easily readable format.[96]

Versioning: Versioning is important for long-term research data management where metadata and/or files are updated over time. It is used to track any metadata or file changes (e.g., by uploading a new file, changing file metadata, adding or editing metadata) once you have published your dataset.

96 The supported formats are listed here: https://guides.dataverse.org/en/5.1.1/user/tabulardataingest/supportedformats.html

Widget: The Widgets feature provides you with code for your personal website, so your dataset can be displayed. There are two types of Widgets for a dataset: Dataset Widget and the Dataset Citation Widget.

Templates: Templates are useful when you have several datasets that have the same information in multiple metadata fields that you would prefer not to have to keep manually typing in, or if you want to use a custom set of Terms of Use and Access for multiple datasets in a dataverse.

REFERENCES

GO FAIR. **FAIR Principles**. Available from: https://www.go-fair.org/fair-principles/. Access on: 23 set. 2020.

KING, G. An introduction to the Dataverse Network as an infrastructure for data sharing. **Sociological Methods & Research**, [*s. l.*], v. 36, n. 2, p. 173-199, Nov. 2007. DOI: https://doi.org/10.1177/0049124107306660. Available from: https://journals.sagepub.com/doi/abs/10.1177/0049124107306660. Access on: Mar. 4, 2021.

NATIONAL ACADEMIES PRESS. Reproducibility and replicability in science. **The National Academies of Science, Engineering, and Medicine**. [*S. l.*]: NAP, 2019. Available from: https://nap.nationalacademies.org/resource/25303/R&R.pdf. Access on: Oct. 10, 2020.

ROCHA, R. P. (coord.). **Acesso aberto a dados de pesquisa no Brasil**: soluções tecnológicas: relatório 2018. Porto Alegre, RS: UFRGS, 2018. 75 p. Available from: http://hdl.handle.net/10183/185126. Access on: Feb. 27, 2021.

THE DATAVERSE PROJECT. **About The Project**. Available from: https://dataverse.org/about. Access on: Sept. 30, 2020.

UZWYSHYN, R. Research Data Repositories: the what, when, why, and how. **Computers in Libraries**, [*s. l.*], v. 36, n. 3, Apr. 2016. Available from: https://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml. Access on: Oct. 3, 2024.

WILKINSON, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, [*s. l.*], v. 3, article number 160018, p. 1-9, 2016. DOI: https://doi.org/10.1038/sdata.2016.18. Available from: https://www.nature.com/articles/sdata201618. Access on: Oct. 3, 2024.

How to cite this chapter: REZENDE, Laura Vilela Rodrigues; BARBOSA, Sonia. Using the DATAVERSE Project to move towards FAIR principles. *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. 2, p. 30-45. DOI: 10.22477/9788570131959.cap2.