FAIR PRINCIPLES APPLIED TO RESEARCH DATA MANAGEMENT

BRAZILIAN EXPERIENCES

2024

ORGANIZERS

LUANA FARIA SALES
VIVIANE SANTOS DE OLIVEIRA VEIGA
SILVANA A. B. GREGORIO VIDOTTI
PATRÍCIA HENNING
LUÍS FERNANDO SAYÃO

Editora Ibict

ORGANIZERS

Luana Faria Sales
Viviane Santos de Oliveira Veiga
Silvana Aparecida Borsetti Gregório Vidotti
Patrícia Henning
Luís Fernando Sayão

FAIR PRINCIPLES APPLIED TO RESEARCH DATA MANAGEMENT:
BRAZILIAN EXPERIENCES

Editora
Ibict

PRESIDÊNCIA DA REPÚBLICA

Luiz Inácio Lula da Silva
Presidente da República

Geraldo José Rodrigues Alckmin Filho
Vice-Presidente da República

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

Luciana Santos
Ministra da Ciência, Tecnologia e Inovação

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA

Tiago Emmanuel Nunes Braga
Diretor

Carlos André Amaral de Freitas
Coordenador de Administração - COADM

Ricardo Medeiros Pimenta
Coordenador de Ensino e Pesquisa em Informação para a Ciência e Tecnologia - COEPI

Henrique Denes Hilgenberg Fernandes
Coordenador de Planejamento, Acompanhamento e Avaliação - COPAV

Cecília Leite Oliveira
Coordenadora-Geral de Informação Tecnológica e Informação para a Sociedade - CGIT

Washington Luís Ribeiro de Carvalho Segundo
Coordenador-Geral de Informação Científica e Técnica - CGIC

Hugo Valadares Siqueira
Coordenador-Geral de Tecnologias de Informação e Informática - CGTI

BOOK SCIENTIFIC COUNCIL

BOOK ad hoc SCIENTIFIC COUNCIL

ORGANIZERS

Luana Faria Sales
Viviane Santos de Oliveira Veiga
Silvana Aparecida Borsetti Gregório Vidotti
Patrícia Henning
Luís Fernando Sayão

FAIR PRINCIPLES APPLIED TO RESEARCH DATA MANAGEMENT:
BRAZILIAN EXPERIENCES

Editora
Ibict

BRASÍLIA 2024

This publication has the cooperation of Graduate Program in Information Science at São Paulo State University (UNESP) within the scope of the Project "Expansion and Modernization of IBICT Actions related to the Activities of Collection, Storage, Systematization, Analysis, Dissemination and Preservation of Data and Information Relating to Science, Technology and Innovation" (Prodoc 914BRZ2005) and with the financial support of CAPES - Process #23038.005690/2021-02 - AUXPE/PROEX AGREEMENT No. 0415/2021 - CALL #01/2023 – Translation of articles or book chapters.

EQUIPE TÉCNICA
Organizers
Luana Faria Sales
Viviane Santos de Oliveira Veiga
Silvana Aparecida Borsetti Gregório Vidotti
Patrícia Henning
Luís Fernando Sayão

Projeto gráfico, capa e Diagramação
Rafael Fernandez Gomes

Dados Internacionais de Catalogação-na-Publicação (CIP)

F163 FAIR principles applied to research data management: brazilian experiences /
Organizer Luana Faria Sales... [*et al.*]. - - Brasília, DF: Editora Ibict, 2024.
1 recurso online [220 p.] : il.

220 p.: il. Color.
ISBN: 978-85-7013-195-9
DOI: 10.22477/9788570131959

1. FAIR principles. 2. Data management. I. Sales, Luana Farias. II. Title..

CDD 002:004

Ficha catalográfica elaborada por Fernanda Maciel Rufino – CRB 1/3064

Postgraduate Program in Information Science (PPGCI), developed by the Brazilian Institute of Information in Science and Technology, of the Ministry of Science and Technology and Innovation (IBICT/MCTI) in cooperation with the Communication School at the Federal University of Rio de Janeiro (ECO/UFRJ).

28 / 5.000

How to reference this book:

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. 220 p.

The opinions expressed in this publication are the sole and entire responsibility of the authors and do not necessarily express the point of view of the Brazilian Institute of Information in Science and Technology or the Ministry of Science, Technology and Innovation.

Address:

Ibict - Instituto Brasileiro de Informação em Ciência e Tecnologia

Setor de Autarquias Sul (SAUS), Quadra 05, Lote 06, Bloco H – 5o. andar

CEP: 70.070-912 - Brasília, DF

The research that resulted in this publication was developed by

CNPq
FAPERJ
Capes

& with the support of

IBICT
UNESP
UFRJ
FIOCRUZ

**For those,
who, like
us,
believe
that
science
makes the
world a
better
place to
live.**

The FAIR principles are not magic, nor are they presenting a panacea, but they guide the development of infrastructure and tooling to make all research objects optimally reusable for machines and people alike, which is a crucial step. It is essential that the community continue to discuss, challenge, and refine their implementation choices, within the behavioral guidelines established by the principles. (MONS, 2018)

TABLE OF CONTENTS

PREFACE

Since the emergence of the FAIR movement in January 2014 and the publication of its principles in March 2016, we can observe a significant growth in the worldwide interest in better handling data and other digital objects. This is so they are more easily findable, accessible, interoperable, and reusable. The principles were defined with the main objective of expressing a set of expected behaviors for digital objects to make them more susceptible to the performance of computer systems. We immediately saw several initiatives seeking options on how to create implementations that followed the principles. However, these initiatives tool place independently, without any coordination between them. To avoid wasting resources and reduce the risk of incompatible implementations, which would be contrary to FAIR's original intentions, the governments of the Netherlands and Germany. Initially, and later the government of France, committed to funding the establishment of an entity that could support and coordinate the development of solutions following the FAIR principles. Thus, the GO (*Global, Open*) FAIR movement was born, from the creation of the GO FAIR International Support and Coordination Office (GFISCO).

GFISCO has three headquarters, in Leiden, Netherlands, in Paris, France and in Hamburg, Germany. Naturally, being funded by three European governments, GFISCO's focus should be on European activities and initiatives related to FAIR in the context of the *European Open Science Cloud (EOSC)*. However, the EOSC benefits would only be maximized if they were not established as a European silo, but as the European branch of a global environment, based on FAIR principles. Thus, GFISCO also works to expand the GO FAIR movement internationally. The initial results of this effort were the creation of regional GO FAIR offices in Brazil and the United States.

This book brings a collection of articles that report work done in Brazil and supported by GO FAIR Brazil office. As we can see from the content, the GO FAIR movement has been growing significantly in the country and covers various subjects, ranging from the discussion of data management and research in light of FAIR principles. It also includes proposals for extensions in data repositories to comply with the requirements of the principles.

The COVID-19 pandemic, while bringing enormous problems to all countries in the world, has also served to clarify it that we still have a lot to do in the field of data management. It is increasingly clear that we could have an efficiency gain in almost Every aspect of the response to the pandemic. From the identification of growth in the contagion, to the analysis of the biological processes related to the infection of SARS-CoV2 and the eventual identification of suitable treatments. This global effort in the fight against COVID-19 is also reflected in this volume, with works directly and indirectly linked to the subject.

The effort to have a more FAIR world is just beginning, but if we have the volume and quality of the works presented in this book as indicatives, we can be optimistic that these goals will be effectively achieved. And we will have an environment where digital objects interact in a more efficient and transparent way.

Luiz Olavo Bonino da Silva Santos
International Technology Advisor - GO FAIR International Support and Coordination Office, the Netherlands

PRESENTATION

The present book is an attempt to bring together theoretical and empirical Brazilian initiatives regarding the application of FAIR principles and to serve as another instrument for dissemination of these principles in Brazil, especially in the field of Information and Computer Science research. Thus, after the prologue, we present a history of achievements of the GO-FAIR initiative in the world and in Brazil, from 2017 to 2020. The book is organized considering, in a first section, studies that take place under each of the categories in which FAIR principles are distributed, and in a second section, reports of empirical experiences within the scope of GO FAIR Brazil. These categories are not mutually exclusive, so some chapters could be categorized in more than one section or even all of them, after all, the book is about the application of FAIR principles. However, in an attempt to present a structured form of organization, we offer this organization to our readers.

Thus, regarding the location of data, the book initially presents the role of unique identification of author so that data can be found, through the standardization of authors' names. Next, in the same section, we chose to categorize the chapters that refer to the implementation of FAIR principles in repositories as we consider them an efficient tool to make data findable, either through APIs or through OPI-MH. Also in this section, we insert two chapters on empirical experiences that, although they did not take place within the scope of GO-FAIR Brazil, they fit perfectly within the initiatives aimed at facilitating the location of research data.

Regarding data accessibility, the "Accessible Data" section was used to cover the chapters that deal with the experiences of accessing and using open data in open access platforms and repositories, as well as an interesting theoretical proposal for the application of linked data in open research notebooks.

In the "Interoperable data" section, the book brings to light a study on standards and another one on the notion of data encapsulation to promote data interoperability between systems. As an empirical experience, the section also brings a study on interoperability between repositories of the National Institute of Health (NIH) health data center.

In the context of data reuse, we gathered three chapters focused on this theme. The first one highlights the importance of increasing data value, expanding its visibility, as well as the potential for reuse. For this purpose, the authors present a study on the meanings of the terms *use* and *reuse* and the conditions that are established for the reuse to be effective. The second chapter focuses on the need for data curation so that data is, in addition to FAIR, CARE, expanding the focus of data management to social, ethical and legal issues as well. In fact, FAIR principles cover only one side of data management, which can and should be complemented with principles that also focus on governance in specific domains, thus promoting the conscious reuse of data. The third chapter of this section addresses the creation of data management services to promote its reuse.

We finish the book with a section Where we highlight two relevant initiatives within the scope of GO FAIR Brazil: the implementation of GO FAIR Brazil Health-Nursing Network, and the last chapter presents a rich experience that emerged at the peak of the pandemic context of COVID-19 with the purpose of building, in an agile way, a federated infrastructure for an international, interoperable and distributed data network, offering support in the search for evidence-based answers about cases of viral outbreaks. The international initiative VODAN – acronym for *Virus Outbreak Data Network* (VODAN), was established in Brazil in the context of GO FAIR Health Network, using FAIR principles to manage data and metadata collected during the pandemic.

With these chapters, we hope to support students and researchers who need to venture into the wide world of knowledge necessary to manage research data with quality, offering a mix of theoretical and empirical content on the application of FAIR principles.

Luana Farias Sales
Silvana Aparecida Borsetti Gregorio Vidotti
Patrícia Henning
Luís Fernando Sayão

HOW IT ALL BEGAN?

After publication of the article *The FAIR Guiding Principles for scientific data management and stewardship*[1] in *Nature* magazine, FAIR principles, an acronym for *Findable, Accessible, Interoperable and Reusable*, were internationally recognized, assuming the role of world reference of good data management practices.These principles triggered concerns in the academic-scientific community when they were placed on the agenda of the *High-Level Expert Group on the* EOSC[2] created in 2016, by the European Commission.

This group aimed to present recommendations that would ensure that science, business, government and, eventually, the industry, could benefit from the big data revolution.Since then, several initiatives have gradually emerged aimed at the development of products and services that adopt FAIR principles in the practices of science. The GO FAIR[3] initiative was one of them, that emerged in 2016, autonomously, with bottom-up management model, that is, organized from bottom to top, by the community itself, aiming to disseminate and generate FAIR resources and services.

GO FAIR aims to ensure the proper reuse of data in different contexts, countries and disciplines, contributing to the sharing and reuse of data in the generation of new knowledge and for the reproducibility of research. This initiative was designed to act under three pillars, GO CHANGE: invests in dissemination actions seeking to influence cultural and political changes that make FAIR principles a reference in science; GO TRAIN: operates in training courses of different levels and natures, related to the application of FAIR principles in science practices; GO BUILD: drives the development of technical and operational infrastructure to support FAIR data.

To better expose the performance of the GO FAIR initiative throughout its four years of existence, this report presents an overview of the main achievements during this period, highlighting the Brazilian participation through the GO FAIR Brazil initiative.

[1] https://www.nature.com/articles/sdata201618

[2] https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf#view=fit&pagemode=none

[3] https://www.go-fair.org

2017: YEAR OF EXPECTATIONS AND BIG CHALLENGES

The year of 2017 was a critical year for the GO FAIR initiative to take off and gain worldwide recognition. Its policy of action was motivated and driven by the decisions of the European Commission[4], which began to demand that results of research financed by it, as well as their respective data, were available in open access and in line with FAIR principles. In addition, it was strengthened by do(G7) group composed of Canada, France, Germany, Italy, Japan, the United Kingdom and the United States, when they included FAIR principles in their open science guidelines[5]. Finally, it was legitimized when the governments of Germany and the Netherlands released EOSC positioning statement to support the GO FAIR initiative.[6]

Activities began centered on the GO CHANGE pillar, aimed at investing the generation of FAIR culture. The results began to appear and publications such as the one in the journal *Nature,* entitled *Don't let Europe's open-science dream drift*[7]. This article mentions the GO FAIR initiative as the one that kick-started the European scientific cloud by getting data infrastructures to agree to adopt protocols that make at least some of their data FAIR. Another publication of equal relevance appears in the editorial of the journal *Nature Genetics* entitled *Data models to GO FAIR*, reporting the lessons learned in carrying out three workshops aimed at managing FAIR data[8]. Furthermore, in that same year, the article *A design framework and exemplar metrics for FAIRness*[9] is presented in preprint. The article indicates a basic set of semi-quantitative metrics for evaluating the level of FAIRness of the data.

Also in 2017, the first GO FAIR implementation network (RI) called the FAIR *Metabolomics Network*[10] was launched. The year ends with the opening of the international office of the initiative called GO FAIR *International Support and Coordination Office* (GFISCO)[11]**,** supported by the research ministries of Germany, the Netherlands, and France, based in Leiden, the Netherlands, starting its administrative activities of support to the IRs that were beginning to be structured.

[4] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

[5] https://www.dtls.nl/2017/10/04/g7-science-ministers-suggest-policy-guidelines-related-open-science/

[6] https://www.government.nl/latest/news/2017/05/30/germany-and-the-netherlands-call-for-rapid-action-on-the-european--open-science-cloud

[7] https://www.nature.com/news/don-t-let-europe-s-open-science-dream-drift 1.22179?WT.mc_id=TWT_Nature-News&sf90574403=1

[8] https://www.nature.com/articles/ng.3910

[9] https://www.biorxiv.org/content/10.1101/225490v3

[10] https://www.go-fair.org/2017/03/16/metabolomics-implementation-network-launched-key-element-european-open-science-cloud

[11] https://www.go-fair.org/go-fair-initiative/go-fair-offices/

2018: YEAR OF INTERNATIONAL VISIBILITY

The year of 2018 begins administratively structured with the first meeting of GFISCO with the objective of raising the *status quo* of the office local operations; identify ongoing actions of international relevance; review the progress of adhesions and launch of new Implementation Networks (IN). During this meeting, actions were established within the scope of the GO TRAIN pillar – centered on training activities for the adoption of FAIR principles[12]. It is important to emphasize that the GFISCO was created not only to support the activities of the IN, but also to guide the IRs in the elaboration of funding proposals, as well as helping stakeholders to "speak with one voice", strengthening and unifying the discourse regarding FAIR principles.[13]

The GO FAIR initiative begins to grow and expand its activities around the world. In France, it arises through the Ministry of Higher Education, Research and Innovation, which promoted a meeting with representatives of research organizations and funding agencies, presenting FAIR principles and GO FAIR initiative for the French scientific community.[14] In Germany, supported by the Ministry of Education and Research, it emerges at the University of Leibniz[15], with the first workshop that presents the three pillars of the initiative and analyzes national actions and efforts related to the management of research data in that country.

In the second half of 2018, actions focused on the GO BILD pillar began to be developed. One of them was an event of international repercussion, at the University of Leinden, in the Netherlands, entitled *Metadata for Machine* (M4M)[16]. This event was organized by GO FAIR and *Research Data Alliance* (RDA) initiatives, bringing reflections on different possibilities of convergence of metadata standards and other interoperability tools for FAIR services and data. The event was dedicated to the Exchange of knowledge, with initiatives focused on metadata standards and ontologies of different types and infrastructures, such as those developed by *CLARIN*[17], *FAIR Data Point*[18], *CEDAR*[19] and *FAIRsharing*[20] initiatives. The discussions were very fruitful, resulting in a pilot project in partnership with the main Dutch funding agency, ZonMW. This project, entitled *The FAIR Funder pilot program to make it easy*

[12] https://zenodo.org/record/1168504#.WnwCi5POXOQ

[13] https://www.go-fair.org/2018/02/20/report-meeting-potential-go-fair-implementation-networks

[14] https://www.go-fair.org/wp-content/uploads/2018/03/The-First-French-GO-FAIR-Meeting-For-Future-INs.pd

[15] https://www.go-fair.org/2018/10/08/on-the-road-to-fair

[16] https://digitalscholarshipleiden.nl/articles/metadata-4-machines-help-you-find-and-reuse-relevant-research-data

[17] https://www.clarin.eu/content/component-metadata

[18] https://www.go-fair.org/how-to-go-fair/fair-data-point/

[19] https://more.metadatacenter.org

[20] https://fairsharing.org/

for funders to require and for grantees to produce FAIR data[21], was published in preprint with the participation of representatives from Brazilian institutions.

Another action developed in 2018 was the publication of a study carried out by a group of experts who developed a basic set of semi-quantitative metrics, with universal applicability, to assess the level of data FAIRness. This article entitled *A design framework and exemplar metrics for FAIRness*[22] is on the *bioRxiv* preprint server.

This year, the Initiative works actively within the scope of GO CHANGE pillar, participating in international events, seeking to disseminate its actions and expanding FAIR culture around the world. One of these participations was through the poster called *The GO FAIR Approach: Building the EOSC Bottom-up – Based on Implementation Networks*[23], presented at the *Digital Infrastructure for Research* (DI4R)[24] conference, held at the University Institute of Lisbon (ISCTE). This event raised doubts regarding the infrastructure and development issues proposed in the European Union (EU) projects.

Doubts persisted, demanding clarification at international events. One of them took place in the Netherlands with the presence of eighteen countries to discuss issues and misconceptions related to FAIR principles, such as: What is FAIR? What is not FAIR? What is GO FAIR? How does GO FAIR initiative relate to EOSC and the Internet of FAIR Data and Services? And how is GO FAIR initiative joining up with the other international partners associated with FAIR principles?

IN were constantly being created. That year, six new networks were launched, including: C2CAMP, *Genetic, Research Data Infrastructure*, *Economic and Social Sciences goINg FAIR, Rare Diseases,Discovery.*

At the end of 2018, the GO FAIR Brasil office[25] was created, operating in all areas of knowledge. Headed by the Brazilian Institute of Information in Science and Technology (IBICT) in the person of Professor and Researcher Dr. Luana Sales – The first action for the creation of the Brazilian national office was the sending of invitation letters to all research institutions in the country. The first meeting took place in the city of São Paulo, at the event commemorating the 20th anniversary of the Scientific Electronic Library Online (SciELO), with the participation of Professors Dr. Barend Mons and Dr. Luís Olavo Bonino da Silva Santos, both from GFISCO. The following institutions participated in this meeting: Oswaldo Cruz Foundation (Fiocruz), National Nuclear Energy Commission (CNEN), Federal University of the State of Rio de Janeiro (UNIRIO), IBICT and SciELO. At this meeting, the manifest document was drafted and signed by all, which started the work to disseminate the FAIR Principles in Brazil, the establishment of the national office GO FAIR Brasil, the institution of IBICT as the office headquarters, with the general coordination of Professor Luana Sales and the establishment of a form of management by thematic

[21] https://arxiv.org/abs/1902.11162v2

[22] https://www.biorxiv.org/content/10.1101/225490v3

[23] https://www.go-fair.org/wp-content/uploads/2018/10/GO_FAIR_poster_DI4R.pdf

[24] https://cetaf.eu/digital-infrastructures-research-di4r

[25] https://www.go-fair.org/go-fair-initiative/go-fair-offices/go-fair-brazil-office/

area. In the same year, the first Brazilian network was launched: GO FAIR Brasil Health, under the responsibility of the Institute of Scientific and Technological Communication and Information in Health (Icict/Fiocruz) under the leadership of Professor Dr. Viviane Santos de Oliveira Veiga, with the participation of several institutions in the area of public health, health surveillance, health information and communication, history of the cultural heritage of science and health, oncology, nursing and professional health education.

2019: YEAR OF CONSOLIDATION

The year of 2019 begins with the first annual meeting of the IN of GO FAIR initiative[26], in Leiden, the Netherlands, with the attendance of ninety professionals in the fields of computer and information science, data managers and related fields from various countries of Europe, United States and Brazil. This event aimed at bringing the GO FAIR Community closer in the identification of possible forms of synchronization and convergence of its FAIR undertakings. The GFISCO team presented the governance policy of the GO FAIR initiative, and representatives from the USA and Brazil presented the stage of development of the Initiative in their countries.

The year of was also contemplated with the launch of twelve new IN: BiodiFAIRse, *CO-OPERAS, FAIR StRePo, Chemistry, Novel Materials Discovery, **Personal Health Train**, Food Systems, **GlobAl Integrated EArth Data**, GO Inte, Embassadors, **Data Stewardship Competence Centers** and GO FAIR África*[27], with the presentation of its Implementation Plan for the period of 2019-2020, establishing the connection of Africa with the Internet of Services and FAIR Data.[28]

GFISCO promoted the first meeting for the development of FAIR Implementation Matrix.[29] At the same time, the workshop dedicated to "GO CHANGE" took place with the participation of twenty-eight representatives from centers of competence in research data management from Austria, Denmark, the Netherlands, and Germany. This event sought to establish an academic culture of "FAIR" research data, resulting in a collection of articles and documents related to FAIR principles.[30]

Participation in events remains a priority for the GO FAIR initiative and its IN, which participated in OSFair2019, held in Portugal.[31] Both RI Discovery and RI CO-OPERAS held workshop sessions at this event. The poster session had the participation of RI GO Inter and GO FAIR Brazil[32] initiative. GO FAIR initiative was also present at the 14th RDA Plenary, in Helsink[33], presenting the *FAIR Convergence Matrix* tool, under development by a GO FAIR working group. This event also had the Brazilian participation presenting the newly created group RDA Brazil, with the presentation of the poster entitled *A Proposal of Machine-actionable Data Management Plan for Fiocruz.*[34]

[26] https://www.go-fair.org/2019/01/30/go-fair-implementation-networks-meeting

[27] https://www.go-fair.org/implementation-networks/overview/in-africa/

[28] https://www.gzu.ac.zw/data-science-through-go-fair-in-africa-a-new-generation-internet-of-data-and-services/

[29] https://www.go-fair.org/2019/06/19/fair-implementation-matrix-development-meeting/

[30] https://www.zotero.org/groups/2345721/fair_data_resources/

[31] https://www.opensciencefair.eu/

[32] https://www.opensciencefair.eu/posters-2019/machine-actionable-data-management-plan-for-fiocruz

[33] https://www.go-fair.org/2019/11/06/go-fair-at-rda-p14-in-helsinki/

[34] https://www.rd-alliance.org/14th-plenary-call-posters

Within the scope of the GO TRAIN pillar, workshops have become a recurrent practice. The 3rd *GOes FAIR* workshop [35] took place at the *Institute GESIS Leibniz of Social Science*, in Cologne, Germany, with the presence of several European research institutions that addressed topics related to the provision of FAIR research data in their countries.

The last two workshops of the year were offered by GFISCO, at the ZBW - *Leibniz Information Center for Economics,* in Hamburg, Germany. The first of these was the *Semantic Interoperability of Metadata for Cross-Domain Research of the Future*.[36] The second was *FAIR training and skills*[37], with participation of research data administrators from Germany, the Netherlands, the UK, Switzerland and Greece.

Still in 2019, the general coordinator of GO FAIR Initiative, professor Barend Mons, was interviewed by the Brazilian scientific journal LIINC em Revista[38], in the special edition about Research Data. In this interview, his experience of having participated in the *EOSC Hight Level Expert Group* and his experience as a researcher in the management of research data was reported. Professor Mons talks, among other things, about FAIR principles, GO FAIR Initiative, his book *Data Stewardship for Open Science: Implementing FAIR Principles*[39] and his satisfaction at seeing how GO FAIR initiative managed to cross the borders of Europe to Brazil.

In November of the same year, the first GO FAIR Brazil Health seminar took place at the Institute of Scientific and Technological Communication in Health – ICICT/Fiocruz.[40] The opening session was held by the vice-presidency of Fiocruz, the ICICT board and the coordination board of GO FAIR Brazil Health initiative. The opening lecture was given by the representative of GO FAIR International, at the time, Dr. Luiz Olavo Bonino. This seminar also had the participation of GO FAIR Brazil coordination board, the Brazilian representative of GO FAIR international and the coordinators of GO FAIR Brazil Health working groups.

The meeting at GFISCO marks the end of the year 2019, with the presence of representatives from Germany, Belgium, Brazil, Denmark, Great Britain, Italy, the Netherlands, Poland, Switzerland, and the USA for the official launch of the *Data Stewardship Competency Centers* implementation network centers *(DSCC),* contributing to the exchange of knowledge among GO FAIR communities.

[35] https://www.go-fair.org/resources/go-fair-workshop-series/germany-goes-fair-workshops/

[36] https://www.go-fair.org/2019/12/19/go-build-workshop-report/

[37] https://www.go-fair.org/events/go-train-workshop

[38] http://revista.ibict.br/liinc/article/view/5043/433

[39] https://www.taylorfrancis.com/books/9781315380711

[40] https://portal.fiocruz.br/noticia/seminario-da-rede-go-fair-brasil-saude-acontece-nos-dias-7-e-8-11

2020: THE OUTCOMES

The year 2020 begins with the 2nd annual meeting of GO FAIR IN held in Germany, with the participation of the representative of GO FAIR Brazil and twenty-seven IRs, who came together to explore areas of cross-convergence among the existing domains.[41]

A special edition of the Journal Data Intelligence[42], published by MIT Press, in 2020, presents a compilation of articles dedicated to efforts of GO FAIR community on emergent practices of FAIR services and principles. Brazil participated in this compilation with the article GO FAIR Brazil: A Challenge for Brazilian Data Science.[43]

In the same year, the elective course Introduction to FAIR Data Stewardship[44], took place at Hogeschool in the Netherlands, developed under the GO TRAIN pillar. In this same occasion, professor Barend Mons published in the journal Nature an article entitled Invest 5% of research funds in ensuring data are reusable[45] and Erik Schultes, from da GO FAIR Initiative, published the article entitled A role for medical writers in overcoming commonly held misconceptions around FAIR data[46], in the journal Medical Writing Journal.

Seven new IN were created in 2020: FAIR Microbiome, Marine Data Centres, GO NANOFAB, Materials Cloud, AdvancedNano, GO UNI and EcoSoc IN.

Two courses of international relevance took place this year. The first was Metadata for Machine[47], sponsored by Denmark Infrastructure Cooperation (DeiC), in partnership with the GO FAIR Foundation; the second entitled Introduction to FAIR Data Stewardship[48], took place at the Hogeschool, in Leiden, aimed at professionals interested in research data management.

CO-OPERAS IN released the reports of five workshops organized during 2020 about the theme "FAIR data for Social Sciences and Humanities", available in Zenodo repository.[49]

[41] https://www.go-fair.org/2020/01/28/accelerating-convergence-in-2020/

[42] http://www.data-intelligence-journal.org/p/issue/395

[43] http://www.data-intelligence-journal.org/p/52/#:~:text=Today%2C%20GO%20FAIR%20Brazil%2DHealth,the%20process%20of%20adherence%20negotiation.

[44] https://www.go-fair.org/events/introduction-to-fair-data-stewardship/

[45] https://www.nature.com/articles/d41586-020-00505-7

[46] https://journal.emwa.org/the-data-economy/a-role-for-medical-writers-in-overcoming-commonly-held-misconceptions-around-fair-data

[47] https://www.go-fair.org/2020/07/08/m4m-for-the-danish-e-infrastructure-cooperation/

[48] https://www.go-fair.org/2020/07/23/new-fair-data-stewardship-course-in-the-fall-of-2020/

[49] https://www.go-fair.org/2020/08/28/co-operas-publishes-a-variety-of-workshop-reports-on-fairification-efforts-in-the-ssh/

The year 2020 is marked by sad surprises and great concerns around the world with the new Coronavirus pandemic. GO FAIR Initiative was particularly affected by the challenge of fulfilling its role as a leading data manager in line with FAIR principles.

In the Search for solutions to fight the new Coronavirus, the VODAN was created [50], focused on the urgent need to use machine learning and artificial intelligence approaches to discover significant patterns for data management in epidemic outbreaks. VODAN is formed by the Data Together organization – made up of CODATA, GO FAIR, RDA and WDS initiatives, which together form a federated infrastructure. VODAN aims to make data from epidemic outbreaks available for reuse in new research, monitoring and other purposes, under well-defined conditions, respecting patients' privacy, in accordance with current legislation and supported by FAIR principles.

In the process of disseminating VODAN network for the international scientific Community, professor Barend Mons published an article entitled The VODAN IN: support of a FAIR-based infrastructure for COVID-19[51], in the journal European Journal of Human Genetics, where he presents the VODAN network and the infrastructure created to fight COVID-19. From this initiative, focused on fighting the pandemic, other countries joined the network, forming subnetworks.

VODAN África IN[52] is partnered with universities, hospitals, and ministries of health from Uganda, Ethiopia, Nigeria, Kenya, Tunisia, and Zimbabwe. The project had international help and support from Leiden University, GO FAIR Foundation, Tilburg University, Europe External Program Africa (EEPA) and Philips Foundation. For this reason, Vodan Africa network is managing to act in a structured way, standing out as an international reference. The first installation of FAIR Data Point working group for COVID-19 data was done at Kampala International University in Uganda, Africa.[53]

The VODAN Africa team joins Asia by holding together, in 2020, a technical session with the Leiden University Medical Center for intercontinental consultation on the FAIR Data Point. Kampala International University, Stanford University, Leiden University and GO FAIR initiative receive funding from Google.org for the VODAN network. This funding was intended for the implementation of standards for data sharing and platforms for disease modeling in institutions in countries in Uganda, Ethiopia, Nigeria, Kenya, Tunisia and Zimbabwe, aiming at monitoring and disseminating COVID-19.[54]

VODAN Brazil[55] tries to follow the path of Africa with the voluntary efforts of professionals, researchers, and students from Fiocruz, Federal University of Rio de Janeiro (UFRJ) and Federal UNIRIO and with the participation

[50] https://www.go-fair.org/implementation-networks/overview/vodan/

[51] https://www.nature.com/articles/s41431-020-0635-7

[52] https://www.vodan-totafrica.info/

[53] https://www.kiu.ac.ug/special-news-page.php?i=covid-19-computer-readable-observational-data-installed-at-kampala-international-university_1595432235

[54] https://blog.google/outreach-initiatives/google-org/google-supports-covid-19-ai-and-data-analytics-projects

[55] https://vodanbr.github.io/

of Gaffrée and Guinle University Hospital, São José Municipal Hospital and Israelita Albert Einstein Hospital, as reported in Chapter 17 of this book.

Also in 2020, the II GO FAIR Brazil Health Seminar took place in Brazil in partnership with UNIRIO, entitled International Seminar on Health Research Data Management, with the participation of representatives from GO FAIR International and Brazil.[56]

As for Brazil, the last events of the year were marked (1) by the launch of GO FAIR Brazil Health Nursing Network during the celebrations of 130th anniversary of Alfredo Pinto Nursing School, at UNIRIO[57], under the responsibility of the Graduate Program in Health and Technology in the Hospital Environment (PPGSTEH), with a more in-depth report in Chapter 16; (2) by the Brazilian award for the best poster presented at the 16th Plenary of RDA, in Costa Rica, entitled VODAN BRAZIL - the Brazilian experience at the VODAN[58]; and (3) by the Brazilian participation of VODAN BR in the International FAIR Convergence Symposiun 2020[59], organized by CODATA and GO FAIR initiatives. This event provided the creation of a unique forum for the advancing of international convergence among different domains around FAIR principles

FINAL CONSIDERATIONS

Considering the panorama presented here through the reports of the main events that involved GO FAIR Initiative during the period from 2017 to 2020, it is clear that the Initiative has grown nationally and internationally, becoming one of the references for data management in the world. There were more than twenty-seven IN created over these years; several articles were published, several meetings, events, and trainings were held in different locations around the world. The alignment of its activities with the European Commission's open Science was legitimized by the EOSC, being present in its objectives and performance.

It was also noticed that the actions that in the first years were focused on theoretical/conceptual studies, policy definitions, culture Generation; training and interoperability practices related to applications of FAIR principle, were directed to VODAN implementation network. The COVID-19 pandemic has generated panic and concern worldwide, posing major challenges for GO FAIR Initiative. VODAN network was created to help fight this highly contagious virus, assuming the role of infrastructure developer for managing data of patients infected by the new Coronavirus, in line with FAIR principles.

Brazilian has always participated in the GO FAIR Initiative since the creation of GO FAIR Brazil office, at the end of 2018. GO FAIR Brazil has engaged in the search for new members, among which the Humanities network, Agro network and Nuclear Energy network stand out, among others still under negotiation. GO FAIR Brazil Health

[56] http://www.unirio.br/prae/ppgsteh/noticias-1/seminario-internacional-sobre-gestao-de-dados-de-pesquisa-em-saude-1

[57] https://www.go-fair.org/2020/09/12/launch-of-the-go-fair-brazil-health-nursing-network-on-september-22/

[58] https://www.rd-alliance.org/rda-16th-plenary-meeting-poster-sessions

[59] https://conference.codata.org/FAIRconvergence2020/

network is the most structured so far, with the subnetwork of the nursing field. GO FAIR Brazil Network has been represented in international events by its representative, professor Dr. João Moreira, based in the Netherlands, at the University of Twente.

It is true that FAIR principles express a one-way path, and that they have never been so present in global data management practices as they are today. However, what to expect from a future in which data is increasingly placed as a necessary input for the development of new knowledge, for innovation, and for our survival as human beings and citizens? A lot awaits us, especially in terms of scientific and technological advances. However, it will depend on how data will be managed. In this sense, it is a fact that the FAIR principles are of fundamental importance for the promotion of data sharing so that they can be found, accessed, interoperable and reused. Thus, we hope that with the organization of this book we will be able to disseminate the theoretical and empirical research carried out in Brazil and encourage our readers and researchers to join the GO FAIR movement, participating in some way in the Brazilian IN or simply applying the FAIR principles in managing their research data.

Acknowledgements: The organizers of this book would like to give special thanks to the professor at the University of Twente, Dr. Luiz Olavo Bonino and Dr. João Moreira who together made valuable contributions to the elaboration of this Prologue as well as to our knowledge of FAIR principles.

1. FAIR PIDS: THE ROLE OF ORCID IN STRENGTHENING THE FAIR PRINCIPLES

Paloma Marín-Arraiza[60]
Ana Heredia[61]

1.1 INTRODUCTION

The idea behind a Persistent Identifier System (PID system) is to offer a lasting reference of an entity (physical, digital, or abstract), for instance, a digital document, website, person, or institution. Some well-known PID systems are Archival Resource Key (ARK), Digital Object Identifier (DOI), Handle system, Persistent Uniform Resource Locator (PURL), Uniform Resource Name (URN) and Open Researcher and Contributor ID (ORCID iD), the latest exclusively for people.

A PID has a series of associated machine-readable metadata; therefore, they identify the object but not its location, as it happens with a URL (Dappert *et al.*, 2017). A PID can be implemented following the HTTP protocol, which makes it actionable and allows directing the reader to the page Where the resource can be found (*landing page*) (López-Pellicer *et al.*, 2016; Van de Sompel *et al.*, 2014).

However, it is important to point out that the persistence is related to the service offered by the system and not to the identifier itself. This means that an entity commits itself to keep the identifier resolvable. The identifier leads the users to the services that guarantee reference (Kunze, 2013). For example, the ARKs can be maintained and resolved through the EZID service (University of California); the DOIs are generated by the International DOI Foundation and its correspondent registration agency, such as Crossref and DataCite and data centers; the Handles are managed by the Corporation for National Research Initiatives (CNRI); and the PURL system was developed by the Online Computer Library Center (OCLC).

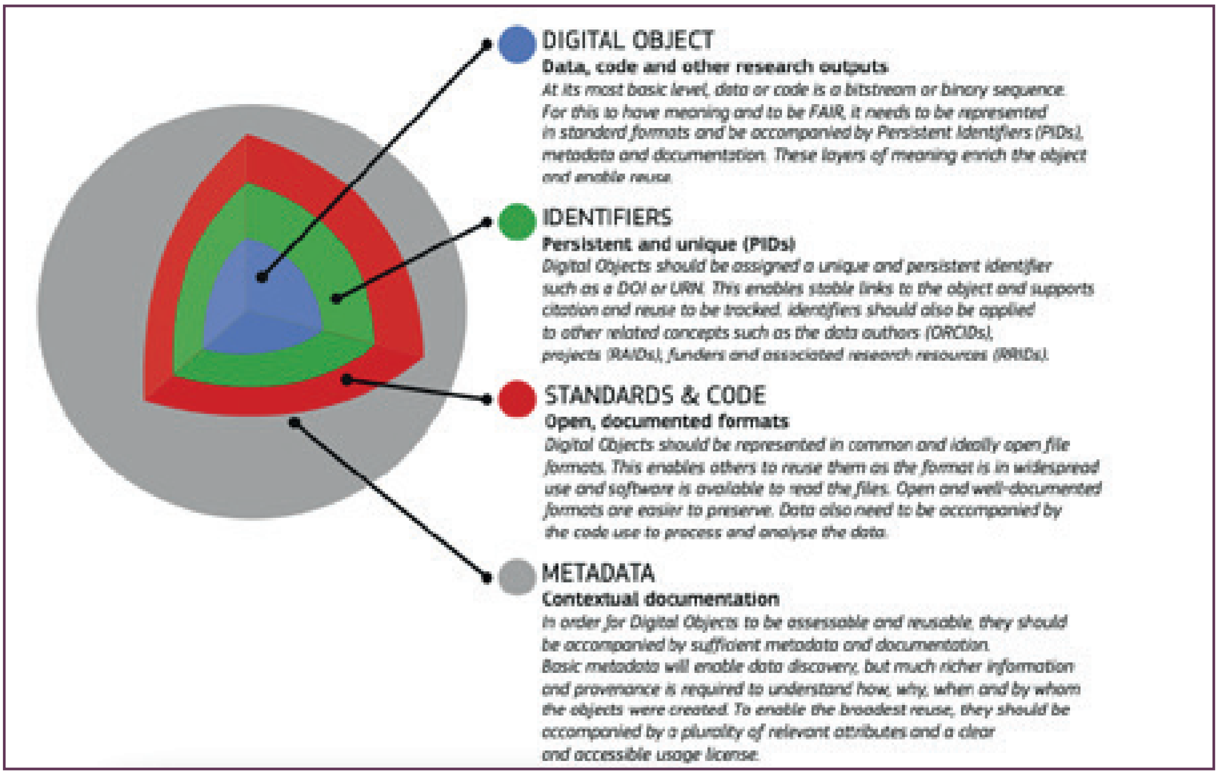
The use of PIDs in archives and research information systems is currently generalized, and PIDs are considered an essential part of the preservation process. For that reason, several research institutions have created data centers to register PIDs, with the aim to preserving their contents and turn them internationally findable and editable. The data center responsible for issuing a PID – for example a research library – must also perform the digital curation to guarantee the maintenance of the resource metadata (Johnston *et al.*, 2018).

[60] Author details: PhD in Information Science (São Paulo State University), Master in Information and Scientific Communication (University of Granada), Bachelor in Physics (University of Granada), ORCID, p.arraiza@orcid.org, https://orcid.org/0000-0001-7460-7794.

[61] Author details: PhD in Sciences (Université Libre de Bruxelles), Master in Cognitive Sciences and Neurosciences (Université Paul Sabatier), Bachelor in Biology (University of Santa Úrsula), Independent information consultant, heredia.a@gmail.com, https://orcid.org/0000-0001-7862-8955.

In fact, current guidelines indicate the use of PIDs, as it is the case of the first FAIR principle: "(Meta)data are assigned as globally unique and persistent identifiers". The report "Turning FAIR into a reality" proposes a FAIR Data Object model (European Commission, 2018), whose layers consist of metadata, standards, identifiers, and data.

Figure 1 – FAIR Data Object Model



DIGITAL OBJECT
Data, code and other research outputs
At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS
Persistent and unique (PIDs)
Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE
Open, documented formats
Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA
Contextual documentation
In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.

Source: European Commission (2018, p.35).

To understand FAIR Data Object, the authors claim that:

> Data needs to be accompanied by Persistent Identifiers (PIDs) and metadata rich enough to enable them to be reliably found, used and cited. In addition, the data should be represented in common – and ideally open – file formats, and be richly documented using metadata standards and vocabularies adopted by the given research communities to enable interoperability and reuse. Sharing code is also fundamental and should include not just the source itself but also appropriate documentation including machine-actionable statements about dependencies and licensing (European Commission, 2018, p.35).

In addition to the identification, PIDs are used to add resources. The research results as a PID are easier to track, which makes the research monitoring activities easier. However, as mentioned earlier, the persistence is not an intrinsic characteristic of a PID, but it is related to the underlying service.

In this matter, it is possible to talk about "reliable identifiers" that are – in addition to persistent – unique, descriptive, interoperable and governed. The ODIN consortium (ORCID and DataCite Interoperability Network) proposed the following characteristics for the reliable identifiers:

> a. unique on a global scale, allowing large numbers of unique identifiers;
>
> b. resolve as HTTP URI's with support for content negotiation, and these HTTP URI's should be persistent;
>
> c. come with metadata that describes their most relevant properties, including a minimum set of common metadata elements. A search of metadata elements across all trusted identifiers of that service should be possible;
>
> d. are interoperable with other identifiers through metadata elements that describe their relationship;
>
> e. Are issued and managed by an organization that focuses on that goal as its primary mission. The organization has a sustainable business model and a critical mass of member organizations that have agreed to common procedures and policies, has a trusted governance structure, and is committed to using open technologies (Ariani *et al.*, 2015, p. 19).

In addition, PIDs work as mechanisms of credit and assignment, when citing the results of the research (McMurry *et al.*, 2017). According to Wilkinson *et al.* (2016), the scientific infrastructures – for example, repositories, supercomputers or physical equipment – can also receive a PID.

1.2 FAIR PIDs AND LEVELS OF MATURITY

PIDs can be internal – when used inside an organization; for example, an employee or student identifier —, owner —When used in a single system; for example, the Scopus Author ID — or open—when they present a complete interoperability with other systems and identifiers; for example, a ORCID iD, a DOI or a Uniform Research Identifier (URI). They allow the establishment of reliable connections among resources.

Demeranville (2018) also defines FAIR PIDs, adding more desirable characteristics to PID systems:

> **FAIR PIDs:** These PIDs are not just resolvable, but can also be used to discover open, interoperable, well-defined metadata containing provenance information predictably. They are openly governed for of the community. Example: DOIs are stored either as URLs "https://doi.org/10.1/123", or simply "10.1/123". We present these to the user as links in the Registry and you can also follow those links to discover metadata describing the linked item. DOIs are governed by the International DOI Foundation and the attached metadata is available under a CC0 license, meaning that it is open to everyone. The metadata contains information about the publisher, the publication, other authors, funding, and affiliation(s), all of which help establish the provenance of the item. Other FAIR PIDs include arXiv identifiers, PubMed and PubMed Central identifiers and most ISBN identifiers. (Demeranville, 2018).

In this matter, it is important to note the maturity of the infrastructure behind these PIDs. We can consider the maturity of an infrastructure when it is in common use in the research Community and across disciples of knowledge. According to the research developed by Ferguson *et al.* (2018) in the framework of FREYA project, Only the "researcher", "publication" and "data" entities currently have mature PID systems.

The following table (Table 1) shows those PIDs whose infrastructure has a high level of maturity.

Table 1 – Entities, types of PIDs and their maturity

Research Entity	Types of PIDs used	Maturity of PIDs infrastructure
Publication	DOI, Accession number, Handle, URN, Scopus EID, Web of Science UID, PMID, PMC, arXiv Identifier, BibCode, ISSN, ISBN, PURL	Mature
Researcher (or scholar)	ORCID iDs, ISNI (also DAIs, VIAFs, arxivIDs, Open IDs, Researcher IDs, Scopus IDs)	Mature
Data	DOI, Accession number, Handle, PURL, URN, ARK	Mature

Source: Adapted from Ferguson *et al.* (2018, p. 9-10).

ORCID iDs are part of this mature research infrastructure and also contribute to FAIRfication of research data, as described below.

1.3 FINDABLE E INTEROPERABLE: THE ROLE OF THE ORCID

FAIR principles guide the entire data releasing process to make them findable - F, accessible – A, interoperable – I and reusable – R.

The role of ORCID in the context of FAIR principles is understood, since ORCID iD acts as an international standard in the persistent identification of authors. Table 2 presents this contribution for each FAIR principle.

Table 2 – The role of ORCID in the FAIR principles

Principle	Description[62]	ORCID Contribution
F1	(Meta)data receives a globally unique and persistent identifier	Providing ORCID iD as PID for “author”/” creator” and “collaborator” .
F2	Data are described in rich metadata[63]	Detail of the provenance information.
F3	Metadata clearly and explicitly includes the data identifier they describe	Inclusion of PIDs in all entries in the ORCID record.
F4	(Meta)data are recorded or indexed in a researchable resource	Availability of ORCID Public API[64] for queries. Annual publication of the ORCID public data archive [65] .
I1	(Meta)data use a formal, accessible, shared and widely applicable language for knowledge representation.	Recognition of ORCID iD as a norm ISO 27729:2012[66]. Use of international standards for the record building (e.g., CASRAI[67]).
I3	(Meta)dados include references qualified to other (meta)data.	Presentation using HTTPS PIDs[68] to discover metadata that describe the linked item

Source: Designed by the authors.

Therefore, the use of ORCID iDs (authenticated[69] if possible), contributes to the FAIRfication process. Furthermore, together with the work to improve the quality and completeness of metadata contained in ORCID records, this will make easier for ORCID to become a reliable source of FAIR data.

1.4 FINAL CONSIDERATIONS

[62] Obtained and translated from https://www.go-fair.org/fair-principles/

[63] It refers to the fact of having relevant atributes and provenance information a

[64] ORCID Public API: https://members.orcid.org/api/about-public-api

[65] ORCID Public Data File 2020: https://doi.org/10.23640/07243.13066970.v1

[66] ISSO 27729:2012. Information and documentation – International standard name identifier (ISNI) https://www.iso.org/standard/44292.html

[67] CASRAI: https://casrai.org/

[68] Identifiers contained in ORCID records: https://pub.orcid.org/v3.0/identifiers

[69] Collection process of an ORCID iD authenticated: https://members.orcid.org/api/tutorial/get-orcid-id

This text intends to present some points about PIDs and their importance in the context of FAIR data, as well as the role of ORCID in the process of data FAIRfication.

ORCID, as a non-profit organization and open infrastructure provider, continues to develop and align its work with improving metadata quality and supporting research communities. FAIR principles also underpin some of this work.

REFERENCES

ARIANI, A.; BARTON, A. J.; BRASE, J.; BROWN, J.; DEMERANVILLE, T.; HERTERICH, P.; MCAVOY, L.; PAGLIONE, L.; RUIZ, S.; THORISSON, G.; VISION, T.; ZIEDORN, F.. **D4.2**: Workflow for interoperability. [*S. l.*]: Figshare, 2015. Available from: https://figshare.com/articles/D4_2_Workflow_for_interoperability/1373669/1. Access on: 26 June 2020.

DAPPERT, A.; FARQUHAR, A.; KOTARSKI, R.; HEWLETT, K.. Connecting the Persistent Identifier Ecosystem: Building the Technical and Human Infrastructure for Open Research. **Data Science Journal**, v. 16, n. 28, p. 1-16, 15 June 2017. DOI 10.5334/dsj-2017-028. Available from: http://datascience.codata.org/articles/10.5334/dsj-2017-028/. Access on: 26 June 2020.

DEMERANVILLE, T. **Blog**: Building a Robust Infrastructure, One PID at a Time. [*S. l.*], 7 Aug. 2018. Available from: https://orcid.figshare.com/articles/Blog_Building_a_Robust_Infrastructure_One_PID_at_a_Time/7008101/1. Access on: 1 Nov. 2020.

EUROPEAN COMMISSION. Directorate General for Research and Innovation. **Turning FAIR into reality**: final report and action plan from the European Commission expert group on FAIR data. Luxembourg: Publications Office, 2018. Available from: https://data.europa.eu/doi/10.2777/1524. Access on: 1 Nov. 2020.

FERGUSON, C.; MCENTRYE, J.; BUNAKOV, V.; LAMBERT, S.; SANDT, S. V. D.; KOTARSKI, R.; STEWART, S.; MACEWAN, A.; FENNER, M.; CRUSE, P.; HORIK, R. V.; DOHNA, T.; KOOP-JACOBSEN, K.; SCHINDLER, U.; MCCAFFERTY, S. **D3.1 Survey Of Current Pid Services Landscape**. Version 1. 17 July 2018. DOI: 10.5281/ZENODO.1324296. Available from: https://zenodo.org/records/1324296. Access on: 1 Nov. 2020.

JOHNSTON, L. R.; CARLSON, J.; HUDSON-VITALE, C.; IMKER, H.; KOZLOWSKI, W.; OLENDORF, R.; STEWART, C.. How Important is Data Curation? Gaps and Opportunities for Academic Libraries. **Journal of Librarianship and Scholarly Communication**, Ames, Iowa, v. 6, n. 1, p. 2198, Apr. 2018. DOI: https://doi.org/10.7710/2162-3309.2198. Available from: https://jlsc-pub.org/article/10.7710/2162-3309.2198/. Access on: 9 Apr. 2019.

KUNZE, J. **The ARK Identifier Scheme**. [*S. l.*]: California Digital Library, 2013. Available from: https://tools.ietf.org/html/draft-kunze-ark-18. Access on: 9 Mar. 2019.

LÓPEZ-PELLICER, F. J.; BARRERA, J.; GONZÁLEZ GARCÍA, J.; ZARAZAGA-SORIA, F. J.; LÓPEZ ROMENO, E.; ABAD POWER, P.; RODRIGUEZ PASCUAL, A. F. El desafío de los identificadores persistentes y accionables. **Mapping**,

n. 180, p. 32-41, 2016. Available from: http://www.jiide.org/Jiide-theme/resources/docs/pdf/articulos/09_art_IAAA_IdentificadoresPersistentesAccionables.pdf. Access on: 10 Mar. 2019.

MCMURRY, J. A. *et al*. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. **PLOS Biology**, *San Francisco*, v. 15, n. 6, p. e2001414, 29 June 2017. DOI: https://doi.org/10.1371/journal.pbio.2001414. Available from: https://dx.plos.org/10.1371/journal.pbio.2001414. Access on: 1 Nov. 2020.

VAN DE SOMPEL, H.; SANDERSON, R.; SHANKAR, H.; KLEIN, M. Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping. **International Journal of Digital Curation**, Edinburgh, v. 9, n. 1, p. 331–342, jun. 2014. DOI 10.2218/ijdc.v9i1.320. Available from: http://www.ijdc.net/article/view/9.1.331. Access on: 9 Mar. 2019.

WILKINSON, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, [*S. l.*], v. 3, p. 160018, Mar. 2016. DOI: https://doi.org/10.1038/sdata.2016.18. Available from: https://www.nature.com/articles/sdata201618. Access on: 9 Mar. 2019.

How to cite this chapter: MARÍN-ARRAIZA, Paloma; HEREDIA, Ana. FAIR PIDS: the role of ORCID in strengthening the air 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. 1, p. 23-29. DOI: 10.22477/9788570131959.cap1.

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.

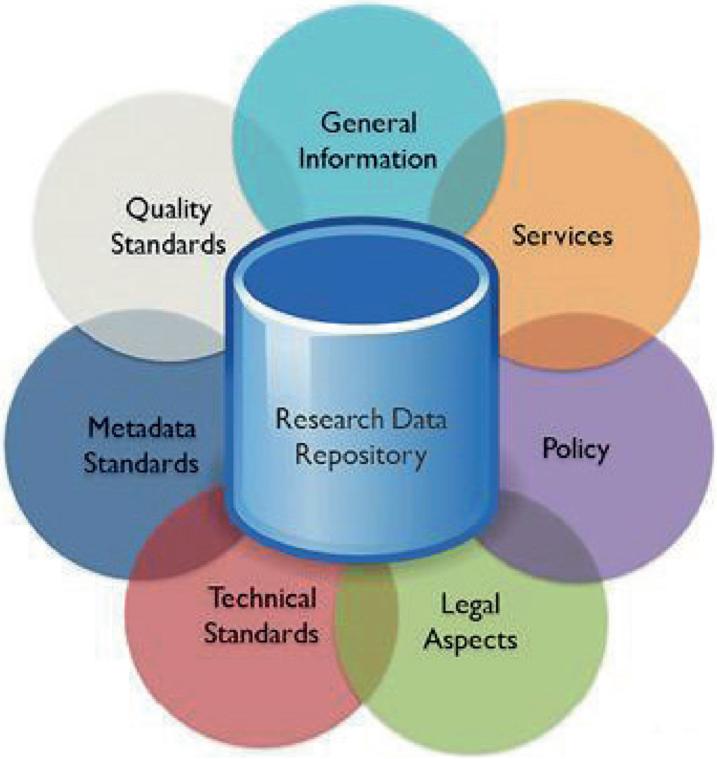
2.2 RESEARCH DATA REPOSITORY: THE DATAVERSE PROJECT

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.

3. TOWARDS THE IMPLEMENTATION OF THE GO-FAIR AGRO BRAZIL NETWORK: THE EXPERIENCE OF A RD&I ORGANIZATION IN THE IMPLEMENTATION OF THE FAIR PRINCIPLES

Debora Pignatari Drucker[97]
Juliana Meireles Fortaleza[98]
Patrícia Rocha Bello Bertin[99]
Isaque Vacari[100]
Carla Geovana do Nascimento Macario[101]

3.1 INTRODUCTION

Agricultural Science is a multi and interdisciplinary Science due to the integrative nature of knowledge about the physical, biological, social and economic environments. This means that the production, trade, and consumption of food are closely related to human and environmental health, and the ability to connect knowledge on these themes, which are often controversial, is essential to overcome complex problems in agriculture (Hilimire, 2016). Considered the basis of the scientific method (Heidorn, 2008), data are essential for the integrated analysis of the several disciplines that are part of agricultural science, and which must be treated as valuable products of the Research, Development, and Innovation (RD&I) activity.

In this context, the amount of data to be analyzed has increased dramatically, with the transition to a much more data-intensive science in all areas of knowledge (Hey; Tansley; Tolle, 2009). In Agricultural Science, large amounts of data are being generated from remote and proximal sensing technologies, which are used to monitor variables of interest in real time – such as soil, climate, atmosphere, or organism variables – as well as from genomic or natural language processing technologies. At the same time, long-tail data, that is, data that are heterogeneous, diverse, poorly structured, and difficult to be obtained (Heidorn, 2008; Borgman *et al.*, 2016) are accumulated for decades — these records of immeasurable value about production systems and environmental assets are at risk of being lost if not properly treated and preserved.

Reliability and reproducibility are also important pillars of the scientific method, which require sound research data management practices to be ensured. In this respect, the FAIR principles (*Findable, Accessible, Interoperable, Reusable*)

[97] Forest Engineer, Embrapa Digital Agriculture, debora.drucker@embrapa.br

[98] Agronomist Engineer, Institutional and Government Relations Division, juliana.fortaleza@embrapa.br

[99] Biologist, Executive Directorship for Research and Innovation, patricia.bertin@embrapa.br

[100] Computer Scientist, Embrapa Digital Agriculture, isaque.vacari@embrapa.br

[101] Computer Scientist, Embrapa Digital Agriculture, carla.macario@embrapa.br

(Wilkinson *et al.*, 2016) are being increasingly adopted as guidance for research data management and enablers of their reuse. Applications and metrics for adopting and evaluating FAIR principles have been developed worldwide through standards, metadata, controlled vocabulary, ontologies and persistent identifiers that bring accurate meaning to data and other research outputs (Henning *et al.*, 2019).

The recognition by the scientific community of the importance of making data management practices adherent to FAIR principles led to the emergence of initiatives such as 'GO FAIR', which articulates communities of practice from thematic and regional Implementation Networks (INs)[102]. One of those networks, named *Food Systems*, aims at supporting the implementation of FAIR principles in agri-food sciences[103]. Brazil is part of this initiative with a national office[104], to which several INs are associated – among them, the GO FAIR Agro Brazil Network, which will contribute to the adoption of the FAIR principles by institutions that produce agricultural data (Go-Change); provide training in partnership with other national INs (Go-Train); and collaboratively build and implement infrastructure and interchangeable patterns (Go-Build).

This chapter aims at reporting the experience of Embrapa in incorporating the FAIR principles into institutional policies, as well as into data governance and management processes and practices. The narrative was built from an exploratory case study, with Embrapa as a single unit of analysis and data collected through documentary research. With its theoretical foundation on conceptualizations of Open Science, e-Science and Research Data Management domains under the perspective of the FAIR principles, the analysis herein serves as a basis for the GO FAIR Agro Brazil Implementation Network, thus benefiting the entire national agricultural RD&I system.

The following chapters present contextual information on Research Data Management (RDM) at the Brazilian Agricultural Research Corporation (Embrapa), explaining its positioning in the global RDM system, describing current internal regulations, and reporting results obtained thus far. Finally, the challenges and future prospects are discussed, to make Embrapa's research data increasingly adherent to the FAIR principles.

3.2 A FEW WORDS ABOUT RESEARCH DATA MANAGEMENT AT EMBRAPA

The mission of Embrapa – a governmental research institution linked to the Brazilian Ministry of Agriculture, Livestock and Supply – is "to create research, development, and innovation solutions to ensure the sustainability of agriculture, for Brazilian society (Embrapa, 2020, p. 16). Organized in 43 research centers geographically distributed nationwide and with strong partnerships abroad, the company generates a large volume of data on the various strategic themes of agricultural research. Aware of the volume, velocity, variety, and value of the research data produced by its activities, Embrapa has mobilized efforts to properly govern and manage these assets throughout their life cycle, to make them findable, accessible, interoperable and reusable.

[102] More information can be found in the GO-FAIR portal: https://www.go-fair.org.

[103] More information on the Brazilian office can be found at: https://www.go-fair.org/go-fair-initiative/go-fair-offices/go-fair-brazil-office.

[104] The portal for data access is available at: https://metabuscador.uspdigital.usp.br/.

Among those efforts, the corporate project entitled "Data and Information Governance for Knowledge at Embrapa: Model and Implementation Plan Development" is noteworthy, which aimed at designing, validating and proposing a systemic model for the governance of data and information at the organization. As a result of the project, several recommendations were made to improve research data governance and management (Table 1), which associate well with the GO Build, GO Change, and GO Train pillars of the GO FAIR Implementation Networks.

Table 1 - Recommendations to improve research data governance and management at Embrapa, as related to the GO Build, GO Change and GO Train pillars of the GO FAIR Implementation Networks

Category	Recommendations
Related to corporate processes	• To model and implement corporate processes for research data management and open data publishing. *Change* • To review the employees' performance evaluation and reward process with the aim of fostering the culture of data sharing and data reuse. *Change* • O develop and implement the processes, competences, tools, and methodologies to enable semantic interoperability. *Build* • To enable the linkage of scientific data to the publications and the projects that generated them. *Build* • To adopt public licenses for digital assets. *Build*
Related to internal guidelines and norms	• To develop and publish a corporate Open Data Plan. *Change* • To develop, review, update, and implement policies and internal rules related to research data and information management . *Change* • To add to the company's Master Strategic Plan, specific guidelines related to research data and information management. *Change* • To establish and sustain a corporate model for research data. *Build*
Related to the organizational culture	• To ensure active participation in national and international forums and networks on research data management. *Change* • To engage Information Science professionals on research data management processes. *Change* • To train and communicate about research data management. *Train* • To require Research Data Management Plans. *Change*
Related to tools, instruments, and technologies	• To develop and implement technological infrastructure for research data management through consistent and interoperable platforms. *Build* • To adopt persistent identifiers for data, datasets, and authors. *Build* • To ensure the alignment of data management plans and information architectures with epistemologically systematized and globally used conceptual models for agriculture. *Build* • To build an open data infrastructure, interconnected with the Brazilian Open Data Portal. *Build* • To implement terminology management and conceptual alignment technological tools. *Build*

Category	Recommendations
Related to structure, roles and responsibilities	• To define rules and responsibilities for research data management within the organization. *Change*

Source: Authors (2024).

It is evident that cultural changes (Go Change), training (Go Train), and building and implementing infrastructure and interchangeable patterns (Go Build) permeate the 19 recommendations in Table 1. The recommendations related to corporate processes align to the Change (1 and 2) and the Build pillar (Build: 3, 4 and 5), while the normative recommendations fit the Change (6, 7 and 8) or the Build category (9). Recommendations related to the organizational culture associated with the Train (12) or the Change pillars (10, 11 and 13). All recommendations related to tools, instruments, and technologies align with the Build category (14, 15, 16, 17 and 18), and finally, the recommendation related to structure, roles, and responsibilities categorizes as Change. In total, nine recommendations were categorized as building and implementation (Build), the other nine as cultural change (Change), and one as training (Train). One of the recommendations categorized as cultural change (10 - "To ensure active participation in national and international forums and networks on research data management") is explored in the next section.

3.3 INTRODUCING EMBRAPA IN THE GLOBAL RESEARCH DATA MANAGEMENT ECOSYSTEM

The Big Data phenomenon and the new e-Science and Open Science paradigms have promoted a transformation in the scientific system, with practices, rules, and behaviors reconfiguration especially, in research data organization and management (Algabli *et al.*, 2015). To better understand and benefit from this transformation, Embrapa has exchanged knowledge through participation in national and international initiatives, networks, groups, and forums addressing RDM. As part of these efforts, Embrapa coordinated, between October 2018 and July 2020, the Commitment 3 of the 4th National Action Plan for Open Government, known as the 'Brazilian Commitment towards Open Science', which aimed at "establishing mechanisms for scientific data government for the advancement of Open Science in Brazil" (Brasil, 2018a, 2018b). The Commitment was carried out in partnership with several government agencies and civil society, including the Ministry of Science, Technology, and Innovations (Ministério da Ciência, Tecnologia e Inovações - MCTI), Oswaldo Cruz Foundation (Fundação Oswaldo Cruz - Fiocruz), Brazilian Institute of Information in Science and Technology (Instituto Brasileiro de Informação em Ciência e Tecnologia - Ibict), Ministry of Agriculture, Livestock and Food Supply (Ministério da Agricultura, Pecuária e Abastecimento - MAPA), National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq), Coordination for the Improvement of Higher Level Personnel (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Capes), National Education and Research Network (Rede Nacional de Ensino e Pesquisa -RNP), University of Brasília (Universidade de Brasília - UnB), and Open Knowledge Brazil (OKBR). Among the results of this Commitment, the following stand out a diagnosis of the Open Science developments in the world and in Brazil, a guidance document about interoperability patterns; indicators for evaluating organizational maturity for research data openness; the development and implantation of pilot institutional data repositories; awareness and training actions on the topic; liaising with scientific editors and funding agencies; and the creation of an inter-institutional network on the topic.

The Open Government agenda turned out to be a suitable environment for strengthening and expanding partnerships among the various actors of the national scientific system, thus avoiding duplicate efforts.

Furthermore, noteworthy is Embrapa's participation in the working group for the implementation of a Network of Scientific Data Repositories in the State of São Paulo, created by the Foundation for Research Support of the State of São Paulo (Fundação de Apoio à Pesquisa do Estado de São Paulo - Fapesp), contemplating data and information from state public universities[105]. Embrapa Digital Agriculture, one of Embrapa's research centers, integrates the WG, focusing on sharing data and information generated by the company, which can be used for scalability tests and integration of agricultural data to repository networks. Furthermore, the Company shares technical knowledge and contributes to speeding up the activities developed by WG in matters such as scalability tests, evaluation of tools or data curation procedures, among others.

In the international context, Embrapa integrates the GODAN (*Global Open Data for Agriculture and Nutrition*) initiative, which aims at promoting global efforts for providing, accessing and reusing relevant data in agriculture and nutrition[106]. As part of the network with more than 1,110 partners, the Company has worked on the translation to Portuguese of instructional materials about open data management in agriculture. The company has also contributed to the *Research Data Alliance* (RDA[107]) - a global initiative started in 2013 to foster open sharing and reusing of research data. Currently, RDA has more than 11,000 members, including data producers, users, and managers who contribute to the development of RDM solutions and good practices. Specialists meet in thematic groups, one of which is the *Improving Global Agricultural Data (IGAD) Community of Practice* (IGAD), coordinated by FAO, USDA, GFAR (Global Forum on Agricultural Research and Innovation), and Embrapa representatives. Another international group with Brazilian coordination is *Professionalizing Data Stewardship*, which brings together professionals from all continents to achieve a common goal regarding the professionalization of research data stewardship. The Data Observation Network for Earth (DataOne[108]) is one more initiative that has contributed to the development and adoption of strategies and best practices for RDM at Embrapa. DataOne aims at a deeper understanding of life on Earth and the environment supporting it; it is held by the community and provides data for several of the members' repositories, promoting the best practices in data management through educational resources and materials. The interface between Science and politics is the focus of action of *Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* (IPBES[109]), which has a task force for data and knowledge that, among other actions, proposes the data policy of the platform and monitors its adoption. Jointly, such initiatives helped to disseminate and adopt RDM best practices and thus, data can be reused, leading to the advance of frontiers of knowledge and subsidizing decision-making in various spheres. The participation of Embrapa in these discussion forums makes it possible to exchange experiences, continuous update and review of strategies and activities, described in the following sections.

[105] More details on the Godan initiative can be found in: https://www.godan.info/

[106] RDA portal contains information on members and work groups: https://www.rd-alliance.org/.

[107] More information in: https://www.dataone.org/.

[108] IPBES portal contains details on the platform: https://ipbes.net/.

[109] More information in: http://www.embrapa.br/siexp.

3.4 EMBRAPA DATA GOVERNANCE, INFORMATION, AND KNOWLEDGE POLICIES

One of *GO Change* recommendations, among the normative measures presented in Table 1, is number 7: "Elaborate, update and implement policies and internal rules for the management of data and information generated during research developed by the Company". An essential action in this regard was the enactment of Embrapa Data Governance, Information, and Knowledge Policy that establishes principles, guidelines, and responsibilities to "strengthen the mechanisms for generating, organizing, treating, accessing, preserving, retrieving, disseminating, sharing and reusing Embrapa's information assets" (Embrapa, 2019, p. 10).

Guided by the principles of the Constitution of Federative Republic, the Declaration of Human Rights and the precepts of the Open Science movement, the policy established 17 principles for Embrapa's data, information and knowledge management, namely: (1) Data, information, and knowledge as corporate assets; (2) Strategic alignment; (3) Development of capacities and competences; (4) Federated infrastructure; (5) Analysis, intelligence, and innovation based on data; (6) Efficiency and economics; (7) Compliance and risk mitigation; (8) Interoperability; (9) Licensing; (10) Preservation and memory; (11) Privacy, protection, and confidence; (12) Safety; (13) Quality and integrity; (14) Epistemological specificity; (15) Organizational learning, continuity and knowledge retention; (16) Openness and transparency; (16.1) Open Access to scientific information; (16.2) Open Data; (17) Monitoring and responsibility when disseminating relevant information.

Although FAIR principles are not directly stated in the guidelines and principles of the policy, the document's core is "well organized, documented, accessible and accuracy and validity checked data and information are easily shareable and reusable", providing several advantages to the administration (Embrapa, 2019). Notably, the publication responsible for introducing the FAIR principles (Wilkinson *et al.*, 2016) is one of the basic references of the policy and constitutes transversal elements to all its content. For instance, the following guideline should be highlighted: "implement and support processes that ensure that data and information produced by Embrapa are reliable and easily retrievable, accessible, interoperable and reusable" (Embrapa, 2019).

Among the principles of the policy, 'Interoperability' is the one which better means the need of applying the FAIR principles. To be covered by RDM best practices, this principle requires the use of semantic tools and widely established and widespread data and metadata standards . This principle is strengthened through the guideline described from the technological perspective that directs to the innovation and to the use of technologies allied to international trends, such as data sharing and reuse with broad adoption of interoperability services.

Thus, FAIR principles encompass a central reference for the development of the GDP Corporative Program, outlined in item 8.1 of the policy strategic guideline: "implement, support and monitor a Research Data Management Corporate Program and guide the development of data management plans in the context of Research, Development, and Innovation projects" (Embrapa, 2019, p. 13).

3.5 RESEARCH DATA MANAGEMENT (RDM) CORPORATE PROGRAM: ACTIONS IN PROGRESS

3.5.1 Diagnosis on data management practices

Throughout the history of Embrapa, several data management practices and information systems were created, according to the specificities of diverse thematic areas in different research centers at Embrapa. Thus, one of the initial actions of the RDM Corporate Program was to carry out a survey to diagnose them. To this end, an electronic questionnaire was elaborated to describe important points related to research data: data characterization, collection and documentation, data storage, backup, accessibility, sharing and reusing, and research data repositories. The questionnaire was answered by 854 data producers distributed among 43 research unities with the aim of supporting and guiding RDM corporate improvement actions of Emprapa, according to the better international practices and trends of data organization and publication.

3.5.2 Development and Implementation of a Reliable Data Repository and Persistent Identifiers Attribution

The current context requires a comprehensive strategy to establish the roots to make Embrapa data FAIR, considering the importance of strengthening solutions that already exist in the Company, such as the Embrapa Experiment Information System (SIExp[110]) and Embrapa Spatial Data Infrastructure – GeoInfo (Drucker *et al.*, 2017), as well as the need to accommodate data for which proper solutions have not been implemented yet. The complexity and multidisciplinary of Agricultural Sciences require technological solutions that allow adherence to FAIR data and, at the same time, accommodation of data from different domains and their representation models and standards, to obtain a central description core, as well as the treatment of specificities via small extensions of this core, so that the interoperability of scientific data repositories is enabled in general at different levels. Considering the best practices adopted worldwide, it was decided to implement a reliable research data repository as a solution for the organization, treatment, preservation, and publication of data produced by Embrapa.

Embrapa Data Repository, Redape[111], was launched in April 2022 and is based on *Dataverse*[112] open data software . Additionally, a computational infrastructure for storing research data was acquired and enabled, and a team responsible for the technical administration of the repository, located at Embrapa Scientific *Data Center*, was nominated. Embrapa Scientific *Data Center* is based in Campinas, SP, and has essential characteristics of information security, such as: restrict access to computers, servers and data storage disks, as well as defense against attacks to repository, prevent access to unauthorized individuals, among others.

[110] Available from: https://dataverse.org.

[111] Available from: https://www.redape.dados.embrapa.br/

[112] PhD in Information Science. Graduated in Engineering. Associate Professor at Unirio where he works in the Department of Technical-Documentary Processes and in the Postgraduate Program in Library Science.

Redape supports the assignment of persistent identifiers, one of the fundamental requirements to make data products available to the scientific community. A persistent identifier (PID) enables the unique identification of a digital object and is addressed to be a permanent way of identifying and accessing this specific resource. The most widely known PID in the scientific community is the *Digital Object Identifier* (DOI), which generates a persistent link that points to the repository or to other digital location when including the URL in the metadata. It provides a system for persistent and actionable identification, as well as for interoperable exchange.

3.5.3 Knowledge Representation

According to Meadow *et al.* (2007), information retrieval is a communication process between record authors, creators, and readers. This process depends on a proper language control (code) between the sender and receiver and between users' documents and requirements (Janaite Neto; Ferneda, 2016). Building controlled vocabulary aims at information indexing, storing and retrieving activities, representing meaningful concepts of some domain of knowledge and, if possible, engaging with types and even subtypes of domain (Chandrasekaran *et al.* 1999; Cintra, 2002; Jacob, 2003; Fujita, 2004).

According to Lattes Platform (CNPq), Agrarian Sciences can be subdivided in the following subareas: Agronomy, Forest Resources and Forest Engineering, Agricultural Engineering, Zoo technics, Veterinary Medicine, Fishery Resources and Fishing Engineering, Food Science and Technology (CNPq, 2021). This diversity of subareas contributes to the high number of terms that can be used for indexing, storing and retrieving information. Agrovoc Multilingual Thesaurus, for instance, accounts for 33,388 main terms and 2,254 alternative terms, including food, nutrition, agriculture, forestry, fishing, scientific and common names of animals and plants, environment, biological concepts, plant cultivation techniques, among others. Agroterms – controlled vocabulary built by a permanent work group at Embrapa – gathered nearly 245 thousand terms pertinent to the agricultural knowledge domain from the gathering of terminologies in Portuguese found in national and international agricultural thesaurus. The expectation is of expanding Agroterms to a conceptual space of Brazilian agricultural knowledge and promoting a better interoperability between internal and external information systems.

3.5.4 Data management Plan

A Data Management Plan is one of the most important stages in the research development process, for it is at that moment that data treatment throughout its lifecycle is discussed. It is also discussed how to ensure that data is freely available – respecting privacy – and how it can be reused, under specific conditions and licenses clearly defined, and which can be properly cited and used as reference. A Data Management Plan (DMP) is, therefore, an essential tool so that best data management practices are applied during the research development until data publication. Aware of DMP importance, development agencies in the United States, European Union, United Kingdom, Australia, and Canada have demanded that research projects are accompanied by a DMP in line with FAIR principles (Aventurier, 2017). In Brazil, Fapesp was the first Brazilian funding agency to announce, in 2017, the obligation of having a DMP for research projects funding requests. Research institutions must insert DMP in their research development process, not only to obtain funding, but also to ensure that data is properly managed. At Embrapa, the obligation of attaching a DMP to all research projects started in January 2022, while it has been a practice adopted by Embrapa Digital Agriculture since 2018.

3.6 Final Considerations

This work described efforts that have been applied at Embrapa to implement research data management based on FAIR guiding principles, and it sought to fit the mapped measures according to the pillars of GO FAIR initiative. The section describing the insertion of Embrapa in the DMP global ecosystem showed that there are several actions in course related to this theme, which meet one of the recommendations characterized as cultural change (Go-Change), number 10: "Ensure Embrapa inclusion and active participation in national and international forums and networks in research data management.". The results obtained so far are remarkable and have the potential to be multiplied and expanded in the coming years, based on the strengthening of relationships and ties with partner institutions. This is a feature that substantiates the *Go-Change* pillar, as Community Building underpins the GO FAIR initiative.

Another significant action to encourage the adoption of practices adhering to the FAIR principles and to support the cultural change necessary to promote the pillars of the GO FAIR movement was the enactment of Embrapa's Data, Information and Knowledge Governance Policy, described herein. To incorporate those principles to the everyday activities of the organization, the company is implementing the Corporate Research Data Governance Program, which will ensure the necessary means, services, and tools so that data produced by projects are easily located, accessed, interoperated and reused. Among the actions for the implementation of this corporate program are the following: diagnosis of data management practices; implementation of a reliable data repository, with the attribution of persistent indicators and viability of data discovery, which were disconnected until then; the development of actions to enable the representation of Agricultural Science knowledge and the establishment of practices of elaborating data management plans in research projects developed by the company.

It is worth highlighting that the research data management process was mapped and formally described, allowing for a deeper understanding of the current practices, so that services and solutions are adherent to the organizational and epistemological culture. This process is not elementary, given the multidisciplinary nature of Agricultural Science and the need to involve different actors and competences to achieve a model of this complexity. In combination, training actions are essential for the successful implementation of the DMP Corporate Program.

Another challenge to be faced is to properly contemplate the interoperability principle by adopting data and metadata standards and semantic tools – essential requirements so that data is properly interpreted and, thus, allowing for its reuse. Once more, considering the great diversity and heterogeneity of data created and analyzed in the context of Agricultural Sciences, it is a challenge that requires the participation of several actors from different subjects that make up agricultural research. The attribution of licenses that clarify the terms of data use is also an essential condition.

As challenges and future perspectives, training, and education actions are crucial for the successful implementation of the DMP Corporate Program. Another challenge to be faced is to properly contemplate the interoperability by adopting data and metadata standards, as well as ontologies and semantic tools, essential requirements so that data is properly interpreted and, thus, enabled to be reused. Once more, considering the great diversity and heterogeneity of data created and analyzed in the context of Agricultural Sciences, it is a challenge that requires the participation of several actors from different subjects

Finally, the elaboration of monitoring strategies of the DMP Corporate Program with a view to adherence to FAIR principles, allowing for the incorporation of improvements, is an essential perspective to ensure its success. A reference base is the work performed in the scope of Goal 9 of the Commitment to Open Science of the Open Government Partnership, entitled "Proposition of a set of indicators for measuring maturity in Open Science". Although the FAIR principles do not necessarily entail data opening, the set of indicators for measuring the maturity level of scientific data opening provides objective criteria that can also be used for measuring the success in Governance, Organizational Culture, Scientific Data Management and Technological Infrastructure fields (Fortaleza *et al.* 2020). More specific indicators can be developed, such as metrics for cataloging research data in institutional repositories; quantity of access to available resources; description and implementation of processes; licensing; and establishment of reward mechanisms when sharing and reusing data.

As demonstrated in the case of Embrapa, the adherence to FAIR principles is crucial for the technological and semantic interoperability of data in the agricultural context. The strategy presented herein assumes that data are valuable products of the research activity, and denotes the transition to a praxis in which reusing data from the agricultural research based on the paradigm of FAIR principles is encouraged. In this regard, building the GO FAIR Agro Brasil network is critical so that the collaborative work mutually benefits the communities that manage agricultural data. Thus, encouraging the reuse of agricultural research data will contribute to the solution of problems not only for Brazilian society but also the society worldwide, considering the relevance of the country in the context of food systems.

REFERENCES

ALBAGLI, S.; MACIEL, M. L.; ABDO, A. H. (org.) **Ciência aberta, questões abertas**. Brasília: Ibict, 2015. Available from: http:/livroaberto.ibict.br/handle/1/1060. Access on: 11 oct. 2024.

AVENTURIER, P. Plano de Gestão de Dados: uma introdução. **Publicação Científica** [*Blog*]. Published in: 17 may 2017. Available from: https://publicient.hypotheses.org/1660. Access on: 30 oct. 2020.

BORGMAN, C. L. *et al.* Data management in the long tail: science, software, and service. **International Journal of Digital Curation**, v. 11, n. 1, p. 128- 148, 2016. DOI: 10.2218/ijdc.v11i1.428. Available from: https://ijdc.net/ijdc/article/view/11.1.128. Access on: 11 oct. 2024.

BRASIL. Controladoria Geral da União. Inovação e governo aberto na ciência - monitoramento e execução: compromisso 3. Estabelecer mecanismos de governança de dados científicos para o avanço da ciência aberta no Brasil. 2018a. **CGU** [*Site*]. Published in: 29 oct. 2019 às 14h23. Atualizado em: 18 aug. 2022 às 14h59. Available from: https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/planos-de-acao/4o-plano-de-acao-brasileiro/compromisso--3-docs/inovacao-e-governo-aberto-na-ciencia-monitoramento-e-execucao. Access on: 3 mar. 2021.

BRASIL. Ministério da Transparência e Controladoria-Geral da União. **4º Plano de Ação Nacional em Governo Aberto**. Brasília: CGU, 2018b. Available from: https://repositorio.cgu.gov.br/handle/1/66740. Access on: 6 oct. 2020.

CHANDRASEKARAN, B.; JOSEPHSON, John R.; BENJAMINS, V. R. What are ontologies, and why do we need them? **IEEE Intelligent Systems**, v. 14, n. 1, p. 20-26, jan./feb. 1999. DOI: 10.1109/5254.747902. Available from: https://ieeexplore.ieee.org/document/747902. Access on: 11 oct. 2024.

CINTRA, A. M. M. *et al*. **Para entender as linguagens documentárias**. 2. ed. São Paulo: Polis, 2002.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO - CNPQ. Áreas do conhecimento – Ciências Agrárias. **Diretório de Grupos de Pesquisa no Brasil** [*Site*]. [2021?]. Available from: http://lattes.cnpq.br/web/dgp/ciencias-agrarias. Access on: 3 mar. 2021.

DRUCKER, D. P. *et al.* GeoInfo: infraestrutura de dados espaciais abertos para a pesquisa agropecuária. **RECIIS:** Revista Eletrônica de Comunicação, Informação & Inovação em Saúde, v. 11, p. 1-17, 2017. Suplemento. Available from: https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1083246/1/GeoInfo.pdf. Access on: 1 mar. 2021.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. **Política de Governança de Dados, Informação e Conhecimento da Embrapa**. Brasília: Embrapa, 2019. Available from: https://www.embrapa.br/politica-de-governanca-de-dados-informacao-e-conhecimento. Access on: 3 mar. 2021.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. **Deliberação nº 2, de 28 de janeiro de 2020**. Brasília: Embrapa, 2020. Available from: https://www.embrapa.br/documents/10180/1546282/Regimento+das+Secretarias+da+Embrapa/d629c401-d2e6-fd8d-5154-ccbaaa1e3313. Access on: 30 oct. 2020.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. **VII Plano Diretor da Embrapa**: 2020–2030. Brasília: Embrapa, 2020. Available from: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/217274/1/VII-PDE-2020.pdf. Access on: 9 Set. 2022.

FORTALEZA, J. M.; BERTIN, P. R. B.; DRUCKER, D. P.; ASSIS, T. B.; COSTA, M. P. Conjunto de indicadores para aferição do grau de maturidade de abertura dos dados científicos. Brasília: Embrapa, CNPq, OKBR, Ibict, MCTI, 2020. 14 p.

FUJITA, M. S. L. A leitura documentária na perspectiva de suas variáveis: leitor-texto-contexto. **DataGramaZero**: Revista de Ciência da Informação, v. 5, n. 4, Aug. 2004.

HEIDORN, P. B. Shedding light on the dark data in the long tail of science. **Library Trends**, v. 57 n. 2, p. 280-299, fall 2008. DOI: 10.1353/lib.0.0036. Available from: https://muse.jhu.edu/article/262029. Access on: 14 Oct. 2024.

HENNING, P. C. *et al.* Desmistificando os princípios fair: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados FAIR. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia,** Paraíba, v. 14, n. 3, p. 175-192, 2019. DOI 10.22478/ufpb.1981-0695.2019v14n3.46969.

HEY, T.; TANSLEY, S.; TOLLE, K. (ed.). **The fourth paradigm:** data-intensive scientific discovery. Redmond: Microsoft Research, 2009. Available from: https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf. Access on: 1 set 2020.

HILIMIRE, K. Theory and practice of an interdisciplinary food systems curriculum. **NACTA Journal**, v. 60, n. 2, p. 227-233, 2016. DOI 10.2307/nactajournal.60.2.227.

JACOB, E. K. Ontologies and the semantic web. **Bulletin of the American Society for Information Science and Technology**, v. 29, n. 4, p. 19-22, Apr./May 2003. DOI: 10.1002/bult.283. Available from: https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/bult.283. Access on: 14 oct. 2024.

JANAITE NETO, J.; FERNEDA, E. Ontologia como recurso de padronização terminológica. **Informação em Pauta**, v. 1, n. 1, p. 30-45, 2016. DOI: 10.32810/2525-3468.ip.v1i1.2016.2967. Available from: http://www.periodicos.ufc.br/informacaoempauta/article/view/2967. Access on: 14 oct. 2024.

MEADOW, C. T. *et al.* **Text information retrieval system.** 3. ed. Amsterdam: Elsevier, 2007. Available from: https://diglibrary.weebly.com/uploads/1/8/5/1/18511482/text_info_retrieval_system.pdf Access on: 30 oct. 2020.

WILKINSON, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, v. 3, p. 1-9, 2016. DOI: 10.1038/sdata.2016.18. Available from: https://www.nature.com/articles/sdata201618. Access on: 14 oct. 2024.

How to cite this chapter: DRUCKER, Debora Pignatari; FORTALEZA, Juliana Meireles; BERTIN, Patrícia Rocha Bello; VACARI, Isaque; MARARIO, Carla Geovana do Nascimento. Towards the Implementation ff the GO-FAIR agro brazil network: the experience of a RD&I organization in the implementation of the 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. 3, p. 46-57. DOI: 10.22477/9788570131959.cap3.

4. FAIR PRINCIPLES AND GOVERNMENT DATABASE MANAGEMENT: SHARING VITAL RECORDS DATA THROUGH THE GOVDATA DATA LAKE

Cláudio José Silva Ribeiro[113]
Ana Cristina Meirelles Velho[114]

4.1 INTRODUCTION

Information and knowledge use has boosted investigation projects, both in academic environments and business and govern organizations. We are witnessing the emergence of an informational and increasingly globalized (Castells, 1999).

It can be said that we are in the cyberspace era, where collaboration, instantaneous and digital fluency are increasingly present in everyday life (Barreto, 2014). Drive by the notion of "data avalanche", the efforts to gather information to support business management (*Business Intelligence – BI*) have shifted towards *Big Data/Analytics* solutions, causing reflections on the actions currently in development.

Since Castells, it is clear that the informational economy proposes the use of data and information to obtain results, so it appears that we are not facing a novelty in research in the field of information management. The Information Science field has promoted studies that explore this theme, especially in approaches for information assets management on the Web in managerial, tactical and operational levels (Velho, 2007; Ribeiro, 2008) and towards digital curation (Sayão; Sales, 2013).

The identification of standards for metadata, interoperability, sharing, archiving, access, reuse of collection, as well as the process and intelligent discovery of resources through ontologies and taxonomies, have become part of the information manager's concerns (Ribeiro, 2014; Sayão; Sales, 2013). In addition to this, the need to preform predictive analysis in datasets using statistical approach, besides group processing of big collections with *data mining* and simulation, promoted the displacement of the information manager in *BI* to *Analytics* activity (Velho, 2007; Siegel, 2013), materialized in *data lakes* and *Dataponds* structures (Inmon, 2016). Therefore, it is necessary to establish new paradigms of knowledge to deal with those extensive collections of intangible and virtual supplies (Levy, 1996).

The investigation is characterized as a descriptive, exploratory research with a qualitative approach, with bibliographic and documentary research and case study (Gil, 2002). Starting from a literature review on initiatives to

[113] PhD in Information Science. Graduated in Engineering. Associate Professor at Unirio where he works in the Department of Technical-Documentary Processes and in the Postgraduate Program in Library Science.

[114] Master in Information Science. Graduated in Systems Analysis. Business Intelligence Project Manager at DATAPREV

share government data, this report presents the objective, main components of GovData platform and a clipping for possible uses by different federal entities and public institutions, including Teaching and Research.

In addition to this introduction, this paper presents four more sections: the next one, with the theoretical framework used as reference, followed by the empirical field, the outcomes, and final consideration.

4.2 SHARING GOVERNMENT DATA USING *FAIR* PRINCIPLES

To better understand government data sharing, it is necessary to go back a few years to look for landmarks, for since the launch of the Transparency Portal in 2002, and subsequently the Access to Information Law (LAI – Lei de Acesso à Informação), in 2011, there was the crystallization of processes of dissemination of government information.

Driven by the movement of access to information catalyzed by the "Letter of Service to the Citizen" and by the *Memorandum on Transparency and Open Government* of the American government (Ribeiro; Almeida, 2011), the Brazilian government joined the *Open Government Partnership (OGP)*[115] initiative, improving even more the processes for dissemination, sharing, and reuse of information produced in the governmental sphere (CGU, 2012).

As part of the plans for the implementation of Open Government partnership, the Brazilian Open Data Portal has started, with dataset provision on government actions. As a result of a joint project between the government and the society, represented by standardization bodies, universities and non-governmental organizations, the portal was supported by actions linked to the National Open Data Infrastructure (Infraestrutura Nacional de Dados Abertos - INDA). This last effort presented a set of standards, technologies, procedures, and mechanisms of control necessary to meet the conditions for disseminating and sharing data and public information in the Open Data model, in accordance with the provisions of the Government Interoperability Project (e-Ping project) (Ribeiro; Almeida, 2011).

All these initiatives, directed to the use of technologies aiming at optimizing government internal processes, formed the concept of "electronic government". As of 2015, the government focus becomes "Citizen-centered", a global trend, and it is then called "digital government" (Brasil, [2018?]) .

Within the comprehensive concept of digital govern, the need to implement services in the course of Digital Transformation[116] paved the way for some structuring projects, among which GovData and ConectaGov stand out. They aim , among other things, at facilitating the exchange of information between the public institutions defined by the decrees 8.789 dated in 2016 and 10.046 dated in 2019 (Brasil, 2016; Brasil, 2019), therefore, the interoperability among information under the government institutions. GovData initiative focuses on the intero-

[115] International Initiative for governmental transparency and fight against corruption. Its launch was led by Brazil and the Unites States (OGP, 2011; CGU, 2012).

[116] Term used in Brazil and in Other countries and that encompasses government projects with the use of digital technologies aiming at increasing the State capacity to offer services to the citizens and data sharing.

perability through database and ConectaGov, through real-time data exchange. Within the scope of this report, the authors' concentration will be given in the evaluation of GovData databases.

4.2.1 GovData Initiative

Decree 8.789, dated June 29, 2016, established the following:

> The bodies and entities of direct and indirect public administration and other entities directly or indirectly controlled by the Union that are holders or responsible for the management of official databases will make available the access to data under their management to the bodies and entities of the direct, autonomous and foundational federal public administration interested in accessing data under their management, under the terms of this Decree (BRASIL, 2016, *online*).

Later, it was revoked and replaced by decree 10.046, dated October 9, 2019, which maintained the same guidelines relevant to this report, namely, the sharing of databases between the institutions and the creation of Central Committee for data Governance[117]. It can be seen that the orientation for sharing databases, in addition to raising awareness on efficient management, has acquired legal force.

The report by the Ministry of Planning, Government Transition 2018-2019, presented 15 structuring themes, including Digital Government. It supports the presentation of the scenario that led to the proposal of GovData as a solution to enable the continuous access to databases by the institutions, without requiring the permission protocols through an agreement that would be necessary for each action.

The history of Brazil in providing technological solutions at the service of society is in line with the world scenario, in which several countries have virtually approached the population via remote channels. Some European countries and the U.S. maintain strategies of Digital Transformation for more than a decade, according to ONU. Furthermore, in this preparatory report for, at the time, the future government that would take over the country in 2018, Denmark was the country of reference, for the amplitude of its initiatives, from fully digital services to data sharing and open government.

It should be noted that in 2020, Brazil ranked first in Latin America as provider of digital services and, in the Americas, it was second Only to the United States (Brasil, 2020).

GovData initiative had as its motivation to meet the need of public managers to define policies based on sufficient information, reducing the empiricism of such decisions (Brasil, [2018?]). Thus, if the State gathers relevant databases that can support analyzes to subsidize actions of its managers, then treating the dispersion of such bases and offer them in a viable way via technologies of access optimize the use of resources and provide more efficiency. This perception of existence of relevant information, not yet adequately available, and the lack of information by managers on the other side, form the core of the justification of this initiative.

[117] It is not the objective of this report to discuss possible similarities and differences between the terms data governance and data curation.

The interoperability of information in the federal government was the theme of the 4th Nation Forum of Union Transfer – Sharing, analyze and safety (Brasil, 2019a). At the time, the three components of GovData solution were presented: *data lake*, the access tools and the data science. *Data lake* corresponds to the data lake, a group of databases that, according to legal determination, must be available for access. The access tools (HUE, Qlik, RStudio and MicroStrategy) correspond to resources available so users can work with the databases, enabling from the production of large cross-checking data to dashboards. And, as a methodological support for the use of databases, the data Science techniques, to guarantee the expected results in the data analysis cycle.

In relation to the actors involved in GovData management, the Digital Government Secretariat of the Special Secretariat of Debureaucratization, Management and Digital Government of the Ministry of Economics is responsible for the role of Executive Office of the Data Governance Central Committee, which coordinates the activities of the (Brasil, 2019b). For issues that transcend the scope of data sharing, the Central Committee reports itself to the Interministerial Committee of Governance, established by decree 9.203, dated in 2017 (Brasil, 2017). Among the defined roles are those related to data holding (from data manager, data custodians, interoperability platform manager), and those related to the use (data receiver, data requester) (Brasil, 2019b).

When analyzing GovData features, it is possible to infer that the sharing assumptions pointed out by FAIR principles can be used with the aim of aligning its structures to the reuse of dataset proposed for C&T field.

4.2.2 *FAIR* Principles and the requirements for evaluation

The motivation for proposing principles for sharing and reusing data drove the formulation of *FAIR* principles (*Findable, Accessible, Interoperable and* Reusable). In this report, it is assumed that these principles are already disseminated, as recently, there have been different studies covering the *FAIR* theme in Brazilian universities' context, as observed in Henning *et al*. (2018); in Moreira *et al*. (2019); in Ribeiro (2019) and in Monteiro and Santana (2020).

In essence, *FAIR* principles aim at data interoperability and reuse. This is carried out by meeting the requirements below:

a. data and its metadata using persistent and universal identifiers;

b. data and its metadata represented by a formal, accessible, shared and widely applicable language for the representation of knowledge;

c. data and its metadata must have qualified references for other metadata and data. These elements and their relations need to be described semantically;

d. data and its metadata must have its origin indicated for use and reuse, as well as its transformation process and its history;

e. data and metadata must be clearly licensed;

f. use standards shared by communities.

These requirements were carefully analyzed considering initiatives for evaluation identified in (2019), Monteiro and Santana (2020) and Taco de Bruin *et al*. (2020). The list of questions used for the investigation was structured according to the Organizational, Digital Content and technological view.

The Organizational view covered aspect linked to the management infrastructure and data governance, including the existence of specific politics and profiles suitable for performing management activities

The Digital Content view covered aspects linked to data and metadata, identification strategies, semantic description and indexation.

The technological view covered aspects of standardization, technological infrastructure and communication protocol, in addition to collection and provision services.

In addition, when adopting *FAIR* principles, *GOFAIR* ([20–]) presents *FAIRification Process* as a way to enable institutions to align their *datasets* to those principles. This process is organized in the following stages:

a. gather and analyze datasets;

b. define and represent the semantic model for datasets;

c. link data;

d. verify data licensing;

e. establish metadata for the sets;

f. publish resources as *FAIR* data.

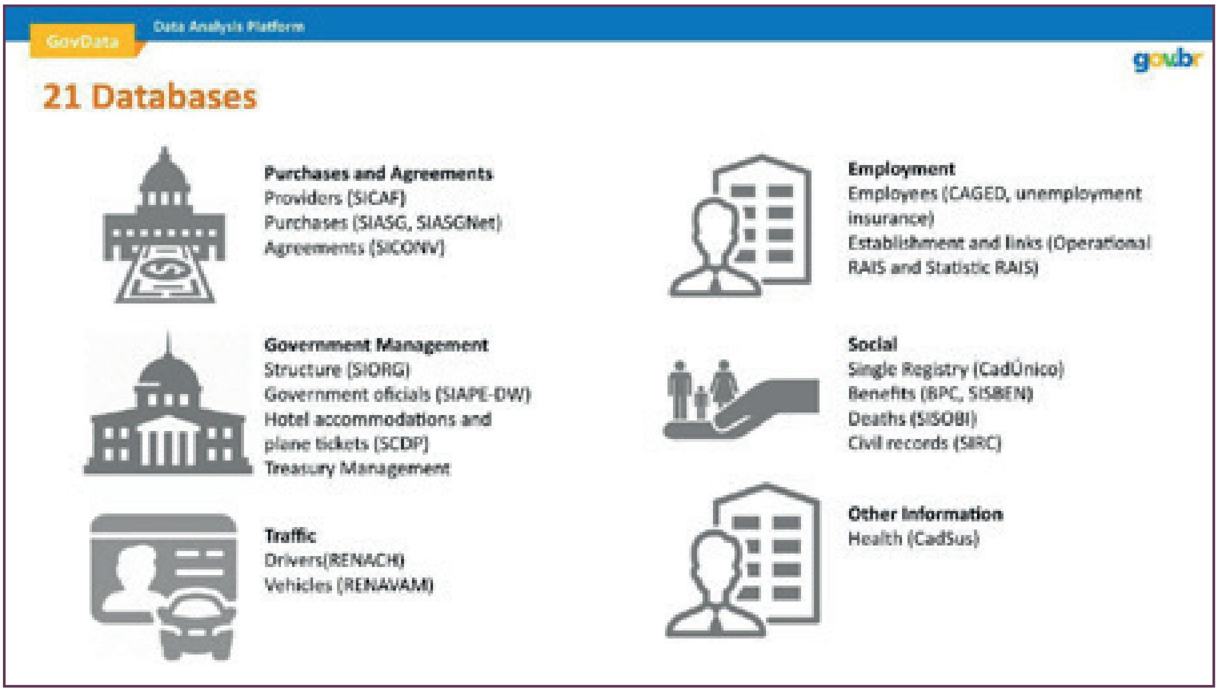
4.3 EMPIRICAL FIELD: SNAPSHOTS OF DATASETS ANALYZED

The Brazilian Social Welfare provides the citizens information about the access to its services, from appointments and orientation to confirmation and granting of benefits. The virtualization of access to these services via the Internet, mobile devices and call centers brings even greater enrichment to this topic. To fulfill its purpose, in addition to the information inherent to its service management, it manages large registers of social information, which are made up of databases of individuals and legal entities in the country, their civil events (birth, marriage, death), their working life and related events. These databases have their origin both in their sources and in external sources provided by other government institutions.

In face of the diversity and correlations between its information, its metadata management is part of its database curation.

At the aforementioned event, 4th National Forum of Union Transfer – Sharing, analysis and safety (Brasil, 2019a), 21 databases available in GovData platform were presented at the time, among which, the Civil Records (SIRC), database who's the analysis snapshot will be used.

Figure 1 – Database presentation



GovData

Data Analysis Platform

gov.br

21 Databases

Purchases and Agreements
Providers (SICAF)
Purchases (SIASG, SIASGNet)
Agreements (SICONV)

Employment
Employees (CAGED, unemployment insurance)
Establishment and links (Operational RAIS and Statistic RAIS)

Government Management
Structure (SIORG)
Government oficials (SIAPE-DW)
Hotel accommodations and plane tickets (SCDP)
Treasury Management

Social
Single Registry (CadÚnico)
Benefits (BPC, SISBEN)
Deaths (SISOBI)
Civil records (SIRC)

Traffic
Drivers(RENACH)
Vehicles (RENAVAM)

Other Information
Health (CadSus)

Source: Brasil (2019a).

The National Civil Registry Information System (Sistema Nacional de Informações de Registro Civil – SIRC) is the digital means to obtain civil data on birth, marriage, deaths and stillbirth, which are sent by the registry offices aiming, among others, at eradicating the under-registration in the country, qualify other governmental databases, subsidize public policies and help reduce frauds in granting of benefits and crimes as falsification and human trafficking (Brasil, [2019?]). GovData presents the following data collection:

a. births: identification of the individual, date, and place of birth, date and place of registry, place of residence, gender, identification of family relation;

b. deaths: identification of the individual, nationality, date and place of death, date and place of registry, marital status, gender, identification of names of parents, cause of death, date and place of birth, place of residence, identifier of the social security benefit, occupation;

c. marriage: identification of spouses, nationalities, date and place of the ceremony, date and place of registry, name of spouses' parents, system of marriage, information on the religious marriage, place of residence, information on the dissolution of marriage, occupations;

d. history: previous versions of birth, death, and marriage certificate that have been altered in some way;

e. operational: collection data encoding (for instance, code and description of the system of marriage), services/ registry office register, data on the operations of file upload to feed the database.

Data collection related to birth and death certificates were used for analyzes, due to the need to snapshot this report as well as its importance as acts that represent the demography of the country.

This investigation did not aim at verifying the correlation of information available to the legislation relevant to the Civil Registry in the country. Nevertheless, the importance of future studies with this aim is recognized.

The description of dataset available on birth and death certificate is available with the name of the attribute, its format and extended description, therefore, technical metadata. Information on its update and its custodian are also available. There is no additional information on related data, not being possible to infer if there is no other relations or if it is only its omission (CKAN, [201-]).

In relation to present the dataset, it is possible to verify that the last update occurred more than one year ago, possibly indicating its deactivation. However, it is understood that this fact does not compromise the analysis proposed in this article, since it is an analysis of the implemented initiative.

The available datasets available on the selected themes, births and deaths, have a significant degree of use for understanding these phenomena and its dimensions. These are official administrative records that can support population studies together with other sources from institutions that are recognized in the country, such as IBGE (IBGE, [201-]). Data related to the place of events can show the population movement in relation to the birthplace and place of residence. The occupation, age, and gender according to the place of residence have the statistical potential for historical analysis of deaths. The cause of death, age, gender, and place constitute a group of data that, when combined, can generate indicators of interest to deepen the reality.

4.4 OUTCOMES

Datasets were analyzed considering what was presented in Section 2. The datasets were categorized and gathered for analysis by the themes described. Table 1 gathers the main considerations, resulting from the evaluation process.

Table 1 – Analysis of the compliance to requirements by GovData

View	Requirement	Analysis
Organizational	Organizational and personnel structure defined	Yes. *Datasets* analyzed are protected by Dataprev. The company has specific area for data governance.
	Roles and responsibilities defined	Yes. *Datasets* analyzed are protected by Dataprev. The company has specific area for data governance.
	Current policies include *FAIR* principles	No.
	Teams dedicated to data management, metadata, and data science.	Yes. *Datasets* analyzed are protected by Dataprev. The company has specific area for data governance and for activities related to data science.
Digital Content	Data and metadata with persistent identifier	Partially. Data and metadata identifiers cannot be framed in the concept of persistent ID.
	Metadata enrichment	Partially. The metadata model follows the established by CKAN tool.
	Data and metadata retrievable by the identifier and with standardized, open and free protocol.	Yes. HTTP. Especially for metadata - DCAT (CKAN).
	Metadata available, even after datasets are removed.	There was no evidence.
	Data and metadata include references qualified for Other elements (data and metadata)	There was no evidence.
	Data and metadata are licensed	Metadata licensed with Creative Commons. There is no indication for data.
	Data and metadata with detailed provenance	It does not have.

View	Requirement	Analysis
Technological	The protocol allows procedure of authorization and authentication, when necessary.	Yes.
	Data and metadata with formal, accessible, shared and applicable language for knowledge representation.	DCAT/RDF metadata with Harvesting Java protocol and comparable to OAI-PMH. REST architecture. Data with HUE/Hadoop and RStudio.
	Data and metadata use vocabularies that follow *FAIR* principles	No
	Data and metadata meet domain standards	Partially. They follow local standards to the custodial institutions, but the alignment with the VCGE[118] was not recognized.

Source: elaborated by authors (2024).

The outcome obtained in the context of the organizational view was satisfactory, since the institutions (Dataprev and Serpro) that manage the datasets have high specialization, experience and technological resources for the adequate governance.

There is also a need to prioritize projects that deal with the vision of digital content, since the requirements linked to the metadata representation and the use of semantic models need to be met in its totality to allow the adequate reuse of sets available. Another important requirement is linked to the use of persistent identifiers for accessing data and metadata. The debate about the generation of identifiers for research data can be extended to encompass government datasets as well.

Similar to the organizational view, satisfactory results were obtained for the technological view. Only the lack of alignment with the VCGE was pointed out. According to Ribeiro and Pereira (2015), this vocabulary was created in 2011 and could be in use to improve data semantics.

Finally, after cross-check the requirements in Table 1 with the datasets under analysis, and based on *Fairification Process* presented in section 2.2, it as verifies that it can be possible to search for alignments with *FAIR* principles. The datasets analyzed and described in section 3 demonstrate the wealth of the relations that can be built from the information of the civil registry of births and deaths registered in the country's registry offices.

[118] VCGE - Vocabulário Controlado de Governo Eletrônico (Controlled Vocabulary of Electronic Government). Available from: https://www.gov.br/governodigital/pt-br/governanca-de-dados/vocabulario-controlado-do-governo-eletronico. Access on: 15 Oct. 2020.

It is understood that the datasets require some level of anonymization to prevent the identification of individuals, but this transformation is possible in face of the inputs already available and it does not reduce the potential that these databases have to produce new information.

4.5 FINAL CONSIDERATIONS

Data sharing is essential in the development of collaborative actions for research increase. Sharing data is essential in the collaborative development of actions to increase research. Discussing a topic of recent interest, the increase in the speed to obtain results is explicit when analyzing the context of projects linked to the development of coronavirus vaccine (SARS-CoV-2 - COVID-19) and in particular the VODAN project[119].

Reusing data is key element in this increment, therefore, the descriptions in metadata and semantic models are essential items for proper understanding available datasets.

In addition to the context of Science and Technology, it is possible to infer that *FAIR* principles are also applicable in data sharing and government information context. In that matter, the efforts mentioned in this report can serve as a starting point for better data and information dissemination for the organized Society and Teaching and Research institutions.

At the time of writing this report, GovData platform is in the institutional site of the Economy Ministry with its three components already mentioned [see 2.1], however, with no reference to databases available in *data lake*. This fact seems to indicate to the authors that there is priority in offering technological resources that provide access and analysis of data Science than the availability of government datasets.

Finally, this report aims at incorporating to the scientific debate the possibility of engaging Society in the actions of the open government and in the development of open science. It is possible to infer that it is a long path; the first steps taken toward *Citizen Analyst* (Allemang, 2010; Ribeiro; Almeida, 2011) can now be followed searching for *Citizen Data Scientist* (Banker, 2018).

REFERENCES

ALLEMANG, D. **Modalities:** The Magazine of Semantic Web. n. 9. 2010. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.190.216&rep=rep1&type=pdf. Access on: 7 Oct. 2020.

BANKER, S. The Citizen Data Scientist. **Forbes**, New York, 21 Jan. 2018. Innovation. Available from: https://www.forbes.com/sites/stevebanker/2018/01/19/the-citizen-data-scientist/. Access on: 2 Oct. 2024.

BARRETO, A. A. A aventura de perceber significados. **DataGramaZero**, Rio de Janeiro, v. 15, n. 3, jun. 2014.

[119] *Virus Outbreak Data Network* - project based on *FAIR* network for sharing data about COVID-19. Available from: https://www.go-fair.org/implementation-networks/overview/vodan/. Access on: 7 Oct. 2020.

BRASIL. **10 - Governo Digital** [2018?] Available from: https://transicao.planejamento.gov.br/wp-content/uploads/2018/11/10_Governo-Digital_versão_para_publicação.pdf. Access on: 2 Oct. 2020.

BRASIL. **Decreto nº 10.046, de 9 de outubro de 2019.** Dispõe sobre a governança no compartilhamento de dados no âmbito da administração pública federal e institui o Cadastro Base do Cidadão e o Comitê Central de Governança de Dados. Brasília, DF: Presidência da República, 2019b. Available from: http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2019/decreto/D10046.htm. Access on: 10 Oct. 2020.

BRASIL. **Decreto nº 8.789, de 29 de junho de 2016.** Dispõe sobre o compartilhamento de bases de dados na administração pública federal. Brasília, DF: Presidência da República, 2016. Available from: http://www.planalto.gov.br/CCIVIL_03/_Ato2015-2018/2016/Decreto/D8789.htm. Access on: 2 Oct. 2024.

BRASIL. **Decreto nº 9.203, de 22 de novembro de 2017.** Dispõe sobre a política de governança da administração pública federal direta, autárquica e fundacional. Brasília, DF: Presidência da República, 2017. Available from: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2017/Decreto/D9203.htm. Access on: 2 Oct. 2024.

BRASIL. **IV Fórum Nacional das Transferências da União – Compartilhamento, análise e segurança.** Brasília, DF: Ministério da Economia, 2019a. Slide Share. 43 slides. Available from: https://www.gov.br/plataformamaisbrasil/pt-br/ajuda/Apresentacoes/arquivos-e-imagens/governanca_de_dados_-_sgd_me-1.pdf. Access on: 2 Oct. 2024.

BRASIL. Brasil em primeiro lugar na América Latina. **Govbr**. Brasília, 22 Jul. 2020. Available from: https://www.gov.br/pt-br/noticias/financas-impostos-e-gestao-publica/2020/07/brasil-esta-entre-os-20-paises-com-melhor-oferta-de-servicos-digitais. Access on: 2 Oct. 2024.

BRASIL. **SIRC** – Sistema Nacional de Informações do Registro Civil. [*S. l.: s. n.*], [2019?]. Available from: https://sirc.gov.br. Access on: 2 Oct. 2024.

CASTELLS, Manuel. **A Sociedade em rede**. São Paulo: Paz e Terra, 1999.

CGU. DEFESA – Gestão da informação será o foco da Defesa no "Governo Aberto". **Govbr**, Brasília, 5 Dec. 2012. Available from: https://www.gov.br/defesa/pt-br/assuntos/noticias/ultimas-noticias/05-12-2012-defesa-gestao-da-informacao-sera-o-foco-da-defesa-no-governo-aberto. Access on: 2 Oct. 2024.

CKAN. **CKAN-GOVDATA**. [*S. l.: s. n.*], [201-]. Available from: https://ck.govdata.gov.br/. Access on: 9 Oct. 2020.

GIL, A. C. **Como elaborar projetos de pesquisa.** 4ª ed. São Paulo: Atlas, 2002.

GOFAIR. **Fairification process**. [*S. l.: s. n.*], [20--]. Available from: https://www.go-fair.org/fair-principles/fairification-process/. Access on: 15 Oct. 2020.

HENNING, P. C. *et al*. Desmistificando os princípios fair: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados fair. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 19., 2018, Londrina. **Proceedings** [...]. Londrina: UEL, 2018. Available from: http://hdl.handle.net/20.500.11959/brapci/103506. Access on: 2 Oct. 2024.

IBGE. **População**. [*S. l.: s. n.*], [201-]. Available from: https://www.ibge.gov.br/estatisticas/sociais/populacao.html. Access on: 2 Oct. 2024.

INMON, B. **Data Lake architecture:** Designing the Data Lake and avoiding the garbage dump. New Jersey: Technics Publications, 2016.

LEVY, P. **O que é virtual?** São Paulo: Ed. 34, 1996.

MONTEIRO, E. C. S. A.; SANTANA, R. C. G. Repositórios de dados científicos na infraestrutura de pesquisa: adoção dos princípios fair. **Ciência da Informação,** Brasília, v. 48, n. 3, 2019. Available from: http://hdl.handle.net/20.500.11959/brapci/136407. Access on: 2 Oct. 2024.

MOREIRA, J. L. *et al*. Towards findable, accessible, interoperable and reusable (fair) data repositories: improving a data repository to behave as a fair data point. **Liinc em revista,** Rio de Janeiro, v. 15, n. 2, 2019. DOI: 10.18617/liinc.v15i2.4817. Available from: https://revista.ibict.br/liinc/article/view/4817. Access on: 2 Oct. 2024.

OGP. Declaração de governo aberto. **Open Government Partnership**, Sept. 2011. Available from: www.opengovpartnership.org/open-government-declaration. Access on: 2 Oct. 2024.

RIBEIRO, C. J. S.; ALMEIDA, R. F. Dados Abertos Governamentais (Open Government Data): instrumento para exercício de cidadania pela sociedade. *In:* ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 12., 2011, Brasília. **Proceedings** [...]. Brasília: UnB, 2011.

RIBEIRO, C. J. S. Diretrizes para o projeto de portais de informação: uma proposta interdisciplinar baseada na Análise de Domínio e Arquitetura da Informação. 2008. 298 f. Tese (Doutorado em Ciência da Informação) - Instituto de Arte e Comunicação Social, Universidade Federal Fluminense / IBICT, 2008.

RIBEIRO, C. J. S. Big Data: os novos desafios para o profissional da informação. **Informação & Tecnologia (Itec),** João Pessoa/Marília, v. 1, p. 96-105, 2014.

RIBEIRO, C. J. S. Digital repositories maturity model: a way to its adoption in research data management. **Liinc em revista,** Rio de Janeiro, v. 15, n. 2, 2019. DOI: 10.18617/liinc.v15i2.4816. Available from: https://revista.ibict.br/liinc/article/view/4816. Access on: 2 Oct. 2024.

RIBEIRO, C. J. S.; PEREIRA, D. V. A publicação de dados governamentais abertos: proposta de revisão da classe sobre Previdência Social do Vocabulário Controlado do Governo Eletrônico. **Transinformação,** Campinas, v.

27, n. 1, p. 73-82, Apr. 2015. DOI: 10.1590/0103-37862015000100007. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862015000100073&lng=en&nrm=iso. Access on: 2 Oct. 2024.

SAYÃO, L.; SALES, L. DADOS DE PESQUISA: contribuição para o estabelecimento de um modelo de curadoria digital para o país. **Pesquisa Brasileira em Ciência da Informação,** João Pessoa, v. 8, n. 2. 2013. Available from: https://pbcib.com/index.php/pbcib/article/view/18634. Access on: 2 Oct. 2024.

SIEGEL, E. **Predictive Analytics.** New Jersey: John Wileu & Sons, 2013.

TACO DE BRUIN *et al.* Do I-PASS for FAIR. A self assessment tool to measure the FAIR-ness of an organization. **Zenodo**, [*s. l.*], 2020. DOI: 10.5281/zenodo.4080867. Available from: https://zenodo.org/records/4080867. Access on: 2 Oct. 2024.

VELHO, A. C. M. **A tomada de decisão na Previdência Social: uma reflexão das ações do produtor de informações da Dataprev.** 2007. 137 f. Dissertação (Mestrado em Ciência da Informação) – Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal Fluminens, Rio de Janeiro, 2007.

How to cite this chapter: RIBEIRO, Claúdio José Silva; VELHO, Ana Cristina Meirelles. FAIR principles ans government databse management: sharing vital records data through the GovData data lake. *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. 4, p. 58-70. DOI: 10.22477/9788570131959.cap4.

5. OPEN DATA OF THE LATTES PLATFORM ACCORDING TO FAIR PRINCIPLES: EXAMPLES OF EXTRACTOR AND UFSC INFORMATION OBSERVATORY

Adilson Luiz Pinto[120]
Thiago Magela Rodrigues Dias[121]
Fábio Lorensi do Canto[122]
Washington Luís Ribeiro de Carvalho Segundo[123]

5.1 INTRODUCTION

The term *Open Access* (OA) is a concept related to the free access to scientific information on the Internet, especially peer reviewed scientific papers or papers published in a specialized scientific magazine. Open access is based on the premise that the scientific research is mostly financed with public resources; therefore, its results should be available and accessible with no cost to society. It considers that researchers do not write for financial reasons, but to maximize the visibility, use, and impact of their research results. The open access movement also argues that, although the process of editing and disseminating an article involves costs, these must be incorporated into the general costs of the research and not passed on to readers. (Shavell, 2010; Freire, 2011).

In the late 1990s, there were several demonstrations in favor of open access. Among the reasons that drove the creation of this movement, the scientific journal's price crises stands out, a phenomenon that limited or prevented the access to scientific information from countries and institutions lacking resources to pay for subscriptions and licenses. Alternatives were sought to provide a broader access, forming consortia to acquire content to be available in portals and databases (Fladung, 2007).

With the development of tools for building repositories and databases, the open access model gains consistency. Therefore, several declarations in favor of this model are published, intensifying the implementation of a basic infrastructure of open access in national and international levels (Kuramoto, 2006).

Open access drives the return of efforts put on researches with public investment, making the results more accessible. Regardless of the meanings the term contains, open access must be discussed based on different aspects, among which access to literature or the knowledge in it stands out (SUBER, 2007). It is important to note

[120] Author details: PhD in Documentation, Professor and Researcher at the Federal University of Santa Catarina (PGCIN/UFSC), adilson.pinto@ufsc.br

[121] Author details: PhD in Mathematical and Computational Modeling, Professor and Researcher at the Federal Center for Technological Education of Minas Gerais, thiagomagela@cefetmg.br

[122] Author details: PhD student in Information Science at PGCIN/UFSC, fabio.lc@ufsc.br

[123] Author details: PhD student in Computer Science (UnB), Researcher at the Brazilian Institute of Information in Science and Technology, washingtonsegundo@ibict.br

that open access to scientific knowledge refers to both formal and informal aspects of the scientific communication process (Leite, 2016).

Recently, open access guidelines have also been applied to data management plans, since those are instruments that guide practices to promote accessibility and reuse of research data. To make data more findable, accessible, interoperable and reusable, FAIR principles are used, an acronym for *'Findable', 'Accessible', 'Interoperable' and 'Reusable'* (Veiga *et al.*, 2019).

Currently, FAIR principles are considered the guiding elements for good practices in the whole research data management process. They aim at implementing a metadata set defined to be used by both automated computational mechanisms and by people. If properly adopted, FAIR principles facilitates the interoperability among different data environments (Henning *et al.*, 2019).

FAIR principles have been part of discussions and contemporaneous practices of data science since the beginning of 2014. They had their application consolidated in 2017, when the European Commission required that data management plans be adopted based on those principles in projects funded by its resources. Since then, the principles have been used to guide the discovery, access, interoperability, sharing, and reuse of research data (Henning *et al.*, 2018).

In Veiga *et al.* (2019) it is possible to find a summary of the FAIR principles:

Findable: (a) data and metadata need to have a single persistent identifier; (b) data should be described with rich metadata; (c) have the persistent identifier for the dataset described in the metadata, and; (d) metadata and data must be retrievable through trustworthy repositories;

Accessible: (a) data and metadata must be retrieved by its identifier using standard communication protocols; (b) the protocols must be free, open and support authentication and authorization, and; (c) metadata must be accessible even when data is not available anymore.

Interoperable: (a) data and metadata must be coded using agreed standards of representation, and; (b) data and metadata must use vocabulary aligned to FAIR principles and include relevant references.

Reusable: (a) data and metadata need to be associated to relevant attributes; (b) data and metadata must be released with use license clearly defined; (c) metadata and data must be associated to their origin in a detailed way, and; (d) data and metadata must meet the community standards.

From the consolidation of these four principles in the context of open data and open science, it is presented one of the forms through which the Federal University of Santa Catarina (UFSC) have used data available in Lattes Platform in its internal management systems.

5.2 METHODOLOGY

The data source defined for this work was the UFSC faculty résumé database registered in Lattes Platform (2,581 UFSC permanent faculty in May 2020). Résumé in the Lattes system were chosen because they have a lot of information. It is a resource that allows the integration of scientific, professional and academic data, and data update can be done permed by research. Among the main information in the résumé, academic qualification, research field, professional performance and academic orientation are those that stand out, in addition to technical and scientific productions.

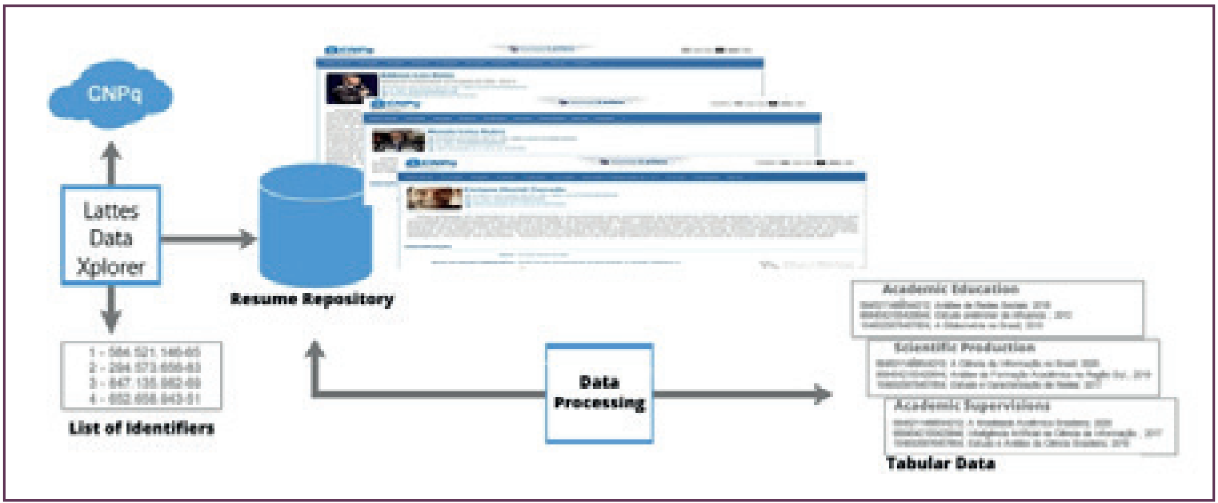
The Résumé in Lattes system became a national standard used for individual evaluation of scientific and academic activities. They add data from researchers from all fields of knowledge, making the platform a relevant source for analysis and understanding of research groups behavior (Digiampietri *et al.*, 2012).

Although the data from the résumé is freely available, they can only be viewed individually through a query interface provided by the CNPq. However, this interface is limited with no possibility of grouping, analysis and comparisons with other résumé. That said, techniques and tools for extracting data are necessary for analyzing large sets of curriculum data.

LattesDataXplorer framework (Dias, 2016) was used for extracting and treating data. It is a tool developed with the aim of collecting and treating curriculum data from Lattes Platform, with low computational cost.

LattesDataXplorer is responsible for encompassing the whole set of techniques and methods for collecting, treating and analyzing data used herein. The extraction is performed by a component that searches and retrieves each faculty résumé from the single indicator in their résumé in Lattes Platform. Consequently, with all résumé stored in XML format, the institution handles its data in UFSC information Observatory (https://observatoriodainformacao.ufsc.br/indicadores-cnpq/ufsc/) making data management possible for the Dean of Research and Graduate Studies of the Institution.

Figure 1 - Model of the system used by UFSC for extracting and treating data coming from the Lattes Platform



Source: Research data (2020).

With *LattesDataXplorer* it is possible to group a set of resumes based on predefined parameters. In the process of selecting résumé based on parameters, regardless of the section in which they are found, they are selected and grouped for analysis. Data is organized in a list of selected resumes, which would not be possible without the strategy adopted.

Subsequently, having as input a specific list of resumes obtained through a specific query or even through using a global listing with the entire local resume repository, it is possible to process data with specific computational routines for each type of analysis. This processing aims at extracting relevant information from the resumes and grouping them in preprocessed data files. Such a strategy aims at generating sets of specific data for application of metric analysis, making it no longer necessary to access the entire set of resumes in each new analysis. Thus, resumes that have a large quantity of data to be processed are accessed and treated only once. As examples of preprocessed data files, there are the ones that gather information about orientations, academic qualification, collaboration and scientific production, as well as about research projects and technical production. All data is available in tabular form.

5.3 OUTCOMES

The main results can be seen at UFSC Information Observatory - https://observatoriodainformacao.ufsc.br/indicadores-cnpq/ufsc/[124]-, an environment that presents historical data from on the institution, which are: (a) indicators that are cross-referenced with data from CNPq, such as scholarship holders of Research Productivity, Technological Development, Research Groups, Research Technical Support and Technical Support Extension; (b) UFSC general indicators (2012-2018), such as Distribution of scientific/technical/artistic productivity and orientation, Scientific production by typology, Technical production by typology, Artistic production and Orientations; (c) Technological indicators (2000-2018), such as patents in general, Thematic patents, Patents by large areas, Patents by inventor, Patents by repository country, Collaborative patents by areas, Patents by institutions collaboration and Patents by areas; (d) indicators by departments per person between 2012-2018; (e) indicators of visibility in Web of Science database (2012-2018), aiming at Typologies of publication, Areas of publication, partner institutions; Main Sponsors, partner countries and most notable authors.

However, this study aims at the feasibility of this entire set of information in open access based on FAIR principles, in which it is intended to determine in which way the Extractor and UFSC Information Observatory datasets are represented

Meeting the ***Findable*** principle, it was possible to identify a context of unique data, the Unique resume identifier (http://lattes.cnpq.br/4767432940301118). Another aspect to be pointed out is that these data have several resources and features, such as tracking system, training levels and work functions, levels of production typologies, either scientific, technical or artistic, as well as tutorials in all training levels. The permanent identification of the dataset can be retrieved

[124] The Observatory is managed by SETIC/UFSC and because it is in constant maintenance during COVID-19 pandemic, it is suggested to right-click to open the link in a new tab to access the spreadsheet. Thus, any maintenance problem of the system is solved, and the content can be accessed.

any moment, since the character sets are unique and constantly surveyed, in which no researcher can have two ways of input. The key point is that the system is updated by this entry point Every weekend.

For the ***Accessible*** principle, the system developed by the Federal University of Santa Catarina from the Extractor and the Information Observatory shows that even if data is retrieved, a standard must be kept for identifying protocols in the resumes in the Lattes system (4767432940301118), which is also open access. Access to data indexing is exclusive to researches through their resumes; however, data extraction is free, especially as it is a government record, maintained by the Ministry of Science and Technology. Finally, data are available even if researches do not update them, including in the case of death.

Regarding ***Interoperable***, data is a set of metadata that each researcher adds to his/her resume. Some fields may use filters. In addition, as they are formatted using the XML standard, the interoperability with other datasets is facilitated. Several other systems were developed for similar treatments in Brazil, such as *Script Lattes* (Mena--Chalco; Cesar Junior, 2009).

For ***Reusable*** data, UFSC provides a system for data gathering in XML format, which goes to a systematic extractor. This system organizes the information according to each professor's functional data, not worrying about overlapping of departments at first. Later, institutional overlapping is possible. Data using license is from the government, so as they are data of national need, it is required to be updated for possible evaluations (Productivity Scholarship, Public notice projects, among others).

5.4 DATA FUNCTION FOR INSTITUTIONAL MANAGEMENT CONTROL

The Federal University of Santa Catarina, as a higher education teaching, research, and extension institution, needs to produce indicators of its scientific, technical, social and internationalization development, as well as follows the historical evolution of these contents. For this reason, there has been investment in projects of this nature, whether for extraction and even for its applicability in new services, research, and products in which it is worth investing.

Considering this process, the actions performed are aimed at generating skills and identifying specialists. The resource presented herein works as a basis for finding talents in the institution through resumes registered on Lattes Platform and on other open platforms.

For the Deans of Graduate Studies and Research, this directory of data helps to identify the specialists of the institution and even the collaborators of certain study themes. As example of practical application of this resource, the case of COVID-19 pandemic is mentioned, making it possible to verify the internal researchers and their main collaborators in the development of public health, health safety and even data science. There are other levels that can also be explored, such as issues of guidance in undergraduate, master's and doctoral levels.

There are four levels of content that can be explored to condense in a context of identifying talents through resumes in the Lattes Platform, such as (i) specialists in scientific productivity; (ii) specialists in guidance and participation in thesis/dissertations in themes/subjects examining commission; (iii) leaders of research groups, and; (iv) specialists in technological production.

The identification of this scenario can be seen from a general framework, combining all possible identifications, aforementioned in the previous paragraph, giving a percentage margin for each one. For instance, 25% for each item or studying which ones are more important and dividing the percentages according to the relevance of each item studied.

However, regardless of the order or more relevant item, it is possible to identify particularities in each one of the items in this possible talent pool. This is because they are traceable and accessible data, which have a standard use interoperability, which can be reused to determine department standards and research centers, as in a system for generating ranking and per person indicators.

When it comes to verification from orientation and participation in theses and dissertation examining commission, it can be based on absolute numbers of orientation, participation in examining commissions or on the relation between both.

The process to achieve this particularity of analysis has the advantage of being able to identify scientific families, such as identifying the professors who manage to guide their advisees since undergraduate course, going through the Master's program until doctorate. This can also be applied to the specialists who manage to guide their advisees in higher academic levels, verifying if the advisees were granted scholarships during this period (Costa; Pinto, 2016).

Following this line of reasoning, we have the open content of the Directory of Research Groups of CNPq, which holds a large data collection about the development of research groups, their participants, and collaborators and finally, the most valuable item, the leaders of the research groups. That information can be useful for identifying specialists in fields, themes, and subjects.

This data can be used separately, specifically to identify the researchers' profiles, as it can be blended in the content of the researchers' resumes.

This fusion results in other items in analysis, which deal exclusively with the productive dynamic of researchers in scientific issues (journal articles, works presented in records of events, published/edited books and chapters of books), as well as the technological issue (patents, technological models, graphic designer, among others).

The scientific capital can also be seen by identifying the most cited authors within their respective fields, providing more guidance in the context of expertise for the field, theme, or subject. The citation index can be generated through databases available in Capes Journal Portal (free for federal systems) or through Academic Google.

Finally, for the identification of these specialists and for building this talent pool, it is possible to verify the researchers' performance in open access publications, as well as in commercial contents such as magazines indexed in databases. The focus of this means of data verification is also to identify if the studies of a group of researchers have adherence at the internationalization level.

5.5 FINAL CONSIDERATIONS

This type of service is essential for UFSC, as well as for any other education institution, and its importance can be summarized in six fundamental points:

(1) It is used as support for Graduate Programs to follow, in real time, the evolution of scientific, technical, and artistic data, and the orientation of its professors. It is worth noting that the Information Observatory also provides services aimed at monitoring the development of a certain Graduate Program in the perspective of collaborators with the other programs in the same field in Brazil, as explained in the article "A bibliometric analysis of the scientific production and collaboration between graduate programs in manufacturing engineering in Brazil" that monitored the main Graduate Programs in Production Engineering from 2008 to 2017 (Dutra *et al.*, 2019);

(2) Explores and interoperates with FAIR applications within the institution, accessing valuable data for planning graduate programs for the Dean's office, as an alternative to add information and indicators of C&T in the Search portal;

(3) Similar to otter Lattes Platform *Scripts* extraction, it is accessible, easy to handle and interoperable. From resumes in SML format, it can reuse data, as seen in UFSC Extractor and Observatory;

(4) It works as informational basis to the Research Dean, and to the project of developing a monitoring laboratory in C&T at UFSC, inside the Information Observatory,

(5) Data generated by both UFSC Extractor and Observatory can work as reference to other public institutions, concerning monitoring per person of researchers' productivity and visibility, since UFSC is the best ranked in citation per capita in the country, when considering the elimination of auto citations, and;

(6) Finally, it is reinforced that the organization of data collected from Lattes Platform in compliance with FAIR principles works as a model for the construction of other systems and services that allow an easy information retrieval, as well as the projection of new indicators on C&T of an institution, its levels of national and international collaboration, and the provision for automatic collection of raw data openly gathered; whenever there is no legal restriction.

REFERENCES

COSTA, Airton; PINTO, Adilson Luiz. **De bolsista a cientista:** a experiência da UFSC com o Programa de Iniciação Científica no processo de formação de pesquisadores (1990 a 2012). Florianópolis: EdUFSC, 2016.

DIAS, Thiago Magela Rodrigues. **Um Estudo Sobre a Produção Científica Brasileira a partir de dados da Plataforma Lattes**. 2016. 181 p. Thesis (Doctorate in Mathematical and Computational Modeling) - Graduate Program Course in Mathematical and Computational Modeling, Federal Center for Technological Education of Minas Gerais, Belo Horizonte, 2016.

DIGIAMPIETRI, Luciano Antônio *et al*. Minerando e caracterizando dados de currículos lattes. *In:* BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2., 2012, Curitiba. **Annals [...]** Curitiba: Brasnam, 2012.

DUTRA, Silvana Toriani *et al*. A bibliometric analysis of the scientific production and collaboration between graduate programs in manufacturing engineering in Brazil. **Informação & Sociedade**, v. 29, n. 1, p. 117-136, 2019. Available from: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/44852. Access on: Sept. 27, 2020.

FLADUNG, Rainer. ***Scientific communication:*** economic analysis of the eletronic journal market. Stuttgart: Ibidem-Verlag, 2007.

FREIRE, José Donizetti. **CNPq e o acesso aberto à informação científica**. 2011. 275 p. Thesis (Doctorate) - Graduate Program Course, Faculty of Information Science, University of Brasília, Brasília, 2011.

HENNING, Patrícia Corrêa et al. Desmistificando os Princípios FAIR: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos Dados FAIR. *In:* XIX Encontro Nacional de Pesquisa em Ciência da Informação, 19, 2018, Londrina. **Annals [...]** Londrina: UEL, 2018.

HENNING, Patrícia Corrêa et al. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, v. 25, n. 2, p. 389-412, 2019. DOI: https://doi.org/10.19132/1808-5245252.389-412. Available from: https://seer.ufrgs.br/index.php/EmQuestao/article/view/84753. Access on: Sept. 27, 2020.

KURAMOTO, Hélio. Informação científica: proposta de um novo modelo para o Brasil. **Ciência da Informação**, v. 35, n. 2, p. 91-102, 2006. Available from: https://www.scielo.br/j/ci/a/RcPCvVSyQ6dx7RcmJFLnbxL/abstract/?lang=pt. Access on: Sept. 27, 2020.

LEITE, Fernando César Lima. **Gestão do conhecimento científico no contexto acadêmico: proposta de um modelo conceitual**. 2016. 240 p. Dissertation (Master in Information Science) - Department of Information and Documentation Science, University of Brasilia, Brasília, 2016.

MENA-CHALCO, Jesús Pascoal; CESAR JUNIOR, Roberto Marcondes. ScriptLattes: An open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31-39, 2009.

SHAVELL, Steven. Should copyright of academic works be abolished?. **Journal of Legal Analysis**, n. 1, v. 2, p. 301-358, 2010. Available from: http://jla.oxfordjournals.org/content/2/1/301.short. Access on: Oct. 01, 2020.

SUBER, Peter. **Open Access Overview: focusing on open access to peer-reviewed research articles and their preprints**. Creative Commons, 2007. Available from: http://legacy.earlham.edu/~peters/fos/overview.htm. Access on: Aug. 9, 2020.

VEIGA, Viviane Santos de Oliveira *et al*. Plano de gestão de dados fair: uma proposta para a Fiocruz. **Liinc em Revista**, v. 15, n. 2, p. 275-286, 2019. DOI: https://doi.org/10.18617/liinc.v15i2.5030. Available from: https://revista.ibict.br/liinc/article/view/5030. Access on: Aug. 9, 2020.

How to cite this chapter: PINTO, Adilson Luiz; DIAS, Thiago Magela Rodrigues; CANTO, Fábio Lorensi do; CARVALHO SEGUNDO, Washington Luís Ribeiro de. Open data of the lattes platform according to FAIR principles: examples of extractor and UFSC information observatory. *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. 5, p. 71-79. DOI: 10.22477/9788570131959.cap5.

6. ANALYSIS OF THE DATASETS AVAILABLE IN THE COVID-19 DATA SHARING/BR REPOSITORY IN CONFORMANCE TO THE FAIR PRINCIPLES

Anderson Rafael Castro Simões[125]
Renata Lemos dos Anjos[126]
Guilherme Ataíde Dias[127]

6.1 INTRODUCTION

The constant use of Digital Information and Communication Technologies (DICTs) by various individuals and sectors of society, including the academic-scientific community, contributes to a growing and continuous production of data. The academic-scientific scenario is configured both as a large producer and consumer of these, which are now considered primary sources for new scientific investigations, providing the development of science.

In this perspective, Sales *et al*. (2020) affirm that the advancement of science in various areas of knowledge is strongly linked to the reuse of scientific data, which points to a demand for managing and preserving them through digital curation activities for as long as necessary, to enable their effective reuse in future research.

These data, in addition to gaining value and importance in social, political, and economic scenarios, become crucial components in facing serious social and environmental challenges of the 21st century, through their sharing (Leonelli, 2019).

Meanwhile, on March 11, 2020, the World Health Organization (WHO) declared a COVID-19 pandemic caused by the SARS-CoV-2 coronavirus. As a result, issues regarding scientific data sharing and collaboration among different sources of investigation have gained prominence and have been evidenced. The pandemic presents a real and urgent need to gather efforts on a global scale, especially in the scientific field, so that the gaps about the new coronavirus are quickly and effectively resolved (Almeida *et al*., 2020).

Therefore, considering this reality, the São Paulo State Research Foundation (FAPESP), in cooperation with other institutions, announced the creation of the country's first open, anonymized Research Data Repository (RDR) re-

[125] Master in Management of Learning Organizations from the Professional Master's Program in Management of Learning Organizations at the Federal University of Paraíba – MPGOA/UFPB. E-mail: anderson.simoes@estudantes.ufpb.br.

[126] Master in Information Science from the Postgraduate Program in Information Science at the Federal University of Paraíba – PPGCI/UFPB. E-mail: renata.anjos@academico.ufpb.br.

[127] PhD in Communication Sciences (Information Science) from the School of Communication and Arts of the University of São Paulo - ECA/USP. Professor of the Department of Information Science of the Federal University of Paraíba. E-mail: guilhermeataide@ccsa.ufpb.br.

lated to COVID-19, the **COVID-19 Data Sharing/BR**[128], which aims to make available COVID-19 related research data in Brazil, to contribute to this topic (FAPESP, c2016).

In this context, COVID-19 Data Sharing/BR materializes the conception that sharing, using, and reusing datasets available in a repository efficiently help to solve the challenges posed by the pandemic, which emphasizes the current role of scientific data.

Some initiatives were proposed to contribute with effective data reuse by other researchers and scientific investigations, among which FAIR principles stand out. Those principles stand out as being an approach that aims at making data easily ***F****indable*, ***A****ccessible*, ***I****nteroperable* and ***R****eusable*. Such principles encourage, among other practices, the use of metadata that, when properly used, can contribute to increase findability, access, interoperability, and reuse of different datasets (Dias; Anjos; Rodrigues, 2019).

Given the importance of collaboration among scientists during the pandemic and the potential of adhering to FAIR principles to increase the sharing of scientific datasets available in digital repositories and maximize their use and reuse, the following research question is posed: **How well do the datasets available in the COVID-19 Data Sharing/BR repository conform to the FAIR principles?**

To answer the research question, the following general objective was elaborated: evaluate the adherence of datasets available in COVID-19 Data Sharing/BR data repository according to FAIR principles.

6.2 EXPANDING THE ACCESS TO SCIENTIFIC DATA IN THE CONTEXT OF COVID-19 PANDEMIC

The COVID-19 Data Sharing/BR data repository was created with the aim of promoting sharing and collaboration, and was established through a partnership between FAPESP, the University of São Paulo (USP), the Fleury Institute, as well as Sírio-Libanês and Israelita Albert Einstein hospitals, all located in the state of São Paulo. The repository contains demographic data on 75 million patients, 1.6 million clinical and lab exams, and data on the outcomes of 6,500 patients, all of which can be used to support scientific research on COVID-19. The datasets available in the repository are classified into three categories: demographic data (gender, year of birth, and region), data on clinical and/or lab exams, and data on the patient's history (hospitalization, recovery, and death) (FAPESP, c2016).

Regarding this matter, the FAIR principles initiative was created to serve as guidelines for academic-scientific, industry, financing agencies, and publisher scenarios that aim to improve their data support infrastructure and promote the use and reuse of their data. Unlike other initiatives that focus on improving data usage for humans, the FAIR principles seek to improve machine capacity to automatically find and use data, thus enabling its reuse by users (Wilkinson *et al.*, 2016). The FAIR principles aim to provide distinct considerations for contemporary data

[128] Available from: https://repositoriofapespcienciasaude.uspdigital.usp.br/. Access on: feb. 25, 2021.

release environments, including support for data input, exploration, sharing, and manual and automated reuse (Wilkinson *et al.*, 2016).

The first FAIR principle (***F****indable*) addresses the need to make data findable, which is an essential prerequisite for the effectiveness of the other three FAIR principles. A dataset must have a unique and persistent identifier, allowing its discovery at any time. Additionally, data must be described with rich metadata in a way that the researcher can find the desired data, even if they do not have access to its identifier (Dias; Anjos; Rodrigues, 2019).

The ***A****ccessible* principle focuses on the need to make data and metadata more accessible from the moment they are found. These entities must be accessible to users and/or machines at all times. For this purpose, it is important to use open, free, and universally implementable protocols (Go FAIR, [202-?]). It is recommended that metadata be available and accessible, even when the dataset does not allow free access to its content (Wilkinson *et al.*, 2016; Go FAIR, [202-?]).

The third FAIR principle (***I****nteroperable*) addresses the need to make data and other digital assets more interoperable. This issue is related to the need to integrate data with other datasets and with a wide range of applications throughout their lifecycle. To make interoperability among datasets possible, it is important to have instruments to semantically standardize the systems involved in this process, such as thesaurus and ontologies (Go FAIR, [202-?]).

The fourth and final FAIR principle (***R****eusable*) addresses the need to make data reusable. The implementation of data reuse requires a multifaceted approach and enables data to be reused by new communities of users for new needs and applications. In this regard, data can become more valuable to individuals in a wide range of organizations, including open-source communities and private organizations (Wise *et al.*, 2019). It is recommended that policies for data and metadata access be explicit, thus ensuring understanding of access, use, and reuse rights, as well as details that indicate the origin of these objects (Go FAIR, [202-?]).

In this perspective, it is noteworthy the importance of implementing FAIR principles, which, when implemented, can result in several developments, including the possibility of process automation through the capacity of machine-automated data and metadata reading. This contributes to increasing their reuse and scalability, besides providing a more rigorous data and metadata management with potential to benefit the entire academic community. Thus, FAIR principles become a premise to support scientific findings and innovation (Wilkinson *et al.*, 2016; Wise *et al.*, 2019).

6.3 METHODOLOGICAL PATH

The research aims to evaluate the adherence of datasets available in the COVID-19 Data Sharing/BR data repository to the FAIR principles. From the objective standpoint, it is characterized as an exploratory and descriptive study. Regarding the problem approach, it is structured, with all stages of the investigative process previously determined. As for the type of investigation, it is a mixed analysis, using qualitative analysis to verify the datasets' adherence to the FAIR principles and quantitative analysis to evaluate the FAIRness adherence score to the FAIR principles (Richardson, 2017).

The *corpus* of this research was constituted by datasets available in COVID-19 Data Sharing/BR repository, in a total of three sets, each one of them coming from one of the collaborating institutions, namely: Fleury Group, Sírio-Libanês Hospital and Israelita Albert Einstein Hospital.

To verify the FAIRness score, we used the *online Self-Assessment Tool to Improve the FAIRness of Your Dataset,* SATIFYD[129], proposed by *Data Archiving and Networked Services* (DANS). This tool serves as a tool for dataset auto evaluation before its publication.

The SATIFYD tool consists of 12 questions that address the FAIR principles, divided equally into sections corresponding to the letters of the acronym FAIR. Specifically, the *Findable* section includes questions one to three; the *Accessible* section includes questions four to six; the *Interoperable* section includes questions seven to nine, and the *Reusable* section includes questions ten to twelve.

As a means of evaluation, the tool provides both a score by letter/principle and a visual representation of the corresponding letter. The more "blue" each letter is, the more adherent the dataset is to FAIR principles in that respective dimension. The tool also provides a general score – FAIRness – calculated from the average score associated with each principle.

At the beginning of the analysis, it was observed that the three datasets in COVID-19 Data Sharing/BR were available in the same manner and followed the same structure. As a result, these datasets received similar scores by the end of the analysis. Therefore, in the presentation and analysis of the results, it was decided to present and analyze the results obtained from only one dataset, as they were deemed similar. The analysis was conducted jointly and simultaneously by the authors, with the goal of providing a comprehensive assessment from different perspectives.

It should be noted that to conduct the analysis, it was necessary to download the datasets, and access the metadata sets available in the respective repository under the option "Completed Record". It is imported to mention that there are no records of changes or updates to the datasets in their respective completed records. Therefore, it is necessary to identify the dates associated with these alterations or updates through metadata related to the transferred files.

6.4 PRESENTATION AND ANALYSIS OF RESULTS

The first section of the tool, corresponding to the *Findable* principle, consists of three questions. The first question pertains to the availability of sufficient metadata, and the tool provides an informative text for each question with an icon **"i"** next to it. The tool presents a list of 13 items indicating the parameters to be met when sufficient metadata is sought. During the analysis, it was observed that the metadata of the analyzed datasets did not satisfy four items on the list: people who contributed to the datasets, target group of the datasets, license indicating data accessibility, and spatial coverage (geographic location in which the research was conducted). It is worth

[129] https://satifyd.dans.knaw.nl/

noting that there is no information available on the individuals who contributed to the research that led to the creation of the datasets. Instead, only the collaborating institutions are mentioned as authors. Regarding the use license, the repository's initial page mentions that all datasets adopt the Creative Commons CC-BY open data license. However, users who access the datasets directly through their persistent identifiers may not have access to information on the adopted license.

The second question discusses the use of standards such as controlled vocabulary, taxonomies (thesaurus), or ontologies for describing sets. As analyzed, the datasets in question did not provide clues about the use of controlled vocabularies. It is noted that not using terminology control resources/tools causes a deficit, both at the moment of finding datasets and in ensuring that other researchers reuse them.

The third question pertains to providing additional documentation, such as a README file. Although it was not found with the same nomenclature, it is noteworthy that the repository clearly shows concern in developing a data dictionary for each of the datasets published. These data dictionaries are considered additional documentation that describes and explains the way data is structured, enabling any researcher and/or institution that can access it to understand and reuse it in their investigations.

Thus, regarding the *Findable* principle, the datasets obtained a FAIRness score of 38%.

The second section, which concerns the *Accessible* principle, also consists of three questions (questions four, five, and six). The fourth question addresses whether metadata is accessible to the public even when data is no longer available. As no information was found regarding this possibility of accessing metadata, even when data is no more available, it was decided to select the option "I can't find this information".

The fifth question addresses whether datasets contain personal data. According to the law Nº 13.709/2018, the Data Protection Law – DPL, personal data refers to natural individuals who are identified or identifiable. Since the data used to create the datasets were anonymized and conformed to the DPL's anonymized data classification, which pertains to data that cannot be traced back to an identifiable individual (Brasil, 2018), this question was answered negatively.

The sixth question discusses which use licenses were chosen to ensure access rights. As previously stated, the repository specifies the use license assigned to all its datasets on its initial page; however, the license is not informed in the sets of metadata. Among the available answer options, we have: open access to all, open access to registered users, restricted access through approval request, restricted access to specific groups and other types of access. It was decided to choose the option of open access to registered users, as users are required to complete a brief registration (name, e-mail, and institution) and agree to the statement of responsibility regarding the ethical use of data and the obligation to give proper credit to the datasets through citations to download the datasets.

It was observed that the complete registrations of datasets do not reveal metadata that facilitates the contact between users and data holders. It should be noted that this contact may be necessary to clarify future and potential questions.

As a result, regarding the *Accessible* principle, the datasets received a FAIRness score of 55%.

The third section, which corresponds to the *Interoperable* principle, also comprises three questions (questions seven, eight, and nine). The seventh question discusses whether the datasets are stored in preferred formats. The tool provides informative texts about these preferred formats, in which, for spreadsheets, the tool states that the preferred formats are ODS and CSV. The analyzed datasets can be downloaded in CSV format, so all the data is in the group of formats indicated as preferred.

The eighth and ninth questions discuss the linkage to other (meta)data and whether they are accessible online and whether there is the provision of contextual information (reference to other sets or publications) about datasets. It was noted that the datasets in question do not contain contextual information or links to other (meta) data, which goes against the recommendation of the Interoperable principle.

As a result, regarding the *Interoperable* principle, the datasets obtained a FAIRness score of 50%.

The fourth section, corresponding to the *Reusable* principle, is composed of three questions (questions ten, eleven, and twelve). The tenth question discusses if there is information about where data comes from, such as data origin, citation for reused data, description of the workflow, history of data processing, and version. In this question, only the option of data origin was selected since it is informed in its description. Concerning citations for reused data, description of workflow, and history of data processing, and version, no information was found.

It is valid to point out the relevance of these other information for datasets. The description of workflow allows other researchers to have a deeper understanding of how data was created or reused through citations of these datasets.

The eleventh question discusses once again the access and use license adopted by datasets, as previously discussed in the sixth question, with the same options of answer, in which open access was selected.

The twelfth question discusses if the (meta)data meets the domain standards concerning standardized data organization. The option of using domain standards was selected, considering the structured and standardized way in which data was organized.

Thus, regarding the *Reusable* principle, the datasets reached the FAIRness score of 74%.

For a better understanding and visualization of the questions and how they were answered, Table 1 was developed with this respective description.

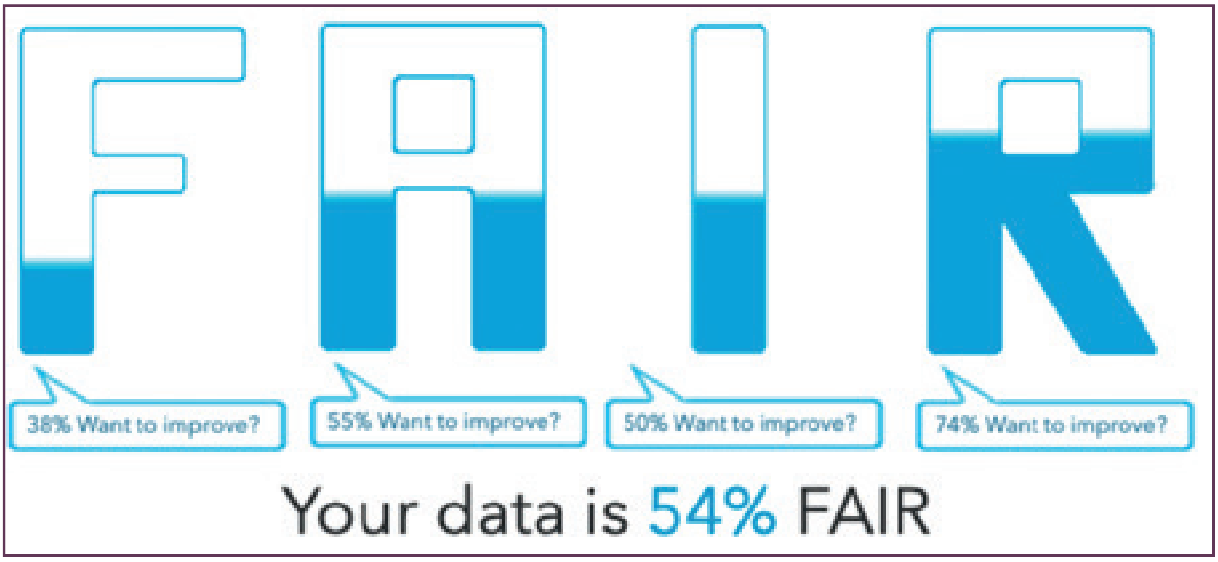
Table 1 – SATIFYD questions and respective options of answers selected.

PRINCIPLE	QUESTIONS	OPTION OF ANSWER SELECTED
FINDABLE Section 1	Did you provide metadata (information) enough about your data so that other people find, understand, and reuse it?	Mandatory metadata fields and some additional fields.
	Did you use standards such as controlled vocabularies, taxonomies (thesaurus) or ontologies to describe your dataset?	No standards were used.
	Did you provide rich and detailed additional documentation?	README file.
ACCESSIBLE Section 2	Is metadata accessible to the public even if data is not available anymore?	Yes.
	Does your dataset have personal data?	No.
	Which use license did you choose to fulfill the access rights attached to data?	Open access (registered user).
INTEROPERABLE Section 3	Is data in your dataset stored in preferential formats?	All data is in preferential formats.
	Do you link to other (meta)data? Can this (meta)data be accessed *on-line*?	No.
	Did you provide contextual information about your dataset?	With no contextual metadata.
REUSABLE Section 4	What kind of information did you provide about your data origin?	Data origin.
	Which use license did you use for your dataset?	Open access (user registered).
	Does your (meta)data meet the domain standards?	Domain standards in metadata.

Source: Data Archiving And Networked Services (c2024); Research data (2020).

In the end, the tool provided a final FAIRness score of 54%, which is calculated as the average score of each principle. Figure 1 illustrates how the tool presents the analysis result.

Figure 1 – Result of analyses in SATIFYD.



Source: Research data (2020).

6.5 FINAL CONSIDERATIONS

It was noted that, despite the datasets adopting some practices proposed by the FAIR Principles, the COVID-19 Data Sharing/BR does not mention or recommend the adoption of the principles before the act of publishing the data by its holders (Santos; Sant'ana, 2019).

As observed, the repository informs, only on its initial page, that all datasets published there adopt Creative Commons CC-BY open data license and that all and any publication or presentation that uses data in the repository must cite it. It is suggested that all datasets inform, in their metadata, the access and use license adopted so that there are no misunderstandings, given the possibility of users directly accessing datasets through their persistent identifiers.

Another point observed was that, repeatedly, only on its initial page, the repository informs that the datasets are periodically updated, so it must be frequently verified for downloading new data. On the other hand, in metadata of datasets published, there is no information about the version history, so it is up to the user to verify, after downloading, its creation date and confirm is there was an update. It is suggested that the version history is informed in metadata.

It is understood that the more evaluation tools available for use, the more opportunities to improve and rethink the ways of evaluating datasets according to FAIR principles; the need to mature the evaluation process is similar to its improvement. A relevant fact for the repositories as well is that the more evaluations are made, more proposals for improvements will be suggested.

During the research, only tools for self-assessment of datasets were found. That is, the researcher himself performs the self-assessment of his datasets before being published in the repositories. Thus, it is important to create tools that enable the analysis of datasets already published to investigate how the adoption of FAIR principles is being taken by the academic-scientific community.

An initiative such as COVID-19 Data Sharing/BR, which in this case was developed in face of a pandemic context, is very welcome; the rationale for the idea and the importance of collaborations from institutions such as FAPESP, University of São Paulo (USP), and the participation of Fleury Institute, besides Sírio-Libanês and Israelita Albert Einstein hospitals, show the importance of making data available in issues regarding public health. However, it is recommended that more efforts are put on increasing the adherence of FAIR principles in datasets stored in this repository.

REFERENCES

ALMEIDA, B. de A. *et al*. Preservação da privacidade no enfrentamento da COVID-19: dados pessoais e a pandemia global. **Ciência & Saúde Coletiva**, [*s. l.*], v. 25, p. 2487-2492, 2020. Available from: https://doi.org/10.1590/1413-81232020256.1.11792020. Access on: 15 aug. 2020.

BRASIL. **Lei nº 13.709 de 14 de agosto de 2018**. Lei Geral de Proteção de Dados (LGPD). Brasília: Presidência da República, 2018. Available from: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Lei/L13709.htm. Access on: 10 oct. 2020.

DATA ARCHIVING AND NETWORKED SERVICES. **Self-Assessment Tool to Improve de FAIRness of Your Dataset (SATIFYD)**. c2024. Available from: https://dans.knaw.nl/en/satifyd/. Access on: 07 oct. 2024.

DIAS, G. A.; ANJOS, R. L.; RODRIGUES, A. A. Os princípios FAIR: viabilizando o reuso de dados científicos. *In*: DIAS, A. D.; OLIVEIRA, B. M. J. F (org.). **Dados Científicos**: perspectivas e desafios. João Pessoa: UFPB, 2019. p. 177-187.

FAPESP. **FAPESP COVID-19 Data Sharing/BR**. c2016. Available from: https://repositoriodatasharingfapesp.uspdigital.usp.br/. Access on: 10 aug. 2020.

GO FAIR. **FAIR Principles**. [202-?]. Available from: https://www.go-fair.org/fair-principles/. Access on: 10 aug.. 2020.

SALES, L. *et al*. GO FAIR Brazil: a challenge for brazilian data science. **Data Intelligence**, [*s. l.*], v. 2, n. 1-2, p. 238-245, 2020. Available from: https://doi.org/10.1162/dint_a_00046. Access on: 20 aug. 2020.

LEONELLI, S. Data-from objects to assets. **Nature**, [*s. l.*], v. 574, p. 317 - 320, 2019. Available from: http://dx.doi.org/10.1038/d41586-019-03062-w. Access on: 20 aug. 2020.

RICHARDSON, R. J. **Pesquisa Social:** Métodos e Técnicas. 4 ed. São Paulo: Atlas, 2017. 424 p.

SANTOS, P. L. V. A. C.; SANT'ANA, R. C. G. Camadas de Representação de Dados e suas Especificidades no Cenário Científico. *In*: DIAS, A. D.; OLIVEIRA, B. M. J. F (org.). **Dados Científicos**: perspectivas e desafios. João Pessoa: UFPB, 2019. p. 53-66.

WILKINSON, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, [*s. l.*], v. 3, n. 1, p. 1-9, 2016. Available from: https://doi.org/10.1038/sdata.2016.18. Access on: 15 aug. 2020.

WISE, J. *et al*. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. **Drug Discovery Today**, [*s. l.*], v. 24, n. 4, 2019, p. 933-938. Available from: https://doi.org/10.1016/j.drudis.2019.01.008. Access on: 15 aug. 2020.

How to cite this chapter: SIMÕES, Anderson Rafael Castro; ANJOS, Renata Lemos dos; DIAS, Guilherme Ataíde. Analysis of the datasets available in the COVID-19 data sharing/BR repositor in conformance to the 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. 6, p. 80-89. DOI: 10.22477/9788570131959.cap6.

7. FAIR PRINCIPLES AND LINKED DATA: PUBLICATION OF OPEN NOTEBOOK SCIENCE

Luciana Candida da Silva[130]
José Eduardo Santarem Segundo[131]

7.1 INTRODUCTION

The way to conduct and communicate scientific research has been going through changes. Such changes are caused by advances in technology, overproduction of data, funding pressures and collaborative culture among researchers.

This new way of conducting science contemplates the dissemination of primary data, preferably, as far as they are generated, and not only in successful cases or cases with consolidated results. The tendency is to allow the open simultaneous collaboration to a broad contribution with the purpose of achieving new results.

Notebooks of science are data annotation instruments, generally experimental, produced in laboratory to base scientific publications that are normally published in its final configurations. For Schnell (2015) notebook science reports hypothesis, experiments and initial analysis or interpretations of the experiments, and it works as a legal record of properties of ideas and results obtained by a scientist. It is observed that despite the importance given to data recorded in journals, in its majority it is not published in a structured way so it is reused by human users and machines.

In this context, Web Semantic and FAIR principles stand out to guide the structuring of open notebook science in a way to make it findable, accessible, interoperable and reusable, besides improving the capacity of machines to automatically find and use data and support its reuse by individuals. Therefore, this group presents semantic elements for publishing open notebooks from the application of FAIR principles, in the Web Semantic and *Linked Data* perspective to support the new scientific practices.

This study is part of a doctoral research that has as its objective to provide semantic guidelines for structuring and publishing open notebook data, aiming at improvements in retrieving and sharing data.

7.2 OPEN NOTEBOOK SCIENCE

The term *Open Notebook Science* was coined by Jean-Claude Bradley in September 2006 to promote debates on open collaboration in science and developing more efficient research techniques. For Bradley (2010) the goal

[130] PhD in Information and Documentation from the University of Barcelona, Professor in the Postgraduate Program in Information Science, Federal University of Rio Grande do Sul. E-mail: fabianocc@gmail.com.

[131] PhD student in Information Science at the Federal University of Minas Gerais, Librarian-Documentalist at the Federal University of Uberlândia. E-mail: marcellomundim@ufu.br.

of open notebook science is making the details of experiments conducted in laboratories freely available on the Web, which does not restrict only successful.

For Schapira and Harding (2019) the opening and sharing of scientific research data recorded in notebook science are efficient and fast ways to disseminate data before it is published in peer-reviewed journals, and they present advantages in relation to the traditional way-release after publication.

> First, by making data accessible in weeks, rather than keeping it hidden for years, it means that others will be able to take advantage of the research and avoid wasting redundant experimental time and resources. Second, open notebook science must include detailed protocols that can be reproduced, which is often not the case in peer-reviewed publications. Third, unsuccessful data, which are seldom released in the current publishing system but are provided in open notebook science, can therefore help to save time, resources, and knowledge (Schapira; Harding, 2019, p. 3).

According to Schapira and Harding (2019) the open notebook science includes several procedures performed during an experimental research in a way to ensure the successful replication of results. These sets of procedures are mentioned in the literature as research protocol, which is set as the main purpose of data dissemination and sharing, for it has a complete description of the procedures performed, the equipment adopted, and the reagents used during the research, besides declaring the intended purposes, the discussion of data found in the research, and it records the results achieved either partial or final. The protocols must be followed by textual documents and spreadsheets in case they are necessary for the interpretation of procedures for materialization of objects of study such as lines, dots, graphics, maps, spectra among others.

The proposal for opening research datasets recorded in notebook science is part of a bigger movement of Open Science called *e-Science,* characterized by the intensive use of technologies and collaborative efforts, which bring the opportunity of thinking about the new contexts and scientific practices. In this context, Foster (2018) classifies the notebook science as integral part of the third dimension *Open Reproducible Research*, of the open Science taxonomy, which constitutes in the act of offering to users free access to experimental elements to allow the reproduction of research regardless of its results.

7.3 FAIR PRINCIPLES

FAIR principles, an acronym for *Findable*, *Accessible*, *Interoperable* and *Reusable*, originated in 2014, during the international conference *Jointly designing a data FAIRPORT,* from a debate among representatives of several fields of knowledge, such as librarians, archivists, editors and research sponsors, members of *The Future of Research Communications and e-Scholarship* (FORCE11), to improve the research data ecosystem and work as guidelines to increase research data reuse in the context of *e-Science* (FORCE11, 2014).

This discussion resulted in four relevant principles, with guiding practices for data release that would be easily findable, accessible, interoperable, and reusable by machines and humans, in the face of the great amount of information generated by intensive contemporary Science in data. These principles incorporate characteristics that define that the resources, tools, vocabularies, and contemporary data infrastructure should be shown to help to find and reusing third-parties (FORCE11, 2014).

According to Wilkinson *et al*. (2016), the elements of FAIR principles are related, but are independent and separable and can be implemented in any combination, incrementally, as data providers evolve their structures to achieve a higher level of FAIR principles purpose. The authors clarified that these principles precede the choices of implementation, and they do not limit Technologies for implementation. Thus, this study associates FAIR principles to Web Semantic and *Linked Data*.

7.4 WEB SEMANTIC AND LINKED DATA

Web Semantic was started in 2001, by Tim Berners-Lee with the collaboration of Hendler and Lassila (2001), from the proposal of defining an efficient way to represent data on the Web and provide improvement in the quality of information retrieval, allowing, according to Santarém Segundo (2012, p. 106) "the users to obtain more accurate results and with information closer to what they really need".

The term *Linked Data* is presented as principles for implementation of Web Semantic technologies to publish and promote the connection of data from different sources on the Web, in order to provide benefits to the data. These principles are: 1– use *Uniform Resource Identifier* (URIs) with names for things; 2–use URI HTTP, so people can look up these names; 3– when someone looks for a URI, provide useful information, using standards such as *Resource Description Framework* (RDF) and SPARQL; and 4– include *links* for other URIs so that the items related can be found. (Berners-Lee, 2006).

The consortium of World Wide Web (W3C) recommends a set of Technologies for releasing open data and connection with the Web, according to *Linked Data* principles. Among the Technologies, *Resource Description Framework* (RDF) and its serialization stand out, with RDF as a standard model adopted for the description of information structured on the Web, allowing machines to legibly represent information about a resource um (W3C, 2014).

W3C recommends a set of 35 (thirty-five) Best Practices (BP) for *Data on the Web Best Practices* (DWBP), to improve the coherence between provider and consumer, encourage and allow the continued expansion of the Web as a way for data exchange and promote reliable data reuse

The thirty-five BP to release data on the Web are distributed in categories and for each best practice a set of benefits is obtained, as presented in Table 1.

Table 1 – Best Practices and Benefits

Category	Best Practice	Benefits
Metadata	BP 1 – Provide metadata for human users and machines	Reuse, understanding, finding, and ability to process
	BP 2 - Provide descriptive metadata	Reuse, understanding and finding
	BP 3 - Provide structured metadata	Reuse, understanding, and ability to process

Category	Best Practice	Benefits
Licenses	BP 4 – Provide information on data license	Reuse and reliability
Data origin	BP 5 – Provide information on data origin	Reuse, understanding, and reliability
Data quality	BP 6 – Provide information on data quality and necessary adaptations	Reuse and reliability
Data version	BP 7 – Assign version for each dataset	Reuse and reliability
	BP 8 – Provide a history of versions	Reuse and reliability
Data identifiers	BP 9 - Use persistent URIs as datasets identifiers	Reuse, interconnection, finding, and interoperability
	BP 10 – Use persistent URIs as identifiers within datasets	Reuse, interconnection, finding, and interoperability
	BP 11 - Assign URIs to datasets versions	Reuse, finding, and reliability
Data formats	BP 12 – Use standardized data formats	Reuse and ease of processing
	BP 13 – Use neutral data representation to locations	Reuse and understanding
	BP 14 – Provide data in several formats	Reuse and processability
Data vocabulary	BP 15 – Reuse vocabularies, preferably standardized	Reuse, processability, understanding, reliability and interoperability
	BP 16 – Choose the correct level of formalization	Reuse, understanding, and interoperability

Category	Best Practice	Benefits
Data access	BP 17 – Allow complete mass access	Reuse and access
	BP 18 – Allow partial access to the datasets	Reuse, access, interconnection, and processability
	BP 19 - Provide data in several formats	Reuse and access
	BP 20 – Allow access in real time	Reuse and access
	BP 21 – Provide updated data	Reuse and access
	BP 22 – Explain the reasons when data is not available anymore	Reuse and reliability
	BP 23 – Provide data through API	Reuse, processability, interoperability, and access
	BP 24 – Use Web standard as APIs basis	Reuse, interconnection, interoperability, finding, access and processability
	BP 25 – Provide documents as far as API is added or changed	Reuse and reliability
	BP 26 – Avoid breaking changes in API	Reuse and interoperability
Data preservation	MP 27 – Preserve the identification and provide information on the field resource	Reuse and reliability
	MP 28 – Evaluate the coverage of a dataset before its preservation	Reuse and reliability
Feedback	MP 29 - Collect *feedback* from consumers	Reuse, understanding, and reliability
	MP 30 – Make the *feedback* publicly available	Reuse and reliability
Data enrichment	MP 31 – Enrich data, generating new data	Reuse, understanding, reliability and processability
	MP 32 – Offer complementary presentations	Reuse, understanding, access and reliability
Republication	MP 33 - Provide *feedback* to the original publisher	Reuse, interoperability, and reliability
	MP 34 – Follow the license terms	Reuse and reliability
	MP 35 – Cite original publication	Reuse, finding, and reliability

Source: Adapted from de Lóscio, Burle and Calegari (2017).

After presenting these BP, it is possible to analyze the semantic elements to describe the open notebook Science concerning data coverage being findable, accessible, interoperable, and reusable, from the application of FAIR principles, We Semantic Technologies and *Liked Data* concepts.

7.5 SEMANTIC ELEMENTS OF DESCRIPTION OF OPEN NOTEBOOK SCIENCE

It presents the result of metadata and vocabularies mapping that aim to describe and individualize the objects that compose the notebook Science ecosystem, especially regarding experimental research for structuring and publication of open notebook Science. The mapping presents a set of semantic elements, according to Table 2.

Table 2 – Mapping of metadata and vocabulary properties

Metadata Spreadsheet (Labels)	Vocabulary Property Schema.org, DC Terms, SKOS and RDA Element Sets
Identification of record	schema: identifier
Date and hour of record	schema:dateTime
Author ^	schema:author
• Date of birth	schema:birthDate
• Date of death	schema:deathDate
• Profession/Occupation	schema:hasOccupation
• Institution associated ^	schema:memberOf
• Department of Institution	schema:department
Contributor /Collaborator ^	schema:participant
• Date of birth	schema:birthDate
• Date of death	schema:deathDate
• Profession/Occupation	schema:hasOccupation
• Institution associated ^	schema:memberOf
• Department of Institution	schema:department
Institution ^	schema:sourceOrganization
Development Agency ^	schema:funder

Metadata Spreadsheet (Labels)	Vocabulary Property Schema.org, DC Terms, SKOS and RDA Element Sets
• Identification of the agent	schema:Identifier
• Controlled access point	skos:prefLabel
• Variable Access Point	skos:altLabel
• Field of activity	rdaa:P50387
• Language ^	schema:inLanguage
• Contact Information	schema:email
Title	schema:name
Subtitle	schema:alternativeHeadline
Language ^	schema:inLanguage
Format ^	dct:format
Type ^	dct:type
Total number of pages	schema:pagination
Spatial coverage ^	schema:location
Research period	schema:startTime
Audience	schema:audience
Intended purpose	schema:description
Results achieved	schema:result
Subject ^	schema:about
• Identification ^	schema:propertyID
• Controlled access point	skos:prefLabel
• Variable access point	skos:altLabel
• More extensive subject	skos:broader

Metadata Spreadsheet (Labels)	Vocabulary Property Schema.org, DC Terms, SKOS and RDA Element Sets
• More specific subject	skos:narrower
Description	schema:description
• Reagents ^	schema:activeIngredient
• InChIKey^	schema:identifier
• Equipments	schema:instrument
• Molecular formula^	schema:identifier
• Molecular weight^	schema:weight
• Measurement technique	schema:measurementTechique
• Chemical names ^	skos:related
• Commercial name ^	skos:related
• Date of creation	schema:dateCreated
• Date of modification	schema:dateModifield
• Closure research period	schema:endTime
• Status of the action	schema:actionStatus
• Error	schema:error
Data source	schema:provider
Declaration of provenance	dct:provenance
Use license	schema:license
Declaration of rights	dct:RightsStatement
Holder of rights	dct:rightsHolder
Size of the application	schema:fileSize
Necessary Software	schema:availableOnDevice

Metadata Spreadsheet (Labels)	Vocabulary Property Schema.org, DC Terms, SKOS and RDA Element Sets
Record of display	schema:RegisterAction
Control of use and user	schema:userInteractionalCount
Type of user interaction	schema:interactionType

Source: Silva (2020).

The first column presents the metadata identified through data modeling and insertion of attributes that describe the notebook Science specificities, plus elements that can be enriched with external vocabularies, such as a certain author's date of birth and date of death when it is the case. The elements marked with a triangle (▯) indicate the importance of reusing information from external *datasets* through URI whenever possible to avoid ambiguities, to ensure standardization and carry additional information to those required through metadata. The second column presents Schema.org, DC Terms, SKOS e RDA *Element Sets* vocabulary properties corresponding to metadata in the first column.

This structure presents an analysis of elements to be considered FAIR, considering that FAIR principles have been required by the academic Community, especially by development agencies, as evaluation criteria for funding research.

7.5.1 Analysis of semantic elements of notebook Science regarding being FAIR

The scope of the elements indicated in the development of guidelines for publishing notebook Science is discussed regarding data being findable, accessible, interoperable, and reusable from the application of FAIR principles, Web Semantic Technologies and *Linked Data* concepts, recommended by W3C.

7.5.1.1 *Findable*

The *findable* principle recommends four practices so that can be findable. The first principle, F1, indicates the assignment of identifiers globally exclusive and persistent, and the third principle, F3, indicates that data identifiers must be included. In this study, notebook Science were structured from the use of persistent indicators to describe names of objects, people, institutions, places and subjects, through the definition of metadata *tags*.

The recommendations of W3C highlight that the finding, use and citation of data on the Web depend fundamentally on the use of URIs in HTTP that can be checked on the internet. Thus, BP 9, 10 and indicate the use of persistent URI for the dataset and URI as identifier of datasets. In this regard, it is possible to observe the *tag* mapping directed to the use of identifiers or persistent URIs, such as people identifiers through vocabularies as ORCID and VIAF (*schema:author*, *schema:funder and schema:sourceOrganization*), indicators for spatial coverage using GeoNames (*schema:location*) vocabulary and subject indicators through LCSH, MeSH and PubChem vocabularies that can be presented from *schema:about* properties and its developments, as presented in Table 2.

In this perspective, the second principle *findable* (F2) recommends that a dataset should be described by metadata rich enough so that, once indexed in a Search mechanism, it can be found even without its persistent identifier. The BP 1, 2 and 3 recommend providing descriptive, structural and administrative metadata. The elements mapped describe data related to the experimental research, and it did not intend to be excessive but describe the information considered sufficient so that the researcher can analyze and choose data reproduction or repeatability.

In the occasion of implementing these guidelines, it is necessary to offer metadata for human Reading, Where W3C recommends providing metadata as part of a webpage HTML and as a separate text file. For machine interpretation, metadata can be provided in Turtle and JSON serialization format, or it can be incorporated in HTML page (HTML-RDFA or JSON-LD), and reuse existing standards and popular vocabularies. For notebook science, the integrated use of Schema.org and Dublin Core metadata standards were chosen to enable the detailed description of metadata values. Schema.org was selected to describe digital objectives regarding notebook science because it has numerous properties corresponding to experimental research, and Dublin Core because it has multiple properties to describe information such as provenance, format, and types of data. In addition to these, the SKOS vocabulary was adopted to refine metadata values.

The recommended practice in F4, the fourth principle *findable*, is that metadata is recorded and indexed in search mechanisms. The BP 12 collaborates with principle F4 when it recommends the use of standardized, machine-readable formats when releasing data on the Web. The indicated formats include, but are not limited to CSV, XML, HDF5, JDON e RDF serialization syntax such as RDF/XML, JSON or Turtle. BP 24 recommends adopting standards as the basis of APIs, and BP 35 suggests citing the original publisher as so it can be easily found. Those guidelines provided *tag schema:provider* to indicate the data source and to provide data reliability.

7.5.1.2 *Accessible*

According to Wilkinson *et al*. (2016) data accessibility is related to the use of standardized communication protocols, open and free, which offer authentication and access to metadata even when it is not available anymore.

W3C recommendation is enabling the access to the complete dataset of a certain research. In order to easily access these datasets, BP 17 recommends that the Web infrastructure must be implemented in a way to allow the mass access of a complete dataset with just a request, avoiding inconsistency in the individual data access over many retrievals, as well as allow the provision of data subsetting (BP18), in case consumers do not need the complete set.

The Web offers access using methods of hypertext transfer protocol (HTTP) for simple mass download of a file. Even if data is distributed in several URIs, it can be organized in a container model using the file transfer protocol to enable mass access to data. The distribution of data in several files allows the retrieval through an application programming interface (API), the most sophisticated retrieval method. BP 18, 20, 23 and 24 mention that an API is the most flexible approach for serving data subsets because they allow to personalize which data is transferred and provides data in real time.

The second principle *Accessible* (A2) suggests that metadata should be accessible, even when data is inaccessible, while W3C suggests, through BP 22, providing explanations for data that is not available, informing how it can be accessed and who can access it. In this regard, recommendations FAIR and BP can complement each other making data available, even when data is not available anymore, and still offer explanatory message.

7.5.1.3 *Interoperable*

The interoperability refers to the capacity of a system to easily communicate with others. For this purpose, it is necessary to adopt actions such as assign interconnected metadata and Web standards as the basis for APIs. In addition, in order to have a formal language, it is necessary to provide human and machine-readable metadata (BP 1 and 2), use persistent indicators (BP 9 and 10) and reuse standardized vocabularies (BP 15 and 16).

For structuring notebook science, vocabularies describing their purposes were selected, such as the description of authorities as refining attributes indicating dates, profession, field of activity, means of communication with the agents and the possibility of linking researchers to institutions and departments in to which they belong. In addition, vocabularies describing the specificities of an experimental research were sought such as names of chemical compounds, chemical properties, procedures carried out during the research, period of realization, status of the research and a way to inform if the research was successful; vocabularies enabling the description of data provenance were also sought. After detailing the attributes that describe the specificities of notebook science, Schema.org, DCTerms and SKOS vocabularies were chosen, in addition to value vocabularies such as GeoNames, VIAF, ORCID among others exemplified in F1 to F3. As a model of representation, the recommended practice is the use of RDF and its serializations, which is also mentioned in F2. Moreover, the use of widely recognized vocabularies enables the benefits of data interoperability, processability, understanding, reliability and reuse.

7.5.1.4 *Re-Usable*

This principle is especially important for the notebook science context because it reflects the application of the previous principles (findable, accessible and interoperable) to the structuring purpose that is data being reused by researchers in new research.

The principle R1 establishes that metadata must be described with plurality of accurate and relevant attributes, therefore, provenance, description of notebook Science specificities, administrative and use metadata were mapped. In order to describe textual values, the use of persistent identifiers is recommended with the purpose of data enrichment.

Principle R1.1, BP4 and BP34 highlight the importance of providing a type of license to avoid limitations in reusing and legally formalizing the provision of data to be reused in other works. Principle R1.2 and MP 5 recommend that metadata must be associated to its provenance. In these guidelines, the tags for data source (*schema:provider*) were mapped to indicate data provenance, provenance declaration (*dct:provenance*) to explain changes in the property and custody of an object, license (*schema:license*) to allow the object use, declaration of rights (*dct:RigtsStatement*), holder of the rights (*dct:rightsHolder*) to indicate the name of the agent that has or manage the rights of the object, date of creation (*schema:dateCreated*) and modification of data (*schema:dateModifield*) to inform the original date and the modifications.

BP 13 suggests providing information on location parameter to avoid difficulties in understanding data that changes from one language to another, for example informing the language on the *tag schema:inLanguage* of the Schema.org standard.

BP 14 recommends, whenever possible, that data is available in multiple formats, when more than one match its intended use.   When implementing these guidelines, it is recommended that data is converted to RDF/XML, JSON and Turtle formats, which meet BP 12 recommendations for using machine-readable format, BP 14 recommendation for multiple formats, BP 15 known and standardized vocabularies.

7.6 FINAL CONSIDERATIONS

The provision of administrative, descriptive, provenance, preservation and use metadata, with the indication of implementation from standardized, human-and-machine readable vocabularies, in addition to the recommendation for attribution of URIs for indicating names and data enrichment, can guarantee that notebook Science data are findable.

Following the recommendation for using APIs will enable data accessibility. In addition to these elements, it is recommended the use of contextual explanation on about data and its metadata in order to provide interoperability of data. It is noteworthy that unifying previous recommendations to the indication of a use license, preferably of public domain, as well as detailing metadata associated to data provenance, enables reusing notebook Science data released based on these guidelines. It is worth noting that not all recommendations related to accessibility principles are covered in these structuring guidelines; however, the recommendation remains for the occasion of implementation.

REFERENCES

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, New York, 17 May 2001.

BERNERS-LEE, T. **Linked Data**. 2006. Available from: https://www.w3.org/DesignIssues/LinkedData.html. Access on: 13 Oct. 2024.

BRADLEY, J. C. The impact of open notebook Science. **Information Today**, Medford, NJ, v. 27, n.8, p. 50-51, set. 2010. Interview by Richard Poyder. Available from: http://www.infotoday.com/it/sep10/Poynder.shtml#top. Access on: 2 Set. 2019.

FORCE11. The Future of Research Communications and e-Scholarship. **Guiding principles for findable, accessible, interoperable and reusable data publishing version B1.0**. 2014. Digital text. Available from: https://www.force11.org/fairprinciples. Access on: 13 Jun. 2018.

FOSTER. **Open reproducible research**. 2018. Available from: https://www.fosteropenscience.eu/taxonomy/term/102. Access on: 2 Ago. 2020.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N (ed.). **Data on the Web best practices**. W3C, 2017. W3C Recommendation. Available from: https://www.w3.org/TR/dwbp/. Access on: 20 Mar. 2020.

SANTARÉM SEGUNDO, J. E. Tim Berners-Lee e a ciência da informação: do hipertexto à Web Semântica. *In*: SANTARÉM SEGUNDO, J. E.; SILVA, M. R.; MOSTAFA, S. P. (org.). **Os pensadores e a Ciência da Informação**. Rio de Janeiro: E-papers, 2012. p. 101-109.

SCHAPIRA, M.; HARDING, R. J. Open laboratory notebooks: good for Science, good for society, good for scientists. **F1000Research Open for Science**, v. 8, n. 87, 2019. Version 1; peer review: 2 approved with reservations. Available from: https://doi.org/10.12688/f1000research.17710.1. Access on: 21 Set. 2019.

SCHNELL, S. Ten Simple Rules for a computational biologist's laboratory notebook. **PloS Computational Biology**, São Francisco, v. 11, n. 9, 2015. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004385. Access on: 28 Feb. 2021.

SILVA, L. C. **Publicação de dados de pesquisa científica**: proposta de estruturação semântica de cadernos abertos de pesquisa frente às dimensões da e-Science. Orientador: José Eduardo Santarem Segundo. 2020. 243 f. Tese (Doutorado em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista Júlio Mesquita Filho (PPGCI/UNESP), Marília, SP, 2020. Available from: http://hdl.handle.net/11449/194341. Access on: 15 Oct. 2024.

WILKINSON, M. *et al*. The FAIR guiding principles for scientific data management and stewardship. **Sci Data**, n. 3, 2016. DOI 10.1038/sdata.2016.18. Available from: https://www.nature.com/articles/sdata201618. Access on: Oct. 2024.

W3C. World Wide Web Consortium. **RDF: Resource Description Framework**. 2014. Available from: https://www.w3.org/RDF/. Access on: 27 Oct. 2024.

How to cite this chapter: SILVA, Luciana Candida da; SANTAREM SEGUNDO, José Eduardo. FAIR principles and linkdes data: publication of open notebook Science. *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. 7, p. 90-102. DOI: 10.22477/9788570131959.cap7.

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.

9. TECHNOLOGIES FOR RESEARCH DATA MANAGEMENT ACCORDING TO FAIR

Milton Shintaku[136]
André Luiz Appel[137]
Alexandre Faria de Oliveira[138]

9.1 INTRODUCTION

Since science emerged as an activity separated from philosophy, it has evolved as the scientific community has been structuring, adapting and influencing technological change. In this matter, technology has always been present, being one of the results arising from science, mainly from rigid sciences, in partnership with engineering. However, it can be said that computer science and the Web, created in the middle of the 20th century, respectively, were landmarks that had a significant impact in the sciences, among other human activities. This is so clear that there are already studies on the so-called artificial science and virtual science, complementing the natural sciences, in which natural phenomena are simulated through computing, or virtual scenarios are created for research.

In scientific communication, the technology, mainly the web, also had such an impact on processes adopted by researchers that, for many thinkers, the Web has the same impact on sharing information as on the invention of Gutenberg's press. In this way of the Web, at the end of the 20th century, two initiatives arose and consolidated, being accepted in the worldwide scientific Community: Open Archives and Access, which are often confused for having the same acronym in English, but they have different aspects.

Open Archives have technological alignment, focused on interoperability, on the possibility of information exchange between computerized systems, initially oriented to digital libraries. This initiative was created by Convenção de Santa Fé (Santa Fé Convention), described by Van de Sompel and Langoze (2000) as oriented to the creation of informational infrastructure for interoperability, mainly for *preprints*. Soon after, it was used by Suleman and Fox (2001) for the creation of digital libraries of thesis and dissertations, for the creation of *Networked Digital Library of Theses and Dissertations* (NDLTD). In Brazil, Open Archives influenced the development of the Digital Library of Theses and Dissertations (DLTD).

Open Access, in its turn, has more philosophical aspects. Originated from the called journal crisis, it defends access with no restrictions to research results. For the implementation of this initiative, Harnad *et al.* (2004) suggested the use of open-access magazines (golden route) and the use of repositories (green route). In Brazil, initiatives such as Bioline *International*, 1993, and *Scientific Electronic Library Online* (SciELO), 1997, were pioneers in supporting the

[136] Doctor in Information Science, Brazilian Institute of Information in Science and Technology, shintaku@ibict.br

[137] Doctor in Information Science, Brazilian Institute of Information in Science and Technology, andreappel@ibict.br

[138] Graduate Student in the Master's program in Information Science at Brasília University, Brazilian Institute of Information in Science and Technology, alexandreoliveira@ibict.br

publication of sets of open-access magazines in electronic environments, representing the golden route. Later, several magazines started to offer free access, and universities created repositories for providing wide access and flow to their scientific production.

Even presenting distinct origins and settings, Open Archives and Open Access merged themselves, as some software sets for the creation of magazines and repositories blend some precepts from both initiatives. However, during some time, in Brazilian universities, libraries of thesis and dissertations and repositories coexisted, for they originated from different initiatives. Due to its conceptual aspects, Open Access comprises open archives in numerous instances.

In this path of opening up science activities, another initiative has been gaining force, the so-called Research Open Data, which acts on the dissemination of data collected or generated in the scope of scientific research. Data sharing through the Internet is not new, so Much that government data has been publicly available, as a way to provide more transparency to the public administration, according to the Access to Information Law (AIL) in Brazil. However, sensitive data resulting from research, still requires attention and discussions, in the same way that the act of sharing requires implementation of computerized platforms specific to the several types and formats of research data.

In this context, Open Data requires discussions for its implementation, even though it presents advantages to the research. One of the points is found in technologies available to be used for data deposit. Another point involves issues concerning protection, restriction, sensibility and many other issues in the scope of research data management, and that need more consensus so that technologies can reach levels of consensus, explicit, for instance, by initiatives as FAIR principles.

9.2 RESEARCH OPEN DATA

In the model of scientific communication proposed by Björk (2007), issues related to research data show up in the processes of research execution and communication of results. Regarding the research execution, the author reports that the research includes four global activities, one of which involves collecting existing data in repositories. That is, in addition to literature review, it is necessary to review existing data, relating them to the context of the research being executed. In the communication of results, in its turn, researchers must deposit their research data in repositories as a way to promote the reproducibility and the reuse in new studies.

This model is aligned to practices linked to open data, as it occurs in research processes and dissemination of data, in a cyclical process, in which existing data support the creation of new data. Murray-Rust (2008), discussing open data in science, emphasizes that they have data sharing for reuse as a principle, in a way to enable new perceptions, regrouping, additions, etc. Thus, the Open Data Initiative represents the removal of barriers when sharing research data, as in an Evolution of Open Access, which removed barriers for the access of articles resulting from research.

Pampel and Dallmeier-Tiessen (2014) advocate for what they call new science, made by sharing and reducing barriers, revealing the need to create strategies for promoting and opening data due to the increasing demand by the community for such practices. For the authors referred, one of the fundamental points for open data success is research data management through infrastructure based on robust policies that support cooperation among scholars, as far as changes in scientific practices only occur with the approval of the academic community.

Evidently, the challenges do not refer only to researchers' behavior. Borgerud and Borglunf (2020), for example, report that in Switzerland, even with federal regulations, opening research data is challenging and raises issues such as the preservation for long periods of time, in a Mertonian view of data archiving, based on accessibility, preservation, verification, and reusability principles. Despite being preliminary, this study reverberates what may countries face, reflecting that it is not enough for the community to accept a new practice, coupled with the government action in implementing laws and other support regulations; it is also necessary studies that support the creation of infrastructure that meets and supports these new practices.

Another point discussed in opening research data is quality. In this matter, Koltay (2020) discusses data reliability with several relations and criteria that help in its verification, such as originality, methods of collection and processing, authenticity, acceptability, applicability, among others. Likewise, the author points out concerns with technical quality in relation to databases shared, aiming at enabling the reuse of data in it. At this point, the author reaffirms the need of data curation to guarantee integrity and authenticity, with the aim of allowing researchers to access reliable and safe databases.

In this context, it is revealed that open data has complexities and challenges in its implementation, involving cultural, procedural, technological, legal issues, among others. Eliminating barriers for research data sharing requires studies in several fields, which must create an extensive conceptual basis that will guide the confrontation of several challenges. Thus, it must be ensured that the actions, methods, and Technologies seek to efficiently meet the purpose of reusing data.

9.3 RESEARCH DATA

The concept of research data, according to Costa (2017), may be affected or influenced by several categories that are assigned to data itself, thus causing the emergence of several definitions that reflect the typology of data, its forms of generation, collection or access, its purposes, research stages in which data is use or generated, etc.

In this matter, research data differs according to the subject, in most cases due to methods used, as rigid sciences generally have a preference for quantitative methods, while humanities, in its turn, prefer more qualitative methods. Thus, research data reflect methodological procedures, generating data more or less structured and composing databases with significant differences, which reflect the nature of the subject, creating different challenges.

Thus, the subjects must have different recommendations, depending on the type of database. Specifically in social sciences, Diaz (2019) recommends that data sharing requires concerns about the current regulations in the country, the code of ethics and laws that guide from data collection to data release, ensuring that data sharing, verification of risks, support from institution in the process are allowed by the researcher, and awareness about ethical responsibilities of the research and collected data.

In the same vein, Sayão and Sales (2016) observe that "the term 'research data' has a range of meanings that change according to specific scientific domains, research objects, methodologies of creation and collection of data and many Other variables".

One of the most recurrent definitions cited in literature is the one presented by the U.S. National Science Board (2011), according to which research data are factual records in the digital environment, generally accepted by the academic community with the purpose of validating research outcomes. Such definition may be understood as restrictive, once it excludes the possibility of data as physical objects and, in this case, including document objects, besides a different category of material records, not digital or computerized, which may be useful in the context of humanities, for instance. Nevertheless, the communication, in general, requires the transposition or different representations for other forms of record, with a heavier logical and conceptual upload. In this matter, the broader definition, presented by Sales and Sayão (2019), helps to accommodate possible variations of context. These authors define research data as "every and any type of record collected, observed or used by scientific research, treated and accepted as necessary for validating the research outcomes by the scientific community" (Sales; Sayão, 2019, p. 36).

The predilection for the digital nature of data is not new, and may refer to the concept of *datalogy*, originally presented by Peter Naur in 1966 to refer to the study and processing of data in already computerized environments (Naur, 2007; Zhu; Xiong, 2015). More recently, Zhu and Xiong (2015) also questioned an interesting distinction, to outline an object of study of data Science, based on the separation between natural phenomena (*real nature*), or observable through the natural or real world of digital phenomena, which other authors define as *data nature*.

Still, in order to define and outline research data, Costa (2017) highlighted some important concepts and practices in this matter, among which the management and life cycle of research data. The author points out that management involves processes of planning, manipulation, storage, and preservation of research data, while Sayão and Sales (2016) discuss actions that collectively permeate the life cycle of research data, in addition to application of standards widely accepted by several academic or disciplinary instances. Sales, Costa and Shintaku (2020) present three potential contexts of research data management application, namely: the context of researchers themselves, the data supporting their research; the context of funding agencies focused on the impact and contribution of research, which is potentially measurable through data sharing; and the context of scientific editors, focused on the verification and reproducibility of research outcomes through data. Data life cycle, in its turn, may include several stages that vary according to the complexity of the cycle, which generally starts with the generation or collection stage, up to reuse or disposal stages, when it is applicable.

Considering the universe of technologies available in data context, it is important to consider different forms or media of data record for presenting a certain reality, such factors influence the capacity or conditions of accessing or using research data. Enabling data to transit among technologies, keeping its qualitative aspects and its reproducibility and preservation aspects, as well as data analysis and interpretation, requires an upload of what might be called data competence or a satisfactory level of so-called data literacy.

Baykoucheva (2015) presents this literacy as skills for data reading, interpretation and understanding, emphasizing the importance of incorporating programs aimed at literacy in data and information, with an index of content and specific competences. The author points out that many aspects that have already been researched with the aim of data literacy are close to or encompass statistical literacy, dealing with applying critical thinking to descriptive statistics. Calzada Prado and Marzal (2013), in turn, point out the emergency of training in competences beyond the statistical scope, encompassing competences for data acquisition, evaluation, treatment and processing, analysis and interpretation tasks. In this matter, it is essentially the understanding and domain of the technology universe that support such tasks.

9.4 TECHNOLOGIES AND THEIR FAIR CRITERIA

According to Björk's (2007) scientific communication model, research data can be sought and shared in repositories, depending on the type and moment in which the research is. Thus, repositories become the most appropriate information system to offer functionalities as database deposit and retrieval. Likewise, They can also offer functionalities for research databases management, involving other services.

However, repositories were originally considered Open Access as the Green Route (Harnad *et al.*, 2004), in which copies of articles published in magazines were freely available, so Weitzel (2006) considered repositories as a second source. With the possibility of managing research data, repositories have a different meaning, adding appropriate functionalities for this type of digital object.

Verification of software suitability for database management can be done in several ways, with several models of evaluation, depending on the purpose. In the study herein, compliance with FAIR principles, acronym for *findable, accessible, interoperable, reusable* is used. For that matter, it is considered:

a. **findable**: enables description by metadata and allows the use of persistent identifiers;

b. **accessible**: can be accessed by humans or machines, offer clear licenses and provide communication protocols;

c. **interoperable**: metadata standard can be understood by machines and functionalities for understanding databases;

d. **reusable**: database described to enable its rescue.

Evidently, FAIR basic criteria have developments, expanding their conceptual coverage. Likewise, data management based on these principles involve certain activities, as the data curation issue, which goes beyond purely tools themselves, involves activities, methods, and standards. Thus, repositories are instruments, as indicated in Bjork's (2007) scientific communication model, for depositing and retrieving databases, taking the form of sharing.

Thus, discussing Technologies under the aspects related to FAIR criteria, each criterium, including its refinements as presented by Henning *et al.* (2019), become requirements that software for repositories must meet. By these means, it is possible to discuss FAIR criteria under the observation of technology, in order to propose an understanding of the theme, for technology offers tools, while criteria represent a conceptual basis, which enables tools evaluation.

The Findable criterium initially treats databases and their metadata. Thus, in order to meet this criterium, repositories have to provide support for depositing databases, regardless of their format or type, and adopt metadata standards that can be accessed by people and machines. Most of the current technologies for building repositories generally meet this requirement, but some observations are required.

Depositing databases requires some curation aspects, such as verifying if there is no virus, or even integrity mechanisms. However, in the findability recommendations, databases have lower emphasis in relation to metadata, mainly with the possibility of having this metadata indexed by search engines. For metadata, basic questions such as the use of persistent identifiers are mandatory. This point is relatively met by identifier systems such as *Handle* or *Digital Object Identifier* (DOI). However, Technologies must enable the application of enriched metadata devoted to some curation activities, which can present challenges for depositors, as Technologies are additional Fields to be implemented. Thus, it is noticed that the findable recommendation affects some points related to metadata quality, which goes beyond the tool.

This view, in which the location has aspects with a lower impact on technology, can be supported by Monteiro e Sant'Ana (2020) study, which presents technological solutions to meet FAIR criteria of accessibility. For the authors, FAIR principles were implemented in the infrastructure of CLARÍN research, including findable aspects through the use of standards currently implemented in several technologies for creation of repositories, such as the use of Dublin Core metadata standards, which have flexibility and minimum standards for interoperability, according to Open Archives Initiative – OAI.

In general, in order to meet the findable FAIR principle, technologies for repositories must allow the indexing through search engines, implement standards of flexible and interoperable metadata, which enable the use of persistent identifiers for databases, in order to the easily found. Thus, in case there are no reviews in these criteria, the majority of repository technologies meets them, to a higher or lower degree, as repositories, in their traditional function in scientific communication, have the function of making the access to their items easier.

For the accessible criteria, the technological aspects are greater, as they have relation to accessing databases through communication protocols or directly for their reusing. Databases deposited in repositories are in digital format depending on their type and, even if they can be considered raw data, they can present several formats. Thus, repositories can provide the files to be downloaded or streamed, depending on the permissions. In the streaming case, repositories need to offer optional functionalities in face of the various formats that research databases can take, such as tables, spreadsheets, audios, videos, texts, and others.

Access to resources occurs through the offer of access, enabling the interaction using communication protocols. These criteria are related to the infrastructure in which repositories will be hosted. This criterium is associated to the use of free tools for creating environments, such as the use of operational systems, application servers and free databases. It is not only using the free software for creating the repository, but having the whole environment built with free technologies.

The only point related to technology for creating repositories and the access criterium are related to metadata preservation, even that the database is not available anymore. This point must guide the removal of databases from repositories by any reason, requiring that metadata be maintained. Thus, it can have an impact on metadata, requiring the presence of field indicating the status of the database.

The criteria related to the interoperability are technological. Since the beginning of the open archive initiative, still at the end of the last century, many softwares have been implementing the *Open Archives Initiative* (OAI) protocol in its versions *Protocol Metadata Harvesting* (PMH) or *Object Reuse and Exchange* (ORE), based on *eXtensible Markup Language* (XML). However, with the technological Evolution, Other forms of notation can be used for the interoperability, such as *JavaScript Object Notation* (JSON) technology, or even with the use of *WebServices* services, with good results and more flexibility in collecting metadata.

In this matter, the idea of interoperability is bigger in FAIR criteria than in open archives, aimed at the possibility of metadata and digital objects exchange. Thus, technologies for implementing research data repository need to be adjusted to meet FAIR. Moreover, it requires developing supporting structures such as ontologies and thesaurus, standardizing content of metadata fields with controlled vocabularies. Likewise, it will require from data producers that They deposit data, data dictionary and, in the future, data narratives.

Even if the interoperability criteria have technological aspects, strong recommendations aimed at metadata are noted, such as in issues of relationship among databases. Thus, if a database is questioned, the original source can be cited indicating its provenance. Likewise, if a database is generated based on other databases, this must be indicated. All this information must be provided in metadata.

For databases to be reusable, complete descriptions are necessary, going beyond metadata. Therefore, FAIR criteria for reuse suggest developing narratives of data that comprise limitations, context, ways of collection, among others. The recommendations for reuse have procedural aspects, in which the information above-mentioned will allow their reuse.

It is worth noting that FAIR criteria guide data producers who desire to share and adopt open data and open science initiatives, researchers who promote transparency through research data dissemination. In that matter, FAIR criteria is presented as a basis for general orientation for all the initiatives that want to adopt data sharing as part of the research activities.

9.5 TECHNOLOGIES FOR FAIR REPOSITORIES

If FAIR criteria have guiding aspects, the *FAIR Data Point Specification (FDPS)* document presents the requirements for developing or evaluating technologies oriented to meet FAIR principles. Thus, according to specifications described in the document referred, a repository will be created to meet metadata and data that can be classified as FAIR. Likewise, the document can be used for evaluating existing software sets, mainly free and open-code softwares, whose purpose is modifications and adaptations.

The repository designed according to FDPS model is, in general, aimed at natural sciences, sometimes focused on biology, which presents certain restrictions. However, with adaptations, it can be used for other sciences bearing in mind the specificities of each field. The intention is to promote technologies that enable the creation of repositories that can interoperate metadata and databases in order to enable distributed research.

Thus, the repository designed by FDPD has two well-established sets of functionalities aimed at retrieval and deposit of databases. These points are in line with current repositories aimed at disseminating technical and scientific documentation. In this matter, conceptually, FDPD data repository presents aspects similar to digital repositories, only specialized in generating databases.

For database retrieval, there are simple processes such as find and access databases. The deposit in databases, in turn, can be manual or automatic. Regardless of processes, whether retrieval or deposit, it is required that the repository offers simple and standardized functionalities, so that users can use the platform for offering or using databases.

The initial requirements to meet FDPS start with the findability features of databases, that is, discovery features, which help users to find out where these databases are, involving issues related to indexing by search engines and others. Similarly, repositories must offer simple search tools, but with indexing made up of functionalities that support the datasets maintained by them.

Once the desired database is found, the access process begins. Thus, the licenses for accessing and using data must be presented, preferably those indicated by FAIR. Likewise, the repository must guide the standardization of format used for data, with strong indication for the use of free formats, so that standardized access is possible. Thus, users who which to aggregate databases from different repositories have their work facilitated, promoting the reuse.

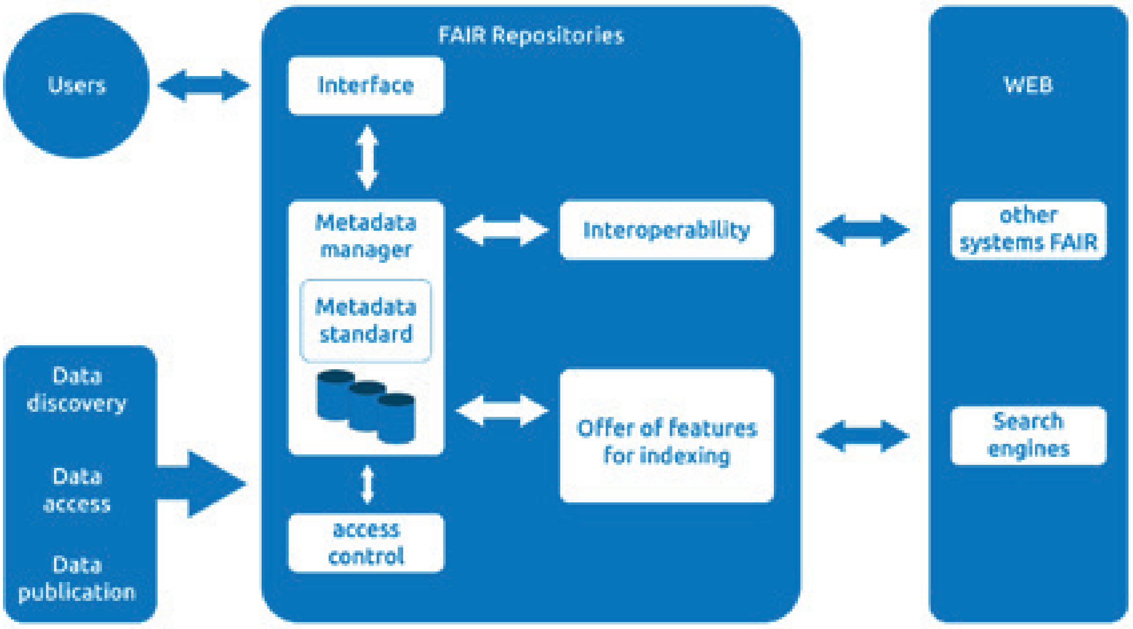
Data deposit, also called data release, consists of a set of functionalities that help authors to provide datasets through repositories, with a great variety of options to provide access, versioning, status, among others. So, it aims at meeting the specificities presented by the databases in relation to the restrictions of use, time of research and others, but allowing users to find those databases.

Finally, FDPS data repositories must enable the Generation of use statistics. Indicators must be offered as strategic information for decision-making related to infrastructure management, for instance. Likewise, it is possible to know the relevance of databases, generating information for the authors about their data reuse. Thus, statistics are useful even to justify the repositories and their investments.

On the other hand, as all repositories, the systems in line with FDP must have an architecture that makes it possible to manage items composed of databases and their metadata, which can be manipulated by both humans and machines. Thus, it must have a Web interface in which users can Search, deposit and retrieve databases through the databases' metadata. Likewise, it must provide access so that machines can retrieve information. The access must be controlled by permissions in order to protect sensitive data. Metadata must follow FAIR standards, offering information on five levels: about the repository itself; about the brochure (database sets that make up the collection); about the database; about distribution; and about data records.

Simply put, the architecture of a repository that meets FAIR principles (Figure 1) helps human users through a Web interface, offering interoperability for Other systems and allowing search engines to index their metadata. The access to metadata and databases must be controlled in order to enable staggered dissemination due to different levels of data sensibility. Metadata follows standards that facilitate the integration, finding, and description of databases. In short, FDPS repository makes data finding, access and deposit easier.

Figure 1 – Architecture of a repository that meets FAIR principles



Source: Elaborated by authors based on *FAIR Data Point Specification* (FDPS).

The specifications for FDPS repositories do not differ Much from the Technologies used for the creation of academic repositories, such as computerized systems. However, databases have specificities that require different treatments. As a result, the differences between databases and academic publications will distinguish from the systems, as well as the types of functionalities proposed.

9.6 FINAL CONSIDERATIONS

FAIR guidelines have features that change the way of analyzing science practices, as they change the focus of research outcomes, represented by articles, books, annals and others, to research data with the aim to store them to be shared. They change, in a certain way, the perspective of research data, as a social asset that must be shared as a way to contribute with other scientists.

For that matter, it requires informational infrastructure, made up of computerized systems, currently helped by research data repositories. Thus, in order to implement these systems, free and proprietary technologies are offered, meeting part of the criteria. However, the task of implementing FAIR principles in research data dissemination is, at a certain point, up to the managers and researchers, as the technologies aimed at the creation of academic repositories meet great part of the demand.

In this matter, it is emphasized that the change for using FAIR principles is behavioral, on the one hand, as it depends on researchers to deposit data from their studies in data repositories with open licenses, with well-described metadata and using free formats. Likewise, it depends on well-structured data repositories that implement FAIR principles, offered by institutions recognized in the academic world, ensuring the necessary curation.

For authors that adopt FAIR principles, the link for data in repositories can be inserted in the methodology section of the article, for example. Another valid option is the adoption of a metadata field in the magazine for inserting the link in the submission process, as many magazines have already done. Regarding the technology, softwares for creating data repositories, magazines, or even formats for article formatting meet the principles.

It may be that in the future not all FAIR principles will be met because the repositories still do not meet all the developments of the principles. Most repositories are based on the sharing process and not on curating their collection. Therefore, the challenges in relation to technological infrastructure still require studies

Possibly, the greatest challenge of research data management in regard to technologies is related to the ones that meet the needs of curation, since the dissemination is only the final part of the process, in which FAIR is contextualized. It is possible that FAIR, in the near future, adapts to the new demands when adopting criteria met by technologies.

REFERENCES

BAYKOUCHEVA, Svetla. **Managing Scientific Information and Research Data.** Burlington: Elsevier Science, 2015, 208 p.

BJÖRK, Bo-Christer. A model of scientific communication as a global distributed information system. **Information Research**, v. 12, n. 2, p. 307, 2007.

BORGERUD, Charlotte; BORGLUND, Erik. Open research data, an archival challenge? **Archival Science**, v. 20, n. 3, p. 279-302, 10 fev. 2020. DOI: 10.1007/s10502-020-09330-3. Available from: http://link.springer.com/10.1007/s10502-020-09330-3. Access on: 3 Nov. 2020.

CALZADA PRADO, Javier; MARZAL, Miguel Ángel. Incorporating Data Literacy into Information Literacy Programs: Core Competencies and Contents. **Libri**, v. 63, n. 2, jan. 2013. DOI: 10.1515/libri-2013-0010. Available from: https://www.degruyter.com/doi/10.1515/libri-2013-0010. Access on: 3 Nov. 2020.

COSTA, Michelli Pereira da. **Fatores que influenciam a comunicação de dados de pesquisa sobre o vírus da zika, na perspectiva de pesquisadores**. 2017. 269 f. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, Faculdade de Ciência da Informação, Programa de Pós-Graduação em Ciência da Informação, Brasília, 2017. Available from: https://repositorio.unb.br/handle/10482/23000. Access on: 9 Sept. 2020.

DIAZ, Pablo. Ethics in the era of open research data: some points of reference. **FORS Guide**, v. 1, n. 3, p. 1-18, jan. 2019. DOI: 10.24449/FG-2019-00003. Available from: https://forscenter.ch/fors-guides/fg-2019-00003/. Access on: 3 Nov. 2020.

HARNAD, Stevan; BRODY, Tim; VALLIÈRES, François; CARR, Les; HITCHCOCK, Steve; GINGRAS, Yves; OPPENHEIM, Charles; STAMERJOHANNS, Heinrich; HILF, Eberhard R. The Access/Impact Problem and the Green and Gold Roads to Open Access. **Serials Review**, v. 30, n. 4, p. 310–314, Jan. 2004. DOI: 10.1080/00987913.2004.10764930. Available from: http://www.tandfonline.com/doi/abs/10.1080/00987913.2004.10764930. Access on: 3 Nov. 2020.

HENNING, Patricia Corrêa; RIBEIRO, Claudio José Silva; SANTOS, Luiz Olavo Bonino da Silva; SANTOS, Paula Xavier Dos. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, v. 25, n. 2, p. 389–412, 26 abr. 2019. DOI: 10.19132/1808-5245252.389-412. Available from: https://seer.ufrgs.br/EmQuestao/article/view/84753. Access on: 9 Sept. 2020.

KOLTAY, Tibor. Quality of Open Research Data: Values, Convergences and Governance. **Information**, v. 11, n. 4, p. 175, 25 mar. 2020. DOI: 10.3390/info11040175. Available from: https://www.mdpi.com/2078-2489/11/4/175. Access on: 3 Nov. 2020.

MONTEIRO, Elizabete Cristina de Souza de Aguiar; SANT'ANA, Ricardo Cesar Gonçalves. Repositórios de Dados Científicos na Infraestrutura de Pesquisa: adoção dos princípios FAIR. **Ciência da Informação**, v. 48, n. 3, mar. 2020. Available from: http://revista.ibict.br/ciinf/article/view/4878. Access on: 3 Nov. 2020.

MURRAY-RUST, Peter. Open Data in Science. **Nature Precedings**, 18 jan. 2008. DOI: 10.1038/npre.2008.1526.1. Available from: http://www.nature.com/articles/npre.2008.1526.1. Access on: 3 Nov. 2020.

NATIONAL SCIENCE BOARD. **Digital Research Data Sharing and Management**. Arlington, Virginia: National Science Foundation, 2011. Available from: https://www.nsf.gov/nsb/publications/2011/nsb1124.pdf. Access on: 3 Nov. 2020.

NAUR, Peter. Computing versus human thinking. **Communications of the ACM**, v. 50, n. 1, p. 85–94, jan. 2007. DOI: 10.1145/1188913.1188922. Available from: https://dl.acm.org/doi/10.1145/1188913.1188922. Access on: 3 Nov. 2020.

PAMPEL, Heinz; DALLMEIER-TIESSEN, Sünje. Open Research Data: From Vision to Practice. *In*: BARTLING, Sönke; FRIESIKE, Sascha (orgs.). **Opening Science**. Cham: Springer International Publishing, 2014. p. 213–224. DOI: 10.1007/978-3-319-00026-8_14. Available from: http://link.springer.com/10.1007/978-3-319-00026-8_14. Access on: 3 Nov. 2020.

SALES, Luana Farias; COSTA, Michelli; SHINTAKU, Milton. Ciência aberta, gestão de dados de pesquisa e novas possibilidades para a editoração científica. *In*: SHINTAKU, Milton; SALES, Luana Farias; COSTA, Michelli (orgs.). **Tópicos sobre dados abertos para editores científicos**. 1. ed. Botucatu, SP: ABEC, 2020. p. 13–21. DOI:

10.21452/978-85-93910-04-3.cap1. Available from: https://www.abecbrasil.org.br/arquivos/Topicos_dados_abertos_editores_cientificos.pdf#01. Access on: 3 Nov. 2020.

SALES, Luana Farias; SAYÃO, Luís Fernando. Uma proposta de taxonomia para dados de pesquisa. **Revista Conhecimento em Ação**, v. 4, n. 1, p. 31–48, 30 jun. 2019. DOI: 10.47681/rca.v4i1.26337. Available from: https://revistas.ufrj.br/index.php/rca/article/view/26337. Access on: 3 Nov. 2020.

SAYÃO, Luis Fernando; SALES, Luana Farias. Algumas considerações sobre os repositórios digitais de dados de pesquisa. **Informação & Informação**, v. 21, n. 2, p. 90, 2016. DOI: 10.5433/1981-8920.2016v21n2p90. Available from: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27939. Access on: 9 Sept. 2020.

SOMPEL, Herbert Van de; LAGOZE, Carl. The Santa Fe Convention of the Open Archives Initiative. **D-Lib Magazine**, v. 6, n. 2, 2000. DOI: 10.1045/february2000-vandesompel-oai. Available from: http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html. Access on: 11 Mar. 2020.

SULEMAN, Hussein; FOX, Edward A. A Framework for Building Open Digital Libraries. **D-Lib Magazine**, v. 7, n. 12, dez. 2001. DOI: 10.1045/december2001-suleman. Available from: http://www.dlib.org/dlib/december01/suleman/12suleman.html. Access on: 3 Nov. 2020.

WEITZEL, Simone da Rocha. O papel dos repositórios institucionais e temáticos na estrutura da produção científica. **Em questão**, v. 12, n. 1, p. 51–71, 2006.

ZHU, Yangyong; XIONG, Yun. Towards Data Science. **Data Science Journal**, v. 14, n. 0, p. 8, 22 maio 2015. DOI: 10.5334/dsj-2015-008. Available from: http://datascience.codata.org/article/10.5334/dsj-2015-008/. Access on: 3 Nov. 2020.

How to cite this chapter: SHINTAKU, Milton; APPEL, André Luiz; OLEIVEIRA, Alexandre Faria de. Technologies for research data management according to FAIR. *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. 9, p. 113-124. DOI: 10.22477/9788570131959.cap9.

10. DATA INTEROPERABILITY AND THE INFORMATION TRANSDUCTION ENCAPSULATED IN DATA ACCESS

Ricardo César Gonçalves Sant'Ana139

10.1 INTRODUCTION

Reducing the distance and barriers between data and informational needs is the challenge herein. And when the data access scenario is considered, this issue also includes the location, interpretation, and use not only by individuals, but mainly by machines.

This demand implies turning these data into readable content, in situations and moments distinct from those in which They were created, generating the need to adopt standards and, consequently, principles and guidelines that, When shared, allow for the reuse of data collection.

Environments with high demand for data access, such as corporative environments or even environments related to public management, are composed of hundreds or even millions of systems, developed either internally or externally, setting a complex and diversified scenario. Such systems require integration of their data so that their effective use is enabled. (Reeve, 2013; Sant'ana, 2009; Dyché; Levy, 2006). However, much of the focus on data management is applied to the processes of collection, storage, and availability in the systems to the detriment of data flow among the different structures (Reeve, 2013). This space among systems tends to increase its complexity exponentially with the increase in data sources considered in the environment, and it is usually served by data interfaces (resources developed to enable data flow among systems).

Thus, the search for data availability in the right place, in the right format, and adherent to the informational needs intended to be met gains increasing relevance, which places data integration as a central condition to the success of data access (Kelleher; Tierney, 2018). This data integration implies, in its turn, in processes of transforming these data in an increasingly automated way so that They can be treated as a single set, generating the need of informational transduction (Sant'ana, 2019), which in its turn, demand definitions that derive from both technical and context knowledge (Reeve, 2013; Shkedi, 2019). In the dimension of knowledge about the context, lines of confrontation to integration challenges emerge, such as those that consider, for instance, the use of ontologies; this issue is beyond the scope of this text, though.

Considering the scenario of multiple systems, for multiple instances, such as in the case of sharing data among academia, industries, funding agencies and academic publishers, it is possible to predict the high level of com-

[139] Associate Professor in Management Information Systems, Associate Professor at the São Paulo State University - UNESP, ricardo.santana@unesp.br

plexity that is presented. Seeking the development of shared means of expanding the reuse of data collections for situations such like this, representatives of these bodies met in a workshop held in Leiden, Holland, in 2014, called 'Joint Designing a Data Fairport' to "design and endorse, by mutual agreement, a concise and measurable set of principles named FAIR Data Principles." (Wilkinson *et al.*, 2016), denomination resulting from the concepts: Findable, Accessible, Interoperable and Reusable. A differential proposed by FAIR Principles guidelines is in the fact that they "emphasize specifically the improvement of machines capacity to automatically find and use data, besides supporting their reuse by individuals"; it is also noteworthy the intention that the guidelines are applicable to "algorithms, tools and workflows that led to these data" (Wilkinson *et al.*, 2016).

FAIR Principles, after adjustments and improvements, are presented with four principles, each one of them with their criteria: four related to Findable concept, two related to Accessible, one related to Reusable and three related to Interoperable. The latter, which is the focus of this chapter, are the following: (Wilkinson *et al.*, 2016):

I1. Metadata with formal, accessible, shared and widely applicable language for the knowledge representation;

I2. Metadata with vocabulary following FAIR principles;

I3. Metadata includes references qualified to other metadata.

These are generic principles, but they point to guidelines that can help increase the interoperability potential of datasets. Aspects such as the need for formalism, that is, meeting pre-established standards, can help to minimize divergent definitions of the same content. The 'accessible' concept, still provided in I1 principle, can also support the need of these standards and rules to be shared among those involved, and that they are also the target of dissemination strategies for their understanding, elements that lead to the concept, also part of I1 principle, 'shared'. All these factors could not be relevant if the feasibility of such definitions were not considered, which leads to the 'applicable' concept that completes I1 principle.

I2 principle aims at the issue of vocabulary used in metadata defined for datasets, recursively pointing to the Other FAIR principles. This issue aims at the issue of vocabularies used in metadata defined for datasets, recursively pointing to Other FAIR principles. It is a quite broad issue and requires remembering that vocabularies cannot always meet the specific needs, which can lead to publications of extensions of existing vocabularies or even the creation of new vocabularies (FORCE11, c2021), which is still a complicating factor.

I3 principle indicates the need of qualified references among metadata (Wilkinson *et al.*, 2016), which points to the need that machine resources can perform operations directly on data collected, which in its turn requires that metadata "must be syntactically parsable and semantically accessible by machine" (FORCE11, c2021). FORCE11 also points out that the syntax and 'semantic' of data models and formats used for metadata must be "easy to be identified and used, analyzed or translated by machines", and this is one of the guiding elements for the argumentation presented in this text

In this same line of analysis, there is a flexibilization provided in the proposal of FAIR Principles through the possibility of definitions emerging in a bottom-up movement,

> if a provider can prove that an alternative data model/format is unequivocally parsable as one of the FAIR formats adopted by the community, there is no particular reason such a format cannot be considered FAIR. Some types of data may simply not be 'capturable' in one of the existing formats and, in this case, maybe only part of the data elements can be analyzed (FORCE11, c2021)

Such flexibilization would increase the potential for adherence to the specificities inherent to the large number of situations and contexts in which data originate, while at the same time carrying with it the complexity from which the motivation for the proposal for FAIR principles originated. This effect is even foreseen by FORCE11 itself When it proposes that: "the ideal situation is restricted FAIR data release to the minimal possible of formats and standards adopted by the community", considering that it would be necessary to offer solutions to the new demands: "FAIRports will offer more and more guidance and assistance in these cases" (FORCE11, c2021).

Even considering that it is not a pre-requirement for determining the data adherence to FAIR Principles (Wilkinson *et al.*, 2016), machine access must be sought with the greatest autonomy possible – access and interpretation to such a point that it is possible to transform data collected in a new dataset more adherent to each need (Reeve, 2013).There is no way not to consider minimum levels of machine processing as a condition without which the data access process would not be able to cope with the data to which we are submitted, or as provided by FORCE11 (c2021), when it states that "providing machine-readable data as the main substrate for knowledge discovery [...] that works smoothly and sustainably is one of eScience greatest challenges".

But Where do such various informational elements necessary so that data can be properly collected and transformed for use come from? Much of it comes from their own fragmented essence, resulting from the necessary structure, native or obtained after treatment, but Always a requirement so that the algorithms can establish in a univocal and detailed way, step by step, what the machine must do in the processing of contents.

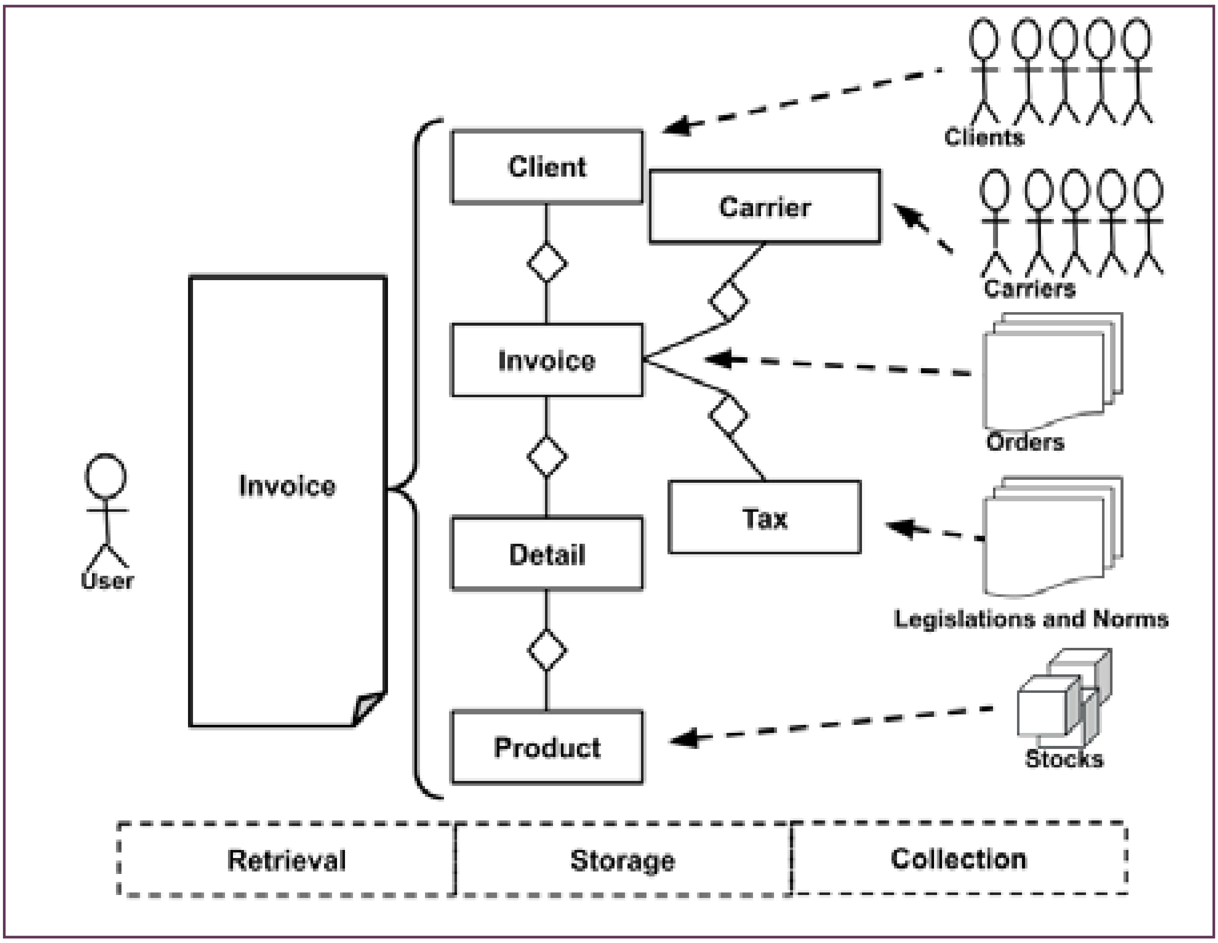
10.2 THE FRAGMENTED NATURE OF DATA AND FAIR PRINCIPLES

By their nature, machine resources are only capable of executing absolutely detailed and precise instructions, hence the need to establish, exactly and formally, what should be done with each informational particle of the contents to be treated. This process algorithmizing leads to a necessary and inevitable fragmented nature of data, so that new informational layers can be established with metadata that support syntactic and semantic elements to each of the fragments, thus composing a minimal structure of meaning – the triad: entity, attribute and value <e,a,v> (Santos; Sant'ana, 2015) – which, in turn, provides the viability of algorithmizing of machine treatment of contents.

This fragmentation generates, in the phase of collection (Sant'ana, 2016), the need to allocate specific values, related to each one of the transactions and registered facts, in specific 'attributes', which in turn, will be linked to 'entities' related to each one of the information identified as relevant. This link emerges from logical mapping of these entities, respecting principles such as those related to data normalization, avoiding redundancies and bringing coherence to datasets.

Thus, if we consider, for example, a document referring to a sales transaction, an invoice (Figure 1), we will have all the data trajectory, from the collection phase (Sant'ana, 2016), with the identification of information about the client, order details, products, and quantities involved, carrier responsible for delivery, classifications and tax calculation, among other information. This information is then recorded, persisted, in the respective semantic structures (entities) with its respective label (attributes), and related to each other in such a way that it is possible for its visualization as a document.

Figure 1 – Fragmented data structure



Source: Designed by author.

From this document, which is available in the retrieval phase, it is possible for the user to visualize the transaction data, therefore, the fact. In addition, it is this result from the composition of respective data from each entity (dataset) available as a document that will be shared with the others involved, such as the client himself, tax authorities, carrier, and the Other systems of the organization itself, such as the financial system, accounting system, stock, among others.

From this fragmentation, two points of reflection emerge in this text: the informational transduction and the encapsulated complexity.

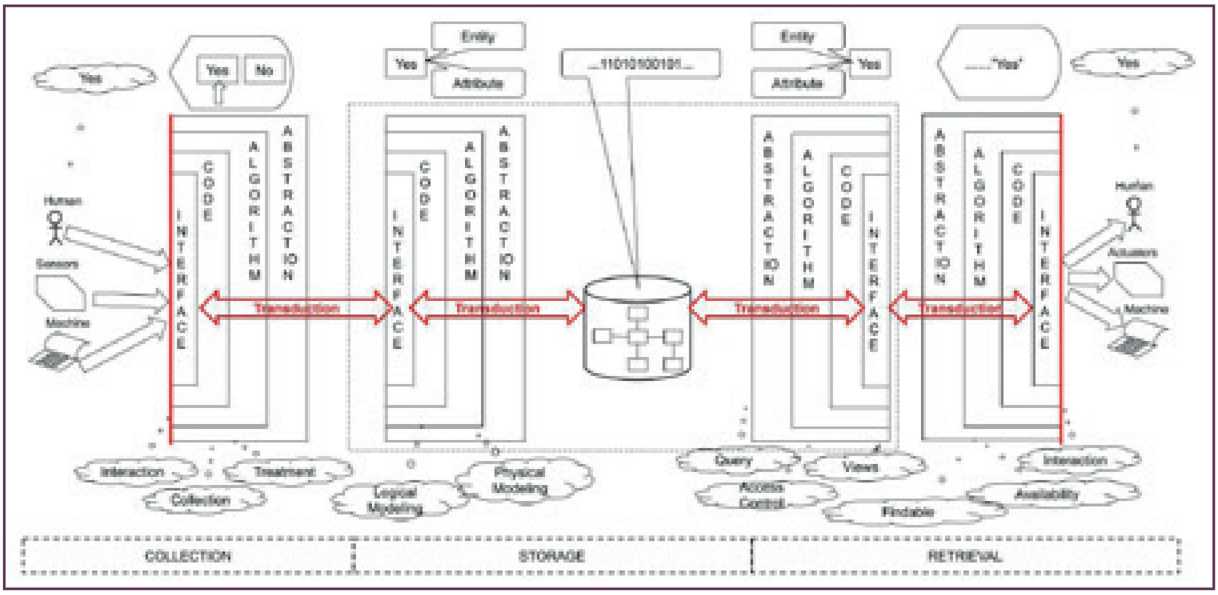
The contents, flowing between source and user, collected, stored and available for retrieval in different data life cycles (Sant'ana, 2016), undergo changes both in energy and format scope and even the content itself. These informational transductions allow for the user's informational needs to be met, meeting the demands through greater adherence to the use context, such as personalization, adequacy, adaptation, resulting in the lowest cost possible when accessing data.

Even though it is fully dominated by those who work in the abstraction layers closest to the machine-analysts, developers and administrators, whether in the programming (software) dimension or in the data dimension – most users do not have, and could not even have, the perception of this fragmented structure to which data are submitted so that they can be used, as such complexity would make the computational resources unfeasible.

10.2.1 Informational Transduction

Data path is long and complex, starting from its apprehension from the fact, going through the several transformations imposed by the interfaces and adjustments to the models of data structure, reaching its records in the digital supports so that, finally, They can be recomposed in a document format. (figure 1). Each one of these changes (Figure 2) implies conversions not only in format and content, but also in energy. Such changes, herein defined as Informational Transductions (Sant'ana, 2019) are necessary so that mediating mechanisms can treat the contents; however, They increase the complexity of the process, making its understanding more difficult, which would make it impossible to use the systems involved

Figure 2 – Informational Transduction in Data Access Source: Designed by the author.



Source: Designed by author.

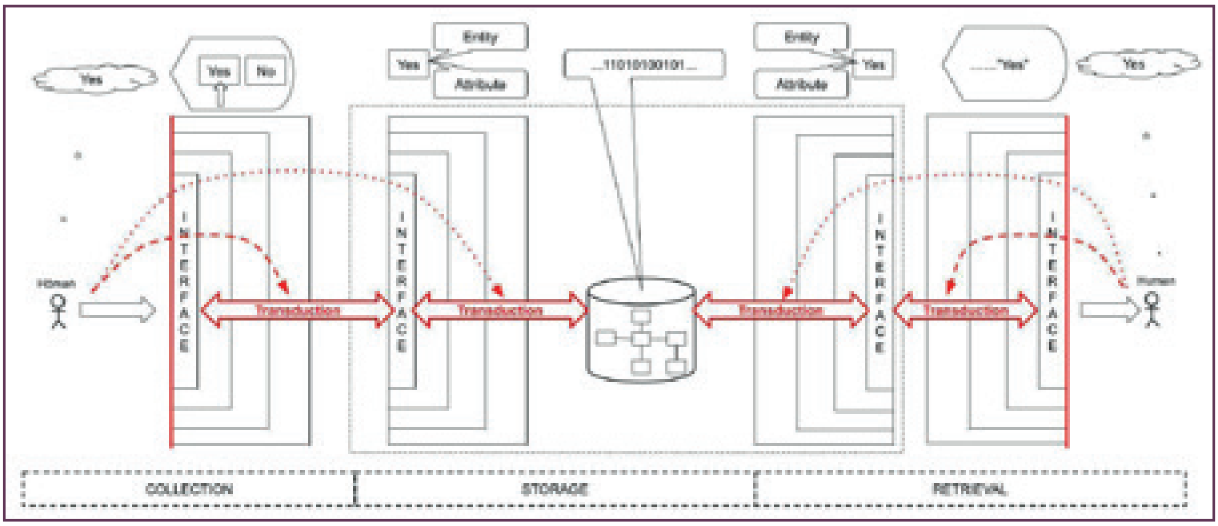
This eventual barrier is overcome by hiding non-essential details from users, which originates the second point of reflection in this text: complexity encapsulation.

10.2.2 Complexity encapsulation

One of the main factors for a system to be viable in its use is the learning curve required from its users. A very illustrative example is the adhesion to the Internet, which in its first years of access to the public, presented a great complexity of use, requiring simple access to a certain content to use complex commands, true lines of code, with information on address and operation, for example, always with a high level of syntactic formalism. This model kept the big public at bay, and only 'beginners' in technology took the risk of using it. This barrier was broken with the proposal of the Web model, which concealed such developments necessary to the operations, through graphic interfaces that allowed a simple user action, in the recent and so far, little used mouse, to be converted 'internally' in complex commands performed by the machine.

This same concealment process occurs in all spheres of technology use and, thus, this complexity encapsulation (Figure 3), allows the user to concentrate only on elements strictly necessary for his interaction with the systems. If we return to the example presented in Figure 1, we can infer that the responsible for data input in the system focuses on data content, without realizing the fragmented way in which they are converted in entities, and even less how they will be physically treated in digital supports. Information such as the hard disk area where the information will be recorded or how the device memory will treat these contents, or even how different programs will interoperate, such as the Invoicing System and the Database Management System (Figure 3).

Figure 3 – User inability on Informational Transduction processes



Source: Designed by the author.

This process has advanced so much that today we have the feasibility of direct customer interaction with the interface systems of companies, the e-commerce, which combines the ease of use and ubiquity of the Internet, linked to the evolution of interface systems, exclude the participation of those who until then were responsible for data selling and input in the systems.

Complexity encapsulation leads to a perception of content, treated by systems, totally based on data visualization, under the form of documents or reports, always aimed at searching for adherence to informational need, concealing, therefore, the structures of entities used and the linking ways that allow relationships among data entities. This user's inability to informational transductions keeps him away from eventual actions and definitions necessary so that the interoperability is viable.

10.3 REFLECTIONS ON ANALYSIS AND FUTURE DEVELOPMENTS

As part of FAIR principles, the increase in the interoperability, so relevant for full data access, is deeply affected by the inability of individuals involved in the process of several physical and logic transductions among the data collection, storage, and retrieval phases.

It is reinforced here the difficult perception, for those involved, of the structurally fragmented nature of data and the need to group them, which in turn, bring their features to layers of abstraction that incorporate semantics to these data and allow, finally, their interpretation.

Logical mappings, in their subsequent layers, incorporate data on data (metadata) that need, among other purposes, to allow the interpretation by humans, and increasingly, by machine treatments, which leads to the need for sharing physical, logical and semantic-complex standards.

Data collected in their respective environments receive treatment and are prepared to be stored, predicting, in most cases, their use in the context established at the moment of collection. However, their use tends to be increasingly disseminated, and it is necessary that these semantic layers, aggregated to the data, can be used by unforeseen or inexistent contexts in the moment of collection and storage. On the other hand, factors such as those related to possible limitations to these data also require that these openings, for unforeseen uses, are explicit, not only to the holders of the resources involved in the data life cycle, but also to users and eventual referenced by these data.

Information Science can, and must, participate in the process of identifying factors involved in informational transductions' encapsulation necessary to the process of data, and collaborate, not only in the search for improvement and increase in the potential for integrating informational content in data, but also, and mainly, contribute for the dissemination, in the Society, of the potential of data use, which once aggregated and meeting interoperability requirements, can represent a higher value than those represented by sets when individually considered.

REFERENCES

DYCHÉ, Jill; LEVY, Evan. **Customer Data Integration**: reaching a single version of the truth. Hoboken, New Jersey: John Wiley & Sons, 2006.

FORCE11. **Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing:** version b1.0. c2021. Available from: https://www.force11.org/fairprinciples. Access on: 07 oct. 2024..

KELLEHER, John D.; TIERNEY, Brendan. **Data Science**. Cambridge: The MIT Press, 2018.

REEVE, April. **Managing Data in Motion**: Data Integration best practice techniques and technologies. Waltham, EUA: Elsevier, 2013.

SANT'ANA, Ricardo César Gonçalves. **Tecnologia e gestão pública municipal**. São Paulo: Cultura Acadêmica, 2009. (Coleção PROPG Digital - UNESP). Available from: http://hdl.handle.net/11449/109104. Access on: 07 oct. 2024.

SANT'ANA, Ricardo César Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, Londrina, v. 21, n. 2, p. 116–142, dec. 2016. Available from: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940. Access on: 29 dec. 2016.

SANT'ANA, Ricardo César Gonçalves. Transdução Informacional: impactos do controle sobre os dados. *In*: MARTÍNE-Z-ÁVILA, D; SOUZA, E. A.; GONZALEZ, M. E. Q. (orgs). **Informação, conhecimento, ação autônoma e big data**: continuidade ou revolução? Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2019. p.117-128.

SANTOS, Plácida L. V. Amorim da Costa; SANT'ANA, Ricardo César Gonçalves. Dado e Granularidade na perspectiva da Informação e Tecnologia: uma interpretação pela Ciência da Informação. **Ciência da Informação**, Brasília, v. 42, p. 199-209, 2015. Available from: https://revista.ibict.br/ciinf/article/view/1382. Access on: 07 oct. 2024.

SHKEDI, Asher. **Introduction to Data Analysis in Qualitative Research**: practical and theoretical methodologies with use of a software tool. [*S. l.: s. n.*], 2019.

WILKINSON, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, [*s. l.*], v. 3, n. 1, p. 1-9, 2016. Available from: https://doi.org/10.1038/sdata.2016.18. Access on: 15 aug. 2020.

How to cite this chapter: SANT´ANA, Ricardo César Golçalves. Data interoperability and the information transduction encapsulated in data access. *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. 10, p. 125-132. DOI: 10.22477/9788570131959.cap10.

11. DEVELOPMENT AND APPLICATION OF INTEROPERABILITY STANDARDS FOR SCIENTIFIC DATA REPOSITORIES: IBICT AND CNPQ REPOSITORIES

Lucas N. Paganine[140]
Washington L. Ribeiro de Carvalho Segundo[141]
João L. R. Moreira[142]

Abstract: *In the 4th Plan of the Open Government Partnership (OGP), specifically in Commitment 3 (Establish governance mechanisms for scientific data for the advancement of Open Science in Brazil), proposed by the Brazilian Agricultural Research Corporation (Embrapa) and coordinated by the Comptroller General of the Union (CGU), comes the landmark 8, "Proposal of interoperability standards for research data repositories", this coordinated by the Brazilian Institute of Information in Science and Technology (IBICT), which resulted in the document Interoperability standards for data repositories of research search. The document aims to develop a minimum set of metadata descriptions for research data, making extensions to specific domains of knowledge. The methodology comprises two stages: the first characterized by the establishment of a general multi-thematic metadata scheme standard from the OpenAIRE Guidelines v4 and the metadata requirements of the Fair Data Point specification; and the second through the development of extensions by area of knowledge, from the Metadata Directory maintained by the Research Data Alliance (RDA), and also through the records of thematic research data repositories in the Registry of Research Data Repositories (re3data). The final results showed a core of 13 mandatory metadata and 3 extensions for specific areas: Biology, Agriculture and Social Sciences. An analysis of the document in question will be carried out and the application cases of the Aleia and LattesData repositories, under development, will be presented. It is noteworthy then that this is not a job that is exhausted in the final application, on the contrary it needs constant development.*

Keywords: *Research Data repositories. Scientific data. Metadata standards.*

[140] Bachelor in Library Science with graduate degree in Public Management, Brazilian Institute of Information on Science and Technology (Ibict) - lucaspaganine@ibict.br

[141] PhD in Computer Science, Brazilian Institute of Information on Science and Technology (Ibict) - washingtonsegundo@ibict.br

[142] PhD in Computer Science, University of Twente - j.luizrebelomoreira@utwente.nl

11.1 INTRODUCTION

Open Science (OS) is an umbrella term that covers several aspects such as Citizen Science and notebook science sharing. It emerges as the development of Open Access Movement (OA), dealing with the opening of scientific processes to the population in general, thus following collaboration and transparency principles. In OS context, it is emphasized the increasing demand for scientific data sharing, using several tools for this purpose, among which scientific data repositories stand out.

These repositories are tools for treating, organizing, dissemination and preserving digital objects, in this case the scientific data. However, due to the needs of the several areas of knowledge and different institutional realities that implement the repositories, a lot of standards for describing datasets stored emerge.

The importance of a research on these different existing patterns and models is highlighted, aiming at developing a central descriptive scheme that also allows meeting the specific needs of different areas of knowledge, through extensions of this central pattern, thus enabling the interoperability among repositories from different thematic domains.

This study was then developed within the scope of the Open Government Partnership (OGP), an international initiative initiated in 2011 with the purpose of encouraging transparency as a government practice, in particular, the access to public information and the active cooperation with society. Such purpose is very much in line with OS.

OGP acts through National Actions Plans committed to Open Govern practices. As a result of the execution of these plans, reports expressing the progress in meeting the proposed goals are elaborated.

In 2018, Brazil developed its 4th National Action Plan, with 11 commitments, among which the Commitment 3, which aimed at "Stablishing mechanisms for research data governance for the development of Open Science in Brazil" (RNP, 2018). This commitment, known as Commitment for OS, was under Embrapa's coordination, but with the participation of several institutions, most of them governmental.

The commitment was organized in nine landmarks, with Landmark 8 described as a "Proposition of interoperability standards for scientific data repository" (RNP, 2018), coordinated by Ibict, but with the collaboration of the National Education and Research Network, Twente University, National Nuclear Energy Commission and National Council for Scientific and Technological Development (CNPq). One of the results from the Landmark was a Guide with the proposal of "Interoperability Standards for research data repositories" which are applicable to any research data repository that wants to promote the interoperability and opening stored scientific data. Interoperability criteria are defined in the document, guiding the building or improvement of scientific data.

Considering this scenario, it is worth it to deeply explore the product of the landmark in question, in special When considering its purpose of "develop and apply a minimal set of descriptions for scientific data, making appendices for specific knowledge domains, based on existing international standards and guidelines" (Paganine *et al.* , 2020).

Therefore, the methodology used herein will begin with the description of the development of the result in question as well as the presentation of results, and at the end, a description of the stages and application process of this document in Ibict and CNPq scientific data repositories is provided.

The document in Paganine *et al.* (2020) is divided in 2 parts, one with a general metadata set, and the other with appendices for specific areas of knowledge. Documents from well-established international guidelines for scientific data repositories are used as main references. They were: the OpenAIRE Guidelines for Data Repositories, concomitantly with the OpenAIRE Guidelines for Repository of Scientific Publications; and the metadata set described by *Fair Data Point* (FDP) *framework.* OpenAIRE guidelines are extensively adopted internationally, however, research dealing with semantic interoperability indicates the importance of extensions to meet FAIR principles FAIR (Wilkinson *et al.*, 2016). This extension also cited by Santos *et al.* (2016) is explored in the FDP *framework*.

For the classification on the extensions to specific areas of knowledge, for organizational reasons internal to Brazil, the CNPq tables of knowledge areas from Frascati (OECD, 2015) research manual and the area division one used by Data Curation Centre (DCC) and RDA, the *Deutsche Forschungsgemeinschaf* (DFG, 2020). The complete comparative table is found in the original document resulting from the landmark.

Regarding the survey of standards for the extensions for specific knowledge areas, it started with Metadata directory[143], a tool maintained by RDA, followed by subsequent analysis of metadata standards used in thematic repositories (they cover only a certain knowledge area) and institutional and/or multi-thematic data found in the Registry of Research Data Repositories (re3data).

At the end, the standards found in the surveyed repositories were checked, in search of which appendices can complement OpenAIRE added to the FDP standard for the 4 areas initially selected: Biology (due to its behavior in the knowledge trees surveyed), Agriculture (due to the importance of the area nationally), Health (due to the importance and pioneering action in Open Access Movement and other aspects related to the research) and Social Sciences (due to the general object of study of the institute where the research was developed, the Ibict). It is noteworthy that during the development of the proposal, specialists in the area were consulted, especially Fiocruz (Health) and Embrapa (Agriculture).

11.2 DEVELOPMENT

The guidelines presented in Paganine *et al.* (2020), also have as references the old DRIVER guidelines that were published in 2007 by the *Digital Repository Infrastructure Vision for European Research (DRIVER)* project and the *Guidelines for content providers: Exposing textual resources with OAI-PMH*, containing initial recommendations for interoperability. These recommendations were complemented by *OpenAIRE Guidelines for Literature Repositories*, which, at the moment this text was being written, were found in version 4 *(*OpenAIRE, 2018).

[143] rd-alliance.github.io/metadata-directory/

OpenAIRE guidelines are organized in three sections: The first one is introductory; the second describes the use of *OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)*, with orientations about it and; finally a general view of a profile for application. This guideline is composed by 4 metadata standards: *Dublin Core;* and its qualified version; *Datacite* and; Oaire (standard elaborated by *OpenAIRE* itself). Some vocabularies controlled for use are also specified, as for example the *Confederation of Open Access Repositories (COAR)*.

Regarding OAI-PMH protocol, it is a tool for exposing metadata through *Hypertext Transport Protocol (HTTP)* and *Extensible Markup Language (XML)* languages that allow the communication and interoperability among databases. It is noteworthy that regardless its use in the interoperability among systems, and its recommendation in OpenAIRE guidelines, it is noted a certain limitation of this tool, in its metadata set (15 elements of *Dublin Core*), which has been used since the beginning of 2000 (Garcia; Sunye, 2003). Other initiatives are increasingly more prominent, especially on Semantic Interoperability, such as W3C: PROV-O (model of generic data of *World Wide Web Consortium*) and the Data Catalog Vocabulary (DCAT) 2.0 (a vocabulary in Resource Description Framework-RDF), which adds classes for describing data services and putting them in line with FAIR principles.

OpenAIRE V4 guidelines establish 4 levels of mandatory fields to be filled out: Mandatory (M), when filling out is mandatory (Applied to 6 fields); Mandatory if applicable (MA), if filling out is mandatory only if the information in the field is part of registry (for instance, the name of the funding body is mandatory, in the case of the dataset results from financing) that is applied to 8 fields; Recommended (R), that is relevant and important but not essential (applied to 15 fields) and; Optional (O), which would only add value to the description, even if it is not necessary (applied to 3 fields). The guideline application profile is, in short, represented by the table that relates the thirty-two Fields of the guideline, with the instructions for filling in and the controlled vocabularies selected to specific fields

The second tool in analysis is the FDP, an independent and open-code web application, developed as implementation of specification reference of FDP itself[144]. These specifications guide softwares for repositories, dealing with metadata management, in particular about semantic Technologies such as RDF, thus being a complementary tool to a data repository software. A repository based on FDP delas with interoperability issues, enabling findability, accessibility, interoperability and reuse (FAIR principles).

The implementation of FDP reference uses an API REST with several functions: creation, storage, release of metadata thus allowing these metadata to be exposed, provided and available in accordance with FAIR principles. It also allows finding metadata of sets available and access them when they have an open-use license. Any data repository can adopt FDP metadata, thus also working as an FDP instance.

One of the main specifications of FDP is the specification of levels of metadata[145], which guides the application of a profile in RDF, reusing standardized semantic models. Therefore, the specification of levels of FDP metadata introduces an organization in four levels of metadata: first, the metadata repository itself; second, the catalog; third, the dataset

[144] github.com/FAIRDataTeam/FAIRDataPoint-Spec

[145] github.com/FAIRDataTeam/FAIRDataPoint-Spec/blob/master/spec.md

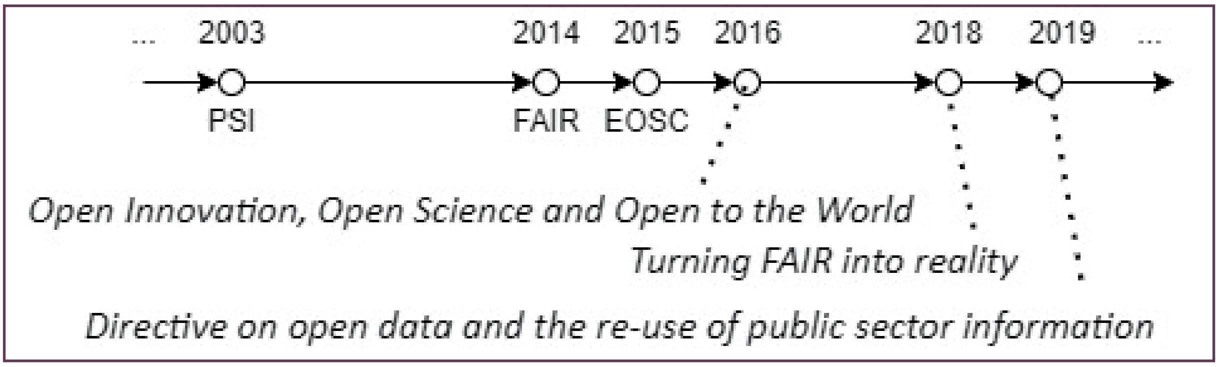
and; fourth, data distribution (the archives belonging to the set). A metadata repository can have one or more catalogs, each catalog can have one or more datasets, and each dataset contains one or more distributions.

FDP metadata standard is based on re3data schema[146] and DCAT vocabulary[147]. As previously noted, the metadata standard is organized in four levels and each property has two possibilities to be filled out: Mandatory, applied to ten fields, and Optional, applied to twelve fields. The levels describe, each one, a type of complex digital object that is possibly described, they are: the level of metadata repository, containing information on the data repository and the FDP itself; the level of metadata in the catalog, containing information on the collection, where each catalog represents a category (generally defined by domain); the level of dataset metadata, containing information on possible serializations of datasets, for instance, the individual archives that compose the datasets.

For example, the data repository B2Share (https://b2share.eudat.eu/) approaches catalogs through communities. Kinder Corona Studies (KiCoS) is one of the communities (catalogs) of B2Share and contains a series of datasets (in the tool represented as *records*); and a dataset can contain a series of distribution (*files*). The implementation of FDP metadata specification as a "semantic *proxy* (*wrapper*)" can add the aforementioned functionalities to the data repository software (Moreira *et al.*, 2019).

Also noteworthy in the European scenario actions and programs of the European Commission (EC) that deal with scientific data sharing and opening, such as: the European Open Science Cloud (EOSC), which dates from 2015; the Evolution of FAIR principles, which started in 2014. In 2016, the document Open Innovation, open Science and Open to the World was published; in 2018, the report turning FAIR into reality is released: Final Report and Action Plan on FAIR Data, and an increasing participation of EC in RDA is noticed; in 2019 there is the transformation of the guidelines of the Public Section Information (PSI), which had been edited in 2003, in the Guidelines for open data and reuse of public sector information (see Figure 1 below).

Figure 1 – Timeline of Other EC initiatives


...
2003
2014
2015
2016
2018
2019
...
PSI
FAIR
EOSC
Open Innovation, Open Science and Open to the World
Turning FAIR into reality
Directive on open data and the re-use of public sector information

Source: Designed by author.

[146] https://www.re3data.org/schema

[147] https://www.w3.org/TR/vocab-dcat-2/

The elaboration of general guidelines for scientific data repositories begins with a comparative analysis between OpenAIRE and FDP sets in Search of differences and similarities among the requirements. The full comparison can be found in the original document resulting from the landmark.

It was noted the possibility of equivalence among most of the fields compared, only the differences in definitions of mandatory levels in the standards on equivalent fields were highlighted. Based on the clarification of these mandatory differences, the search for definitions of minimum mandatory metadata began.

Difficulties were encountered in defining equivalences especially in OpenAIRE Fields related to *Type* subtypes (such as *dateType*), in addition to the adoption of different controlled vocabularies. In particular, OpenAIRE recommends a controlled vocabulary for *resourceType*, named *Controlled Vocabulary for Resource Type Genres (Version 2.0)*, which is a taxonomy for classifying resource genre typologies.

It was decided not to adopt this taxonomy because it was identified that it presents a series of semantic problems in its hierarchical relation, once it is not possible to identify which type of relation is used, for instance, if it is a relation of specialization, such as l *rdf:Type* (or *"is a"*), or if it is another type of relation. Another example is the case When the taxonomy describing an interview is a dataset, which does not seem to make sense once the interview is an intentional action (an event, or *perdurant*), while a dataset is a substance (an *endurant*) that can have different identity principles. This taxonomy also presents reasoning problems in relation to the principles of identity, rigidity and logical disjunction of the categories. For example, the taxonomy presents, in the same level, the *learning object* and *text* elements, which can be (or not) disjoined, and not share the same identity principles.

It is important to point out that adopting a taxonomy of this type can cause a series of difficulties in interoperability, since the machines need a precise description of the types of resources available in data repositories. An incompatibility was also found in the *dcat:distribution* field of the FDP, this field asks for a description of information about the individual archive that composes the dataset (*dcat:Dataset*).

The minimum mandatory central standard developed covers 13 fields with examples of completion and application that can be found in the original document resulting from the landmark in Paganine *et al.* (2020).

Starting from the definition of the core, the design of thematic metadata is approached. A difficulty was the issue of frequent multidisciplinary even in monothematic repositories. For that purpose, the table comparing knowledge areas was used as a guide when choosing and organizing these areas.

It started with the Search for thematic metadata schemes. A page published by DCC in 2020 was used as an initial tool, with different metadata schemes, divided by fields of knowledge. This list was analyzed and deduplicated for the standards of the chosen areas (Social science, Biology and Agronomy). However, the result did not present satisfactory specificity and coverage. Therefore, a list of standards maintained by a group interested in RDA[148]

[148] rd-alliance.github.io/metadata-directory/

metadata was used (RDA, 2020). Finally, the result obtained was refined by checking the metadata obtained with the list displayed in re3data[149] record, measuring which of the selected standards are efficiently used in thematic data repositories related to the selected areas.

Following the level of complexity from the simplest to the most difficult to be treated, the analysis was followed in the tables in the Biology area. Three metadata schemes were analyzed: MIBBI, Darwin Core and ABCD. MIBBI has 40 categories or modules, as they are named, with 23 main fields, but only 17 of them are completely developed, and the remaining were still being designed when the table was collected. Darwin Core has 12 description packets with semantic upload or categories, while ABCD has 38 categories of expandable metadata. While browsing in re3data it is noted that, excluding generic schemes and schemes from different or general knowledge areas, from the 42 results, 2 used MIBBI and only one used Darwin Core.

After the comparative analysis, the metadata of apparent importance in the area were selected, according to their frequency in the standards, eliminating redundancies and generalizing similar terms. The result is also found in the original document, containing 8 fields using Darwin Core, which was chose for the ease of translation and comparison with the traditional Dublin Core:

When focusing on the Social Sciences, which is an area of greater coverage in the chosen standards, a range of standards is also observed. They are: METS, MODS, MARC, CERIF and Dublin Core (DC). METS has 7 classes, MODS, has 20, DC, has 15, MARC, has 9 and CERIF has 22. These classifications and respective Fields or elements of the standards were compared following the same process applied in the Biology standards. The selected result was then adapted to Datacite language, due to its development towards scientific data and the extensive application and compatibility.

Finally, in regard to Agriculture, the same process applied in Biology is used. The schemes AGRIS and AgMES were selected. AgMES has 21 fields based on DC and covers semantic standardizations in agriculture on description, finding resources, interoperability and data exchange in several informational resources. AGRIS has 16 fields and is aimed at the international system of information on agricultural sciences and technologies guidelines, on good practices for information. In re3data record it was not identified the intended use of any of the 2 schemes in thematic repositories of the agriculture area. In order to describe the fields herein, AgMES scheme was chosen due to its proximity with the DC.

Also noteworthy are the negative results obtained in Health, the area presented an unforeseen high complexity of sets description standards. There was a lack of documentation on these, thus, metadata schemes specific to the area were not identified. In the attempt to capture some of the fields frequently used in the area, queries were carried out with thematic data repositories and with some experts from Fiocruz, but the results obtained were still not satisfactory and sufficient for the elaboration of a proposal that encompass the entire Health area.

[149] https://www.re3data.org/

The general multi-thematic metadata standard developed herein has already been applied in the development of Ibict scientific data repositories (named *Aleia*) and CNPq (named *LattesData*).

Aleia and LattesData creations were also motivated by Commitment 3, of the 4th National Action Plan of OGP Brazil, through a technical cooperation agreement (TCA) between CNPq and Ibict, in December 2019. Aleia aims at providing a tool with the functionality of recording, gathering, organizing, disseminating, sharing and preserving scientific data from research carried out by Ibict collaborators and scientific datasets external to the body, but from specific scientific communities. LattesData aims at working as official tool that enables its funded researchers to make inputs in datasets that emerged as a result of projects developed with CNPq resources, being part of the accountability procedures, as well as non-client and partner institutions of CNPq that sign agreements for the collaborative use of the space created.

Both repositories are multi-thematic and intend to cover datasets from researchers from different institutions, from several areas and realities. With this in mind, it was decided to start by just the application of the minimum central description standard with just a few minor changes and additions to better fit its supporting institution.

Aleia additional metadata are presented as follow:

Table 1 – Aleia additional metadata

Field	Description
Author's curriculum in the Lattes platform	Address to access curriculum in the Lattes platform (this field was just adapted from the field "Identifier" of the original standard)
Author's institute of origin	Full name of the institution to which the researcher is linked
Description	General description of dataset and its content
Alternative title	Title of dataset in another alternative language
Contact	Email of the responsible for the dataset.
Dataset language	Language in which the dataset was developed
Reading software and data manipulation	Program used to access and manipulate the dataset archives

Source: Designed by the authors

The additional metadata of LattesData are presented in the following table.

Table 2 – LattesData additional metadata

Field	Description
Contact	Email, preferably institutional, of the responsible for the data
Author's Internal Identifier (IDLattes)	Definition: Identifier number of the curriculum in the Lattes platform (this field was just adapted from the field "Identifier" of the original standard)
Author's external identifier	ORCID identifier number
Author's institution	Name of the institution to which the author is affiliated
Dataset alternative identifier	Another persistent identifier of dataset obtained differently from the main one used in LattesData repository
Notes	Free text that can be used to list/describe archives (related each archive, its type, description and it also informs if it needs a specific software), comments or guidelines for access among other details
Project summary	An explanatory text describing the project and the dataset in a general way, encompassing conclusions, methodology, collection, etc.
Value received	Value, in reais (Brazilian currency) received by the Project from the funding body
Project validity	Start and end date of the project
Materials and other related products	Any product or material related to the dataset Other than scientific publication in formal standard

Source: Designed by authors

Once the additional fields were defined in both repositories, great difficulties were encountered to change them and the form in the chosen software (Harvard Dataverse). As an attempt at a solution, a communication form is being developed with API REST of the Dataverse, for retrieving and filling in metadata externally to the software. Another difficulty encountered is the change of Fields prefixes that comes with the Data Documentation Initiative (DDI) standard instead of the DCAT and Datacite scheme recommended. In the future, it is planned to integrate the filling in of metadata with the development of a Data Management Plan (DMP), so that it also have a machine-readable format.

When analyzing the objectives and results obtained up to now, it is noted that the general core of minimum metadata presents information enough for the interoperability of the chosen guidelines. However, it is interesting to highlight the importance of adding other relevant institutional information to contribute with a more qualified description of the sets deposited, and its association to financed projects, in the case of the CNPq.

Finally, it is also highlighted the magnitude of the complexity of treatment and description of different knowledge areas, in multi-thematic repositories and in special, When considering the possibility of also covering institutions with very diversified realities and contexts.

11.3 FINAL CONSIDERATIONS

The main difficulty faced in the execution of this work was the fact that the repositories do not adopt well-known metadata standards, and the fact that the standards adopted in different areas are not compatible with each other. The content presented as a result from the work carried out in OGP is adequate when dealing with areas that have efforts towards the representation of scientific products, but the description of datasets, which potentially covers any area of knowledge, shows itself to be a work that is constantly changing and updating. Thus, it is noteworthy that this is not a study that ends in its application, on the contrary, it needs constant development in Search of extensions and adaptations to Other areas of knowledge. It is also intended to develop a corpus composed of a set of metadata that describes research data for each specific areas of knowledge. This information will be collected in data repositories of unique themes registered in *re3data* which allow communication. The collection will be automated and will cover: title, keywords, abstract and subject. The Corpus will then be organized and its visualization generated with the VosViewer program for identification of key issues and specific areas highly populated by datasets.

It is noteworthy that the work presented here was prepared in collaboration with IBICT, CNPq and the University of Twente. For future developments, it is intended to continue studies in other areas of knowledge, in special exact sciences and health, as well as in the ontological analysis of genre types of resources to address the problems identified in the classification recommended by OpenAIRE. Future activities also include widespread adoption of FAIR principles and, particularly, the evolution of semantic interoperability among data repositories through application of well-founded ontologies in developing repositories.

REFERENCES

CNPQ. **Tabela de Áreas do Conhecimento**. Brasília: CNPQ, 2020. Available from: http://lattes.cnpq.br/web/dgp/arvore-do-conhecimento. Access on: 28 May 2020.

DCC. **Disciplinary Metadata**. [*S. l.*], 2020. Available from: dcc.ac.uk/resources/metadata-standards. Access on: 28 May 2020.

DFG. **Classification of Subject Area, Review Board, Research Area and Scientific Discipline**. Bonn, 2020. Available from: dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/index.jsp. Access on: 28 May 2020.

FDP. **FAIR Data Point Specification**. [*S. l.*], 2016. Available from: github.com/FAIRDataTeam/FAIRDataPoint/wiki/FAIR-Data-Point-Specification. Access on: 28 May 2020.

GARCIA, Patrícia de Andrade Bueno; SUNYE, Marcos Sfair. O protocolo OAI-PMH para interoperabilidade em Bibliotecas Digitais. Ponta Grossa, p. 1-12, 2003. Available from: https://www.researchgate.net/publication/229039114_O_protocolo_OAI-PMH_para_interoperabilidade_em_Bibliotecas_Digitais. Access on: 28 May 2020.

MOREIRA, João Luiz R.; BONINO, Luiz.; PIRES, Luís F.; SINDEREN, Marten van; HENNING, Patricia. Repositórios para dados localizáveis, acessíveis, interoperáveis e reutilizáveis (FAIR): adaptando um repositório de dados para se comportar como um FAIR Data Point. **Liinc Em Revista**, v. 15, n. 2, 2019. Available from: http://revista.ibict.br/liinc/article/view/4817. Access on: 28 May 2020.

OPENAIRE. **DRAFT:** OpenAIRE Guidelines for Literature Repository Managers v. 4. European Union, 2018. Available from: openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/. Access on: 28 May 2020.

OECD. **Frascati Manual 2015**: Guidelines for Collecting and Reporting Data on Research and Experimental Development, The Measurement of Scientific, Technological and Innovation Activities. Paris: OECD Publishing, 2015. Available from: oecd.org/publications/frascati-manual-2015-9789264239012-en.htm. Access on: 28 May 2020.

PAGANINE, Lucas Nóbrega; CARVALHO SEGUNDO, Washington Luís R. de; MOREIRA, João Luiz R.; SAYÃO, Luís F.; BARRETO NETO, Vanderlino Coelho; CIUFFO, Leandro Neumann; FELICÍSSIMO, Carolina Howard; DIAS, Gustavo Neves; **Padrões de interoperabilidade para repositórios de dados de pesquisa**. Brasília, DF: IBICT , RNP; Rio de Janeiro: CNEN; Enschede: University of Twente, 2020. 44 p. Available from: https://www.gov.br/cnen/pt-br/acesso-rapido/centro-de-informacoes-nucleares/material-didatico-1/padroes-de-interoperabilidade-para-repositorios-de-dados-de-pesquisa.pdf. Access on: 28 May 2020.

RDA Metadata Standards Directory. 2020. Available from: https://rd-alliance.github.io/metadata-directory/. Access on: 28 May 2020.

RE3DATA. **Registry of research data repositories**. [*S. l.*], 2020. Available from: service.re3data.org/about. Access on: 28 May 2020.

RNP. Wiki **Ciência Aberta na OGP Brasil**. 2018. Available from: wiki.rnp.br/pages/viewpage.action?pageId=107315238. Access on: 30 Apr. 2020.

SANTOS, L. FAIR Data Points Supporting Big Data Interoperability. *In*: MERTINS, Kai; JARDIM-GONÇALVES, Ricardo; POPPLEWELL, Keith; MENDONÇA, João P.(org.). **Enterprise Interoperability in the Digitized and Networked Factory of the Future**. London: ISTE Press, 2016. p. 28

WILKINSON. Mark D.; DUMONTIER, Michel; AALBERSBERG, IJsbrand Jan; APPLETON, Gabrielle; AXTON Myles; BAAK, Arie; BLOMBERG, Niklas; BOITEN, Jan-Willem; SANTOS, Luiz Bonino da Silva; BOURNE, Philip E.; BOUWMAN, Jildau; BROOKES, Anthony J.; CLARK, Tim; CROSAS, Mercè; DILLO, Ingrid; DUMON, Olivier; EDMUNDS, Scott; EVELO, Chris T.; FINKERS, Richard; GONZALEZ-BELTRAN, Alejandra; GRAY, Alasdair J. G.; GROTH, Paul; GOBLE, Carole; GRETHE, Jeffrey S.; HERINGA, Jaap; HOEN, Peter A. C 't; HOOFT, Rob; KUHN, Tobias; KOK, Ruben;

KOK, Joost; LUSHER, Scott J.; MARTONE, Maryann E.; MONS, Albert; PACKER, Abel L.; PERSSON, Bengt; ROCCA--SERRA, Philippe; ROOS, Marco; SCHAIK, Rene van; SANSONE, Susanna-Assunta; SCHULTES, Erik; SENGSTAG, Thierry; SLATER, Ted; STRAWN, George; SWERTZ, Morris A.; THOMPSON, Mark; LEI, Johan van der; MULLIGEN, Erik van; VELTEROP, Jan; WAAGMEESTER, Andra; WITTENBURG, Peter; WOLSTENCROFT, Katherine; ZHAO, Jun; MONS, Barend. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, England, v. 3, 2016. Available from: https://www.nature.com/articles/sdata201618. Access on: 28 May 2020.

How to cite this chapter: PAGANINE, Lucas N.; CARVALHO SEGUNDO, Washington L. Ribeiro de; MOREIRA, João L. R. Development and application of interoperability standards for scientific data repositories: Ibict and CNPQ repositories. *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. 11, p. 133-144. DOI: 10.22477/9788570131959.cap11.

12. *INVESTIGATING FAIR PRINCIPLES IN NATIONAL INSTITUTES OF HEALTH (NIH) SCIENTIFIC DATA REPOSITORIES*

Marcello Peixoto Bax[150]

12.1 INTRODUCTION

Data collection and analysis are essential for all the sciences, but They are especially important when it comes to biomedical or health sciences. As the world advances towards the personalized medicine and more speed and agility in the production of drugs, the understanding of data collected in clinical trials and other studies is important to speed up the advance of scientific research. Unfortunately, due to the multiplication of data and formats, the collection, organization, and dissemination of data is becoming increasingly difficult to be efficiently executed. In addition, understanding phenomena and proving a hypothesis in this field require large amounts of data, and few researchers have the resources and means to collect such an amount of information. Clinical trials are expensive, require non-trivial resources and can take years to complete, depending on the study. Obviously, as such, this type of data collection and acquisition is not readily available for most researchers. That is why the health field, and other knowledge domains, are moving towards the widespread sharing of these data through public or private *data centers* available to researchers.

Unfortunately, as a result of inadequate data management, successful data sharing initiatives are very few. Data management is "the main channel for knowledge discover and innovation", promoting data sharing and reuse in scientific communities (Wilkinson *et al.*, 2016). It became important to define a set of common principles that define what a "good" data management should be. These principles, which highlight the findability, accessibility, interoperability, and reuse capacity of datasets, are known as FAIR Guiding Principles. Such principles are increasingly considered a reference for *data centers* and are being used to evaluate and highlight the success of certain initiatives. Numerous publications discuss the adherence to FAIR principles as a way to illustrate the commitment to facilitate data sharing in their respective communities. Examples include *Immune Epitope Database* (Vita *et al.*, 2018), *DisGeNET Platform* (Piñero *et al.,* 2017), *BioSharing Portal* (Mcquilton *et al.,* 2016) and *Omics Discovery Index* (Perez-Riverol *et al.,* 2017).

There are several common problems that prevent data to being considered FAIR. First, few *datasets* can conform to each other; several are closely related, but data are not set the same way; therefore, they do not easily conform and cannot be integrated for analysis. Data harmonization requires the use of common or at least commensurable categories and units of measurement. Second, in the case of research involving different teams, the research coordinator (main investigator) generally knows more details on the structural nature or data collected (their

[150] Post-doctorate in Information Science, School of Information Science SIC – UFMG, bax@ufmg.br

properties and relations), than he can convey cohesively as supplementary information that would accompany the data itself (metadata). Finally, if the amount of data is sufficiently voluminous, automated methods may be the only viable way to generate comprehensive and in-depth analysis of it. However, if data meanings are not explicitly formalized in such a way to be "machine-readable", there will be no automated method that can support this analysis. This is where the concept of semantic "lifting" of data or even of semantic "ingestion" of data in repositories comes in.

12.2 SEMANTIC *LIFTING* OF DATA

The semantic lifting of data is a process by which data are converted from their original tabular representation to CSVs files and/or relational tables, for an ontological formal representation that represents the "knowledge" in the structure of a graph of knowledge (Pan *et al.,* 2017). In this operation, data are not only converted to another format, but "elevated" to the level of "knowledge" as they are represented by ontological models based on description logics that explain their formal semantic. The process transforms data, originally with no explicit meaning, into data potentially interoperable in the semantic web (linked data) and treatable by computer. Data lifting is, therefore, important because it helps to fight all the problems aforementioned that are targets for the treatment of FAIR principles. Data are collected and re-structured in accessible format, guided by metadata and machine-readable. Data in this format can be widely released by the web for subsequent extraction of information and knowledge, preserving its original meaning.

Several data centers are working to increase the *FAIRness* of their data repositories. In some cases, it is done with the development or integration of software platforms that incorporate a process of semantic enrichment of data models. Something that can be achieved by representing the model, or part of it, with a formalism guided by ontologies (semantic lifting) and mapping it to a reference ontology.   As it is a relatively recent phenomena, there is not a consensual method of performing the process of *lifting* and data intake. Therefore, it is important to understand how different organizations are trying to improve the state of the art in terms of "semantic data lifting" as a way for us to learn from each of these efforts.

12.3 DATA CENTERS FUNDED BY *NATIONAL INSTITUTES OF HEALTH*

The National Institutes of Health (NIH) funds hundreds of data centers in different health areas. Some of them have unique data sets for a specific domain, while others host several data sets in various NIH domains and agencies. Although the Institute encourages data centers to use specific domain repositories whenever possible, these repositories are not available for all data sets. When researchers cannot find a data center that maintains a repository for its subject or for data they generate, a general repository may be a useful site to share data. General repositories accept data regardless of type, content or disciplinary focus. NIH does not recommend a specific general repository, but it maintains a non-exhaustive list, provided as a guide to finding repositories. The list contains the following most known general repositories: *Dataverse, Dríade, Figshare, Mendeley Data, Open Science Framework, Vivli and Zenodo*. A comprehensive list of data centers funded by NIH for sharing data was created by *the US National Library of Medicine*, where *Trans-NIH Biomedical informatics coordinating committee* (BMIC)[151] keeps another list with currently 66 data centers (by October 2017) that maintain domain-specific and open repositories funded by NIH. Another 31 domain specific supported repositories include those that have limitations in sending and/or accessing data (sensible data).

Studying all the 97 repositories from these data centers would not be possible. Understanding the technical resources of a data repository is not trivial and generally requires accessing at least some available data. Thus, initially, **about 10 repositories were studied**, whose descriptions stood out for their greater level of detail. These repositories were inspected regarding the technical capacities that differentiate them from the others. The research revealed that, actually, some of these repositories are hosted in software platforms developed by third-parties that contain data from various studies and different institutions. It generates an interesting dynamic in which some data centers create repositories and host their data, while others simply host data for institutions that are interested. Finally, three data centers were selected for a more detailed inspection of their repositories: **ImmPort**, **Synapse** and **NDA** (*National Data Archive*) from the National Institute of Mental Health (NIMH). These three centers were selected due to the possibility of data access. Each one of the centers contains at least some repositories that allow the public access to summarized data, at the very least. Having access to data allowed a deeper understanding of how they is stored and how they can be searched. Another reason they were selected was due to their usability. These platforms had data search mechanisms relatively simple. It should be noted, however, that many other sophisticated data centers exist in many countries, including Brazil, and that this analysis did not intend to exclude their relevant contributions. However, the time and resources restrictions of this research demanded that some platforms easier to access were considered.

12.4 EVALUATION CRITERIA

The analysis of these data centers pointed out four evaluation criteria: 1) how can data be found and searched/consulted? 2) Are they single-domain or multiple cross-domain data? 3) Is the data representation scheme-free or fixed/relational? 4) Does the repository semantically lift the data when ingesting it into a database? These criteria were selected because each one of them works as an indicator of the level of data adherence to FAIR principles. The first criterion aforementioned looks to meet certain findability and accessibility characteristics because the filtration/query resources demand that data are "described with rich metadata", "recorded or indexed in a searchable form" and "retrievable by their identifier using a communication protocol standard" (Wilkinson *et al.*, 2016). The domain or knowledge area of a repository, according to the criterion aforementioned, works as an indication of potential for reuse of data. Although FAIR principles indicate that reusable data "meet community standards relevant to the domain" (Wilkinson *et al.*, 2016), the platforms that can meet the researchers' needs in correlation to domains have the potential to facilitate the research. Data must "use a formal, accessible, shared and widely applicable language for knowledge representation" so that they can be considered interoperable (Wilkinson *et al.*, 2016). Thus, the repositories that have a specific scheme (fixed/relational) use it as a way to facilitate this interoperability; however, fixed schemes can limit researchers in their decisions on which data they can or cannot

[151] https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

store. Finally, machine readability of data was emphasized by proponents of FAIR principles. As such, what we are calling "data lifting" provides inherent mechanisms to meet each one of these principles.

The following sections detail the specific characteristics of each one of the repositories of three data centers researched, regarding the four evaluation criteria aforementioned. The degree of compliance with the criteria can be considered a proxy to understand how FAIR principles are more widely considered by data centers.

12.5 ANALYSIS OF THE STUDIED REPOSITORIES

According to Byrd *et al.* (2020), whenever possible, scientific research data must be shared through domain-specific repositories, which use data widely used in a field. Such specific repositories are ideal data warehouses. They provide long-term access to data through the provision of persistent IDs, such as digital objects identifiers (DOI). They reduce research costs by making large collections of correlated data available in a single location, which can reduce redundant work and encourage the generation of new hypothesis from secondary analysis. Lastly, they allow data to be cited, making data scientists earn credit for sharing datasets. We analyzed two domain-specific repositories below.

12.5.1 Domain-specific repositories

12.5.1.1 *ImmPort*

ImmPort[152] is funded by NIH, focusing on "Bioinformatics for the future of Immunology". It is a "curation and distribution portal" that aims at providing immunological data sharing (Bhattacharya *et al*., 2018); it is "one of the biggest open repositories with curation" of human immunological data (Sansone; Cruse; Thorley, 2018). In its efforts on data curation, ImmPort elaborates guidelines and standards based on suggestions from the immunology research Community, maximizing data accessibility and interoperability of this community. The repository is composed of four components: **private data, shared data, data analysis** and **resources**. Data collected is selected in the private data component, eventually released through the shared data component. The data analysis component uses *Galaxy* tool to allow data analysis in the repository space itself. Galaxy makes the analysis and meta-analysis of cytometry data easier, which is the focus of the portal. Finally, in **resources,** information on ImmPort is gathered, its publications and tutorials.

Unlike other repositories, ImmPort uses ontologies as a way to annotate its data with common and agreed terms, including one Cellular ontology, one Disease ontology, one ontology for Biomedical Investigations, one ontology for Proteins and one for Vaccines. These ontologies were used in the elaboration of *ImmPort Data Model*, which details the variables stored in each table and the relation among them. When uploading data to the ImmPort, the data model provides a set of common terms to be used so that the annotation is consistent with the other data already in the repository. It is done through data upload models and a validation tool. The studies available

[152] https://www.immport.org/

in the repository can be consulted through a basic Keywords Search or by applying filters that include metadata such as if the study were or wasn't a clinical trial, the type of study, the species researched, the type of biological sample and the type of clinical trial.

These search resources do not consider data in itself, only certain metadata provided currently the study is submitted to the repository, which is done through pre-defined templates. Thus, there is no way to consult data by filtering certain criteria in various studies simultaneously. In addition, to visualize data in itself, individual archives must be downloaded by the researcher. However, detailed metadata on stored studies is available directly on the website. Metadata is standardized through templates designed based on the data models.

ImmPort domain is strictly of immunology. It is noted that the platform was adapted specifically to receive research on immunology. The data model itself also has elements very specific to immunology. Although this rigidity is important when data adjust to the model, it limits the use of ImmPort by other research with domains that could cross with immunology. ImmPort also has a specific scheme through the data model and, therefore, it is clearly not schema-free. This limits researchers when they need to store data that does not exactly fit the predefined schema/model.

ImmPort has some degree of data lifting, although it is not exactly clear how meaningful and comprehensive it is. The repository provides to the researchers models to be used when formatting and sending its data and asks the researchers to validate these data in relation to existing models. This indicates that ImmPort aims at standardizing its data so that studies are compatible among them. However, as studies are downloaded file by file, it is not clear whether the portal uses these models to store them in a way that they can be combined, generating information and knowledge. In general, ImmPort shows certain interesting resources, but it's not clear how deeply it applies the lifting process to store data.

Considering specific domain data centers funded by NIH, in addition to ImmPort, *National Data Archive* is an infrastructure to host data repositories in the mental health domain.

12.5.1.2 *NIMH National Data Archive (NDA)*

Initially developed to integrate a set of research data repositories as the *National Database for Autism Research* (NDAR[153]) and Other three in mental health, "it became a platform to share data on mental health and other researchers. The platform has strict restrictions on data use, and the download requires the user to complete a Certification of Use signed by NIH. Although it limits the use of the platform, summary of data is available and can be consulted. NDA has branched out to include other aspects of mental health domain. The repositories included, in addition to NDAR, are *Research Domain Criteria Database* (RDoCdb), *National Database for Clinical Trials related to Mental Health* (NDCT) and NIH *MRI Repository* (PedsMRI). NDA is structured to meet the needs of specific research data on mental health. In addition, the restriction of its access makes it accessible predominantly to participants of communities in the mental health area.

[153] https://nda.nih.gov/

NDA content is organized around the concept of "Globally Unique Identifier" (GUID), which works as a way to identify data of unique individuals (Dan *et al.,* 2018). GUIDs are generated by a tool that requires the researcher to enter specific personal identification information, which is then used to generate a hash code that solely represents the person in the data set. The same personal identification information will ensure that the same GUID is generated, therefore, if the same subject participates in several studies, it will not be duplicated in the system. It allows that all data is internally related to only one person, enabling the NDA to provide sophisticated queries for data extraction.

NDA has six query tools: general queries, data from labs, data from papers, data dictionary, query per concept and query per GUID. Each tool provides its own exclusive resources, which improves the process of data collection and analysis. The general query allows the researcher to select predefined fields to build a query. The results from this query are shown (along with the summary statistics) and the resulting data can be downloaded. In addition, the user can select which exact fields he wants to download, as well as from which source. This is unique because it means that a single query can generate results in all repositories using NDA to store their data (although certification of data use is necessary to download data from each repository). *Data from Labs* and *Data from Papers* tools look into information on NDA collections and NDA studies, respectively. Here, collections and studies can be filtered by different criteria and downloaded using the same download mechanism used by the General Query tool. It is crucial because it allows the researcher to select several collections or studies and extract specific structures, in accordance to the definitions in the Data Dictionary, only for download. The *Data Dictionary* tool allows the researcher to select "data structures" and e "data elements" directly from the data dictionary. This dictionary shows several attributes of each data structure and includes detailed information on its elements. Finally, the *query by Concept* tool allows query through "ontological concepts", according to definition by ASD *Phenotype Ontology*, using the same filtering resources and download available in the platform as a whole. It is important to note, however, that data is not stored using any type of ontological representation, the ontology is used as a filtering tool. In fact, the NDA approach "does not allow easy creation of an ontology, whether it among all data in clinical evaluations in the NDAR or among data in the NDAR and other lexicon" (Dan *et al.,* 2018).

Just as the ImmPort, the NDA uses a very specific scheme, according to the definition in its data dictionary. When a user submits any data set, it has to be validated according to the data dictionary; otherwise, it is not accepted in the system. This validation tool is publicly available for researchers and will warn the researcher about his errors that can be fixed in his data. In addition, all data sets must have a GUID, which restricts them to be related to a single subject (it makes sense for clinical data and mental health, but makes extensibility low among domains).

If the researcher needs a structure not defined by the data dictionary, he can send new definitions to the NDA *Help Desk* for eventual implementation. This means that even if the platform has a very specific scheme, such scheme is in a certain way open to changes and additions. However, this makes changes in the scheme take longer to be implemented because all maintenance if performed manually by NDA employees. This is also applicable to data upload, which generally takes 4 months to be publicly available on the platform. Up to this point, data remain in a private status so that NDA employees can review and ensure its quality.

Due to the rigid scheme and NDA validation tools, the data intake process can be done quickly. The query tolls available suggest that data stored in the NDA are transformed from their original upload status to a format in which all data are stored around GUID. It allows the extraction of knowledge among studies, collections, and repositories in a way many other platforms cannot.

The capacity of selectively manipulate data for download creates many opportunities for exclusive analysis of data. Our investigation was unable to exactly clarify how these data are stored in the "backstage", but it was clear that all data are associated to a single GUID in all the platforms in a way that it can be easily looked into. This makes NDA different from many other repositories; however, there is still room for improvement when it comes to upload automation and data curation.

12.5.2 Archive repositories for general use

Both data centers examined hereto are domain-specific, however, in certain circumstances, particularly at the beginning of the development of a scientific data domain, they may not have specific repositories. In such cases, investigators can still choose to put data in archiving platforms for general use, such as Figshare or Zenodo, along with metadata that precisely describe the archives included and its format. For data that cannot be publicly shared due to privacy issues, the Synapse platform provides a similar archiving platform for general use that supports controlled access sharing (Byrd *et al.* 2020).

12.5.2.1 *SYNAPSE Platform*

Synapse[154] is an open-source software platform for researchers who can use it as a site to store and annotate their data. Different from ImmPort, it does not have requirements for data formatting, serving researchers who just want to store their data somewhere. Even so, many data centers funded by NIH use it as a repository, and the platform itself is funded by several institutes linked to NIH.

The platform allows its users to create personal workspaces, upload different archives, connect themselves through provenience relations, annotate archives for better finding, provide narrative for data, create digital object identifiers (DOI) for any resource and work collaboratively. To register as a Synapse, the user simply has to provide an email, and the download of public data and the creation of content become easily accessible. Note that to store data on human beings (once there are use restrictions for that), the researcher must go through a certification process.

Synapse can be operated through several methods, including Python and R, in addition to the traditional web interface. However, certain functionalities (as downloading a group of files) are available only via Python or through the command line. Each resource in the platform has an exclusive SynapseID and, therefore, can be retrieved. The user must use the web interface to determine the resource SynapseID, but once found, automated tools can conduct data analysis.

[154] https://www.synapse.org/

Each project in Synapse stores its information in relational tables whose schemes are defined by the project owner. This makes a Project queryable using similar SQL language, but in a standard research/filter interface also available. In general, these queries are used to research specific files or multiple files that share a specific characteristic; however, the user must know the scheme to generate a successful query. Synapse supports two different structures: table views and file views. File views allow browsing the uploaded files, view and download them. However, queries cannot be executed in data themselves. A data table can be researched and queried. Queries can be extended by different repositories, as each resource hosted in Synapse has an exclusive ID; however, as table schemes are identified by the Project owner, there is no guarantee that a single query can be successfully extended by several repositories that use different schemes.

As Synapse only operates as a domain independent platform, there is no specific domain connecting all repositories. Any certified user can upload and store data in the platform, a researcher does not need to operate in any specific domain to have the site benefits. Synapse is schema-free in the sense that the user is responsible for deciding which scheme to use to store data. However, each resource in a project must follow a pre-define standard scheme so that the project is uploaded and queryable. If the scheme is not enough for a researcher, he can create his project with his scheme, but it results in separation of an existing repository from which he could be benefited.

Relational tables are primitive when it comes to data lifting, for their rigid structures limit the information that can be extracted. As such, Synapse is limited when it comes to its data lifting resources. However, Synapse has an interesting capacity that allows users to "track the history of analysis and communicate and share a sequence of processing stages". The user himself must define the provenance (preferably When loads or edits an archive). Otherwise, the provenance track is lost. In general, the opening of Synapse and its low entrance barrier make the platform widely accessible, but this freedom is at the cost of making standardized approaches of data exchange among repositories impossible.

12.6 FINAL CONSIDERATIONS

Obviously, data organization and availability for research in health are highly related to the amount of information and knowledge that can be extracted from them. That is why research groups are starting to develop and apply tools to better structure these data so They can be analyzed and shared more promptly. Often, these groups appeal to FAIR guiding principles as a reference to develop their functional resources of data sharing. Making data available in this way, following such principles, lowers the barriers inherent in successfully carrying out research in health, allowing researchers to use and apply it in their research, given that many times data is not collected by them. It also opens doors for cross-domain collaborative studies, a trend that has been steadily increasing recently.

From the current 99 repositories funded by NIH, some have resources that illustrate the movement towards a wider sharing of scientific data by researchers, especially when it comes to data lifting, which is an important foundation, so data can be considered FAIR. All three repositories analyzed present certain attributes that show their movement towards a deeper use of data lifting in their infrastructure, although they are still very limited in the wider adoption of this criterium.

As summarized in Table 1, these repositories were evaluated through four criteria – query structure, scientific domain of action, data representation scheme and data lifting – as a way to understand the level of sophistication towards meeting FAIR data and machine-readable criteria. Although each one of the analyzed repositories has its strengths and weaknesses, it is important to understand what these organizations are doing to improve their resources aimed at sharing data for future research, as a way to understand the information and knowledge in the health research as a whole.

Table 1 – Analytical-comparative synthesis of repositories evaluated

Repositories/Criteria	ImmPort	NDA	Synapse
Query structure	Through metadata terms and filters	Through data from labs, data from papers, data dictionaries and concepts and GUID	Through table and archive
Scientific domain	Immunology	Mental health	any
Data scheme	Repository-specific	Repository-specific	*Project-specific*
Data lifting	Limited	limited	limited

Source: the author.

REFERENCES

BHATTACHARYA, Sanchita *et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. ***Sci Data*****,** v. 5, n. 180015, 2018, p. 1-9. DOI: https://doi.org/10.1038/sdata.2018.15

BYRD, James Brian *et al*. Responsible, practical genomic data sharing that accelerates research. **Nature Reviews Genetics**, v. 21, 2020, p. 615-629. DOI: https://doi.org/10.1038/s41576-020-0257-5

DAN, Hall *et al.* Sharing Heterogeneous Data: The National Database for Autism Research. **Neuroinformatics, v.** 10, n. 4, oct., 2012. p.331–339. DOI: 10.1007/s12021-012-9151-4

MCQUILTON, Peter *et al.* BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. **Database (Oxford)**, v. 2016, p. baw075, jan., 2016. DOI: 10.1093/database/baw075

PAN, Jeff Z. *et al.* Exploiting Linked Data and Knowledge Graphs in Large Organizations. Switzerland: Springer International Publishing, 2017. 266 p. DOI: https://doi.org/10.1007/978-3-319-45654-6

PEREZ-RIVEROL, Yasset *et al.* Discovering and linking public omics data sets using the Omics Discovery Index. **Nature biotechnology,** v. 35, n. 5, p. 406-409, may., 2017. DOI: https://doi.org/10.1038/nbt.3790

PIÑERO, Janet *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, **Nucleic Acids Research**, v. 45, n. D1, p. D833-D839, jan. 2017. DOI: 10.1093/nar/gkw943.

SANSONE, Susanna-Assunta; CRUSE, Patricia; THORLEY, Mark. High-quality science requires high-quality open data infrastructure. **Scientific Data,** v. 5, n. 180027, feb., 2018 . DOI: https://doi.org/10.1038/sdata.2018.27

VITA, Randi *et al.* FAIR principles and the IEDB: short-term improvements and a long-term vision of OBO-foundry mediated machine-actionable interoperability. **Database (Oxford)**, v. 2018, p. bax105, 1 jan. 2018. DOI: 10.1093/database/bax105

WILKINSON, Mark D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v. 3, n. 160018, 2016. p. 1-9. DOI: https://doi.org/10.1038/sdata.2016.18

How to cite this chapter: BAX, Marcello Peixoto. Investigating FAIR principles in national institutes of health (NIH) scientific data repositories. *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. 12, p. 145-154. DOI: 10.22477/9788570131959.cap12.

13. DATA REUSE: FAIR PRINCIPLES AND THE RESEARCH ECOSYSTEM

Sônia Elisa Caregnato[155]
Rafael Port da Rocha[156]
Rene Faustino Gabriel Junior[157]

13.1 INTRODUCTION

The Open Science movement has gained momentum recently due to both technological advancement and society's perception that scientific research is a collective activity, publicly funded, and that needs to return value to the society that supports it.

In this matter, data produced in the course of a research, in addition to publications that contextualized them, must be in open access so they can be shared among scientists and reused in new research, providing feedback to science, whose character is cumulative. Therefore, data sharing has been a demand of governments, funding agencies and research institutions, but in order for this to materialize, planning, management and curation of data sets in repositories are necessary. Such activities occur in the scope of a research ecosystem that involves technologies, people, and institutions.

With the aim of ensuring good practices for research data sharing, FAIR principles establish that data should be findable, accessible, interoperable, and reusable. The intention is, through the principles, to facilitate data reuse both by humans and machines. (Wilkson *et al*., 2016).

Therefore, sharing and reuse represent a pair of concepts that complement each other. However, as observed by some authors (Pasquetto; Randles; Borgman, 2017; Tenopir *et al*., 2011; Wallis *et al*., 2013), the first is much more frequently studied than the second, although the benefits of sharing can only be obtained if data are effectively reused.

Searching for a deeper understanding of this subject and its implications, this paper explores the use of the term research data *reuse*, as well as terms commonly related to it, through a literature review. Initially, research data sharing is approached, then it is related to reuse and FAIR principles. Afterward, the meaning of the terms

[155] PhD by the University of Sheffield, United Kingdom. Professor in Graduate Programs in Communication and Information Science, both at the Federal University of Rio Grande do Sul (UFRGS). Email: sonia.caregnato@ufrgs.br

[156] PhD in Computer Science by the Federal University of Rio Grande do Sul (UFRGS). Professor in Graduate Program in Information Science UFRGS. Email: rafael.rocha@ufrgs.br

[157] PhD in Computer Science by Paulista Júlio de Mesquita Filho State University

(Unesp). Professor in the Graduate Program in Information Science at the Federal University of Rio Grande do Sul (UFRGS). Email: rene.gabriel@ufrgs.br *use* and *reuse* is discussed, as well as their variations to, finally, address the necessary conditions for an effective reuse of research data.

13.2 RESEARCH DATA SHARING

Research data sharing is intrinsically related to its use, whether to validate it or originate new interpretations. However, the sharing act informs very little about the use that will be made of data. There is, indeed, an assumption in the open access policies, that is, that research data are useful for other researchers and that they are going to reuse them (Pasquetto *et al*., 2019).

According to Borgman (2012), sharing concerns the act of opening data in a way they can be reused by other individuals. It is important to point out that the degree of confidence, the use, and the value of these data vary dramatically: while some are structured and curated, others are simply made available.

In part, at least, the quality of data made available depends on how it is shared. Pasquetto, Randles e Borgman (2017) specify that sharing can be done through private exchanges between researchers, deposit in repositories, availability in lab websites, supplement of journal articles and, more recently, data articles, which clarify the provenance and enable the allocation of credit to authors. Thus, for the authors, both direct exchanges between researchers and the open availability of data are understood as sharing. The way it is done, according to the authors, varies in relation to the knowledge area, data type, country, funding agency and other factors.

Boté and Térmens (2019), on the other hand, prefer to differentiate private sharing from public sharing in general or specialized institutional repositories, calling the latter publication. For them, sharing among peers during the period when data is not necessarily publicly available assumes a high level of trust between the parties, as well as an easier way to obtain information through informal channels on how to integrate data in new projects. Data release in repositories, however, demands a greater effort given the need to provide detailed documentation that accompanies data and explain how to use them.

Regardless of the way, there are, evidently, advantages to data sharing. Borgman (2012) identified four rationales for this practice. They are: a) to reproduce or to verify results; b) to make available the results of the research funded by the public sector; c) to enable others to formulate new research questions based on existing data; d) to advance research and innovation.

However, there are also barriers and challenges to data sharing. So much that, as it is an intricate and complex phenomenon, Professor Christine Borgman called it a conundrum in her famous 2012 article. According to her,

> If the rewards of the data deluge are to be reaped, then researchers who produce those data must share them, and do so in such a way that the data is interpretable and reusable by others. Underlying this simple statement are thick layers of complexity about the nature of data, research, innovation, and scholarship, incentives and rewards, economics and intellectual property, and public policy (Borgman, 2012, p. 1.059).

Among the difficulties for not sharing research data are fears of inappropriate use of data and competition, the cost of preparing data and documentation, lack of time, lack of appropriate infrastructure and standards, ethical issues – among them, the use of data for purposes apart from those for which they were collected – and, finally, the fact that raw data are of little use for reuse without significant efforts to make them available in a way that allows further analysis (Borgman, 2012; Kim; Yonn, 2017; Perrier *et al*., 2020; Rowley *et al*., 2017; Tenopir *et al*., 2011; Wallis; Rolando; Borgman, 2013).

The support for research data sharing and the struggle to overcome obstacles to data reuse motivate the search for measures to guide the necessary management actions. FAIR principles are an important milestone in this effort and also in expanding the value of data for its reuse, as will be discussed below.

13.3 FAIR PRINCIPLES AND THE RESEARCH ECOSYSTEM

In addition to being one of the four FAIR principles, reuse is also the purpose of data curation processes. The main feature of FAIR principles is the provision of a concise, high-level set of guidelines that are valid for any domain and that must be applied not only to data, but also to metadata. Accordingly, they present themselves as a means of facilitating reuse.

In a final report entitled Turning FAIR into reality (Collins *et al*., 2018), The European Commission's FAIR Data Specialist Group pointed out that the implementation of principles requires the creation of a new research culture, in addition to a technical ecosystem consisting of appropriate services and infrastructure, which include policies, data management plans, persistent identifiers, interoperability standards, metadata, and repositories. In addition, the authors stress that it is necessary to promote the development of skills to, on the one hand, process and analyze data (data science), and on the other hand, manage and preserve them throughout their life cycle (data curation). It is also necessary to develop indicators for evaluating compliance with the Principles and, ultimately, pursue project sustainability and funding.

The report defines the FAIR ecosystem as a model that indicates the minimum components necessary to promote the creation, curation, and reuse of FAIR digital objects in an effective and sustainable way (Collins *et al*., 2018). The central element of this ecosystem, therefore, is the digital object (research data and other resources), which must be accompanied of persistent identifiers and metadata to enable it to be found, used and cited. FAIR digital objects also need to be represented, ideally in open file formats and use vocabulary common to communities, so that interoperability and reuse are possible. In addition, the associated documentation must include machine-actionable instructions on conditions of use and licensing. Finally, as pointed out by Koers *et al*. (2020), this ecosystem is supported by metrics, certification mechanisms, incentives, funding and training.

An important movement for FAIR principles to become effective is the GO FAIR initiative, which emerged in Europe in 2017, and expanded to other countries including Brazil. GO FAIR proposes the formation of networks for implementing FAIR data and services, so that those interested can work in participative and collaborative ways (Sales *et al*., 2020).

Neither FAIR principles nor GO FAIR initiatives establish the obligation for all research data to be open, that is, data can be FAIR and, at the same time, be shared in a restricted way. This provision is necessary in certain circumstances, for example when including personal, confidential or commercially valuable information. The greatest benefits to science and society, however, occur when data are both FAIR and open, as the absence of restrictions increases the possibilities of large-scale reuse (Collins *et al*., 2018), or, as claimed by Henning *et al.* (2019, p. 394),

> [...] the more open they are, the more they will be used, reused and combined with other data, promoting economic growth, innovation, and development... Information on use licenses, however, must be clearly specified for the data to be considered FAIR. Thus, if the data cannot be opened, or if they can only be used with restrictions, this information must be made explicit.

The FAIR principles establish that (Wilkinson *et al*., 2016):

> 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 detailed provenance
> R1.3. (meta)data meet domain-relevant community standards

That is, to be reused, both data and metadata must be accompanied of information that effectively enables them to be used in different contexts from those to which they were created.

In this matter, Turning FAIR into reality report (Collins *et al*., 2018) suggests 27 recommendations, each one accompanied by a set of relevant actions set to support the realization of FAIR data ecosystem. Among these, there are some directly related to the reuse. Particularly, the authors recommend that research funders should encourage FAIR data reuse, requiring that communities approach existing content whenever possible. This can be done by requiring researchers to show in their projects that FAIR data were searched and/or looked into before proposing the creation of new data, or when acknowledging that the results from research that reused data has the same value of research that created new content, or when financing research that reuse FAIR data.

For reuse to effectively derive from sharing, it is necessary to understand all dimensions of the phenomenon, starting with the definition of the term itself. Thus, the next section deals specifically with the reuse of data, considering its dimensions and characteristics.

13.4 RESEARCH DATA REUSE

As previously mentioned, research data sharing implies its reuse for of science itself, the scientific community and society in general, that is, from the perspective of the research ecosystem. In this regard, it is initially necessary to clarify the meaning of reuse in this specific context and in its relation to the use of research data.

According to Van de Sandt and colleagues (2019), the term *research data reuse* refers to a complex concept that varies according to the knowledge areas. Even so, the authors stress that it is essential to define it because it has been increasingly used by funding and research institutions, showing its importance.

The first distinction frequently found is that between use and reuse, as suggested by Pasquetto, Borgman, Wofford and Pasquetto (2017), Randles and Borgman (2019). Other related concepts are also pointed out in the literature, for example, reproducibility, replicability, integration, and reanalysis (Boté; Térmens, 2019; Curty, 2019;

Van de Sandt, 2019). These aspects, including taxonomy proposals or models to understand data types, will be discussed below.

13.4.1 Defining reuse

Reuse is often defined as the subsequent use, made by other researchers, of data collected for a certain project, or as precisely stated by Boté and Térmens (2019, p. 329), "[...] finding, processing and analyzing someone else's datasets to create new knowledge".

Three elements are essential in this definition: a) it is a secondary use, not originally intended; b) it happens temporally after use; and c) it is done by a researcher or research group different from the one who collected data. Regarding the first and second aspects, there appears to be no divergence in the literature. In relation to the third, however, there are divisions. Pasquetto, Borgman and Wofford (2017) explicitly point out that, if the same scientist returns to the same dataset in a later project, this action would be characterized as use and not reuse. For the authors, reuse occurs when datasets are retrieved by third parties and used in another project. Some authors, however, do not establish this distinction, for example, Custers and Uršič (2016), while others still defend that any subsequent use, even if it is by whom collected the data, must be considered reuse (Curty, 2019).

A distinctive approach is that taken by Van de Sandt and colleagues (2019), who concluded that the discourse characteristics of the subject area do not prove that there is any difference between reuse and use. Based on the etymological analysis of the words *use* and *reuse* and on related concepts, as well as on discourse analysis and the formulation of scenarios, the authors consider four characteristics frequently used to differentiate use from reuse

> a) the character of the data, which refers to the number of reused datasets or their transformation by reuse;
> b) the user, who the literature differentiates from the data producer;
> c) the purpose, which is related to the research question and/or method;
> d) the temporal dimension, in which the original use is evidenced before the second use of the data.

Based on these analyses, authors state: "Therefore, we define (re)use as the use of any research resource, regardless of when it is used, its purpose, its characteristics and its user" (Van de Sandt *et al*., 2019, p. 14). Furthermore, they seek to credit the confusion of terms to the linear model centered on the published article, which would be less dynamic and complex than the current research scenario.

The originality of the work is also in relating this proposal of simplifying the language to the involvement in the open science movement, since, according to the authors, researchers would be more likely to publish and document quality research for a purpose (the use) that is already consolidated, instead of a different purpose (reuse) that is not clearly understood (Van de Sandt *et al*., 2019). A more detailed analysis of the article, however, could reveal that this is a strategy related to the perception that peer recognition is greater when it comes to original work (use) rather than secondary (reuse). This could be manifested in the author's concern about evaluating the impact of research through citations, which is also mentioned in the text.

In any case, much of the literature that addresses the topic does not seem prone to such a change, at least in the short term. Therefore, it is understood that this differentiation will continue, mainly because it is useful in the context of privacy and data protection, as will be discussed later. Next, taxonomies or models that seek to differentiate the ways in which data are reused are discussed.

13.4.2 Types of data reuse

Reuse can be understood as a broader category, which encompasses reproducibility, replicability, reuse, integration and reanalysis, among other terms (Curty, 2019; Pasquetto; Randles; Borgman, 2017; Van de Sandt *et al*., 2019).

Starting from the context of *big data* and not of research data, but without excluding them, after identifying practical, technological and legal barriers, Custers and Uršič (2016) propose a reuse taxonomy composed of three elements: recycling, repurposing and recontextualization. Data recycling refers to using data several times, but always with the same initial purpose; data repurposing refers to reusing data for distinct purposes for which they were primarily collected; and data recontextualization implies using data in different contexts from the ones they were initially obtained. This distinction is particularly important from a legal perspective, in situations that involve privacy and protection of personal data, as normally the research subjects' authorization to use the data is given on the condition that its use is made only within the scope of that study.

Pasquetto, Randles and Borgman (2017), based on previous literature, differentiate reproducibility from replication, on one hand, and integration from independent reuse, on the other hand. Reproducibility occurs when a research problem is formulated again based on the same data and methods, to validate, verify or confirm the research, whereas replication implies the use of new data to answer, with the same methods, a previous question. The independent reuse refers to an external agent that performs the reuse; for example, the reproduction of a study is an example of independent reuse. Integration involves the reuse of datasets combined with other data, whether they are the result of research by others or of new observations.

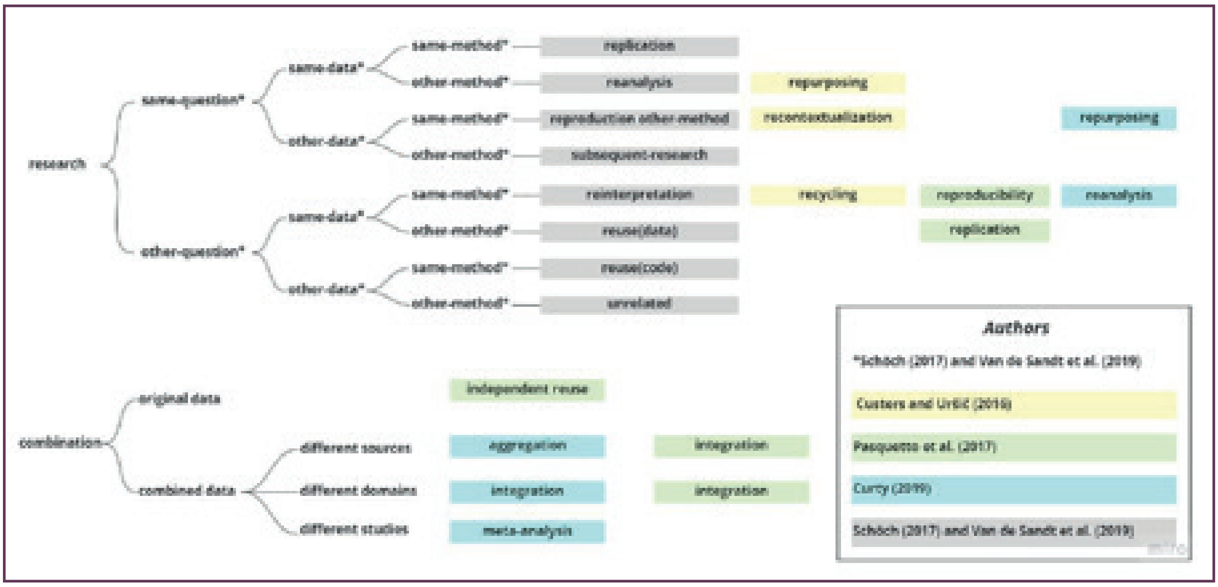
As can be seen, these are not excluding dimensions or that can be easily distinguished from each Other. Thus, in a later study, Pasquetto, Borgman and Wofford (2019) introduce another type of differentiation: reuse is a continuum that ranges from comparative to integrative. Comparative data reuse, as the term implies, involves using data for a specific comparison, which requires interaction experience, that is, knowing enough about the data to assess its quality and value. On the other hand, integrative reuse, such as using data in a new experiment, involves interpretation and, therefore, requires more specialized and in-depth scientific knowledge to be carried out, as well as greater confidence in the quality of data to be reused

Curty (2019) proposes a classification that describes five approaches for research data reuse, as follows, a) repurposing; b) aggregation; c) integration; d) metanalysis; and e) reanalysis. According to the author, in repurposing, data from a single study is fully or partially reused for new analysis, resulting from different research questions, without being complemented or integrated with data from other sources. In reuse by aggregation, data from different studies/sources are gathered to compose a more complete dataset. The reuse by integration is the one that combines data from different types of studies, through variables that connect separate studies. The metanalysis combines data from multiple independent studies, with very similar research questions, integrating them into a broader and more substantial analysis. The reanalysis involves the verification of original results, through new analysis, using the same methods and techniques, that is, it is the concept of reproducibility approached by Pasquetto, Randles and Borgman (2017).

As can be seen, there is no consensus among researchers on a definition for reuse or even on how to categorize its variables and specificities. An approach that seems promising for systematizing the different concepts of reuse based on three dimensions of research (question, data, and method), is that of Schöch (2017). The author derives eight categories of reproducibility in research, two of which (subsequent research and unrelated research) are not types of reuse, unlike the six others, which are directly related to reuse, namely: replication, reanalysis, reproducibility, reinterpretation, data reuse and code reuse.

To relate these different concepts used in the literature, it is proposed to group Custers and Uršič's (2016), PPasquetto, Randles and Borgman's (2017), Curty 's(2019) and Schöch's (2017) and Van de Sandt 's(2019) classifications (Figure 1). This grouping occurs based on the observation of two general criteria: types of reuse determined by the research context and types of reuse determined by the need to combine data. The types of reuse determined by research context are regrouped based on categories proposed by Schöch (2017) and Van de Sandt (2019): same/other research questions, same/other data and same/other research methods.

Figure 1 – Categorization of term related to the concept of research data reuse



Source: authors.

Figure 1 also shows that "repurposing", "recycling", and "recontextualization" by Custers and Uršič (2016), "reproducibility" and "replication" by Pasquetto, Randles and Borgman (2017), and "repurposing" and "reanalysis" by Curty (2019), assign data reuse to the research context. The combination of data with others is a criterion for "integration", "aggregation" and "meta-analysis", by Curty (2019), and "independent reuse" and "integration", by Pasquetto, Randles and Borgman (2017). Interestingly, in Schöch (2017) and Van de Sandt (2019), there is an inversion between term and concept, about "replication" and "reproducibility", in relation to Pasquetto, Randles and Borgman (2017).

It is understood that the representation synthesizes the interpretations of the term reused in the literature about research data, while revealing the complexity of the initiatives to understand it effectively

13.4.3 Conditions for research data reuse

Conceptual definitions are essential, but the effective reuse of research data can only occur if conditions are offered to researchers and appropriate actions encouraged in the scope of research ecosystem.

In a study to evaluate the practices of research data reuse in situations in which such use failed, Yoon (2016) proposed ways of overcoming the problems. The author offers the following suggestions:

> a) ease of reusing, particularly related to the interoperability and access, is an initial condition for successful experiences with data reuse;
> b) understanding data through documentation can be a minor difficulty, at least for experienced researchers, although the process still represents a challenge;
> c) the main component of reuse experience that becomes flawed is the lack of support for data reuse, which shows the need to develop a supporting system for those who reuse research data.

The importance of documentation is frequently emphasized as an essential condition for a successful reuse. For instance, Curty (2019) discusses that to be reused, data needs to be considered relevant, complete, understandable and reliable, and that these attributes can only be observed if data are accompanied by supplementary information and descriptions about its origin and processing, that is, documentation contextualizing them. Although it is not exactly Yoon's (2016) results, the Curty (2019) points out the need to develop data reuse skills. This is also stressed by Estevão and Strauhs' (2020) work, which points to the requirement of informational literacy in data reuse by researchers, many of whom have no experience in the subject.

The results of the study by Kim and Yoon (2017) on the data reuse behavior of scientists show that there are significant variations between disciplines, as well as within them, in data reuse intentions. The usefulness of data, as perceived by scientists, the concern with their quality and the offer of resources in their organizations were considered the most important elements by interviewees for the reuse of data. At the disciplinary level, the availability of data repositories showed a significant positive relationship with the intention to reuse data.

In summary, for the reuse of research data to materialize and for it to fulfill the promise of improving the form and results of knowledge production for of society as a whole, it is necessary to mobilize the entire research ecosystem: people need incentives and training; institutions need to provide the necessary conditions and technologies must be exploited to their full potential.

13.5 FINAL CONSIDERATIONS

Data reuse is the centerpiece of the research ecosystem, and the benefits of data will only be effective if data are prepared with reuse as a principle, as when adopting FAIR principles. Reuse is also one of the goals of digital curation of research data, as it comprises actions that will maintain and add value to reliable data for present and future use. Furthermore, the values of this data can only be determined through a proper understanding of its reuse, in the context of its user community.

The term *research data reuse* refers to a complex concept. To help researchers or data curators to understand data reuse, this work aimed at developing an analysis of this term through its definitions, its relationship with other terms and characteristics that provide distinctions between use and reuse. It also approached ways of reuse, through identification of its types, categories, and benefits.

Further studies on reuse are necessary, as is the understanding of the perspectives and practices of researchers from different scientific communities since research data reuse will not be fully achieved through the simple availability of data in repositories. There is an ecosystem that needs to be mobilized to make it possible to achieve this end.

REFERENCES

BERGHMANS, S. *et al*. Open Data: the researcher perspective - survey and case studies, 2017. **Mendeley Data**, v. 1, 4 apr. 2017. DOI: 10.17632/bwrnfb4bvh.1

BORGMAN, C. L.; SCHARNHORST, A.; GOLSHAN, M. S. Digital data archives as knowledge infrastructures: mediating data sharing and reuse. **Journal of the Association for Information Science and Technology**, v. 70, n. 8, p. 888–904, 2019.

BORGMAN, C. L. The conundrum of sharing research data. **Journal of the American Society for Information Science and Technology**, v. 63, n. 6, p. 1059–1078, 2012. DOI: 10.1002/asi.22634

BOTÉ, J.; TÉRMENS, M. Reusing Data: Technical and Ethical Challenges. **DESIDOC Journal of Library & Information Technology**, v. 39, n. 6, p. 329-337, 2019. DOI: 10.14429/djlit.39.6.14807

COLLINS, S.; *et al*. **Turning FAIR into reality**: final report and action plan from the European Commission expert group on FAIR data. Bruxelas: European Commission, 2018. DOI: 10.2777/1524.

CURTY, R. G. Abordagens de reúso e a questão da reusabilidade dos dados científicos. **Liinc em revista**, v. 15, n. 2, 2019. DOI: 10.18617/liinc.v15i2.4777

CUSTERS, B.; URŠIČ, H. Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection. **International Data Privacy Law**, v. 6, n. 1, p. 4–15, 2016. DOI: 10.1093/idpl/ipv028.

DIAS, G. A.; ANJOS, R. L. D.; ARAÚJO, D. G. A, Gestão dos dados de pesquisa no âmbito da comunidade dos pesquisadores vinculados aos programas de pós-graduação brasileiros na área da ciência da informação: desvendando as práticas e percepções associadas ao uso e reúso de dados. **Liinc em revista**, v. 15, n. 2, 2019. DOI: 10.18617/liinc.v15i2.4683

ESTEVÃO, J. S. B.; STRAUHS, F. R. Letramento informacional para reúso de dados nas ciências sociais: requisitos e competências. **Informação & Informação**, v. 25, n. 2, p. 1-25, 2020. DOI: 10.5433/1981-8920.2020v25n2p1

HENNING, P. C. *et al*. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, v. 25, n. 2, p. 389-412, maio/ago. 2019. DOI: http://dx.doi.org/10.19132/1808-5245252.389-412

KIM, Y.; YOON, A. Scientists' data reuse behaviors: a multilevel analysis. **Journal of the Association for Information Science and Technology**, v. 68, n. 12, p. 2709-2719, 2017. DOI: 10.1002/asi.23892

KOERS, H. *et al*. Recommendations for Services in a FAIR Data Ecosystem. **Patterns**, v. 1, n. 5, 100058, 2020. DOI: 10.1016/j.patter.2020.100058

PASQUETTO, I. V.; BORGMAN, C. L.; WOFFORD, M. F. Uses and Reuses of Scientific Data: The Data Creators' Advantage. **Harvard Data Science Review**, v. 1, n. 2, 2019. DOI: 10.1162/99608f92.fc14bf2d

PASQUETTO, I. V.; RANDLES, B. M.; BORGMAN, C. L. On the reuse of scientific data. **Data Science Journal**, v. 16, n. 8, 2017. DOI: 10.5334/dsj-2017-008

PERRIER, L; BLONDAL, E.; MACDONALD, H. The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis. **PLoS ONE**, v. 15, n. 2, e0229182, 2020. DOI: 10.1371/journal.pone.0229182

ROWLEY, J. *et al*. Academics' behaviors and attitudes towards open access publishing in scholarly journals. **Journal of the Association for Information Science and Technology**, v. 68, n. 5, p. 1201-1211, 2017. DOI: https://doi.org/10.1002/asi.23710.

SALES, L. *et al*. GO FAIR Brazil: a challenge for Brazilian data science. **Data Intelligence**, v. 2, n. 1-2, p. 238-245, 2020. DOI: 10.1162/dint_a_00046

SCHÖCH, C. Wiederholende Forschung in den digitalen Geisteswissenschaften. In: Digital Nachhaltigkeit, 2017, Bern, Switzerland. **Proceedings [...]** Bern: Universitat Bern, 2017. p. 13-18. Available from: https://zenodo.org/record/277113#.X6D7u4hKi01 Access on: 14 out. 2020

TENOPIR, C. *et al*. Data sharing by scientists: practices and perceptions. PLoS ONE, San Francisco, v. 6, n. 6, e21101, 2011. DOI: 10.1371/journal.pone.0021101

VAN DE SANDT, T. *et al*. The definition of reuse. **Data Science Journal**, v. 18, n. 1, p. 1-19, 2019. DOI: https://doi.org/10.5334/dsj-2019-022

WALLIS, J.C.; ROLANDO, E.; BORGMAN, C.L. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology**. PLoS ONE**, v. 8, n. 7, e67332, 2013. DOI: 10.1371/journal.pone.0067332

WILKINSON, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, v. 3, n. 160018, mar. 2016.

YOON, A. Red flags in data: Learning from failed data reuse experiences. **Proceedings of the Association for Information Science and Technology**, v. 53, n. 1, p. 1–6, 2016. DOI: https://doi.org/10.1002/pra2.2016.14505301126

How to cite this chapter: CAREGNATO, Sônia Elisa; ROCHA, Rafael Port da; GABRIEL JUNIOR, Rene Faustino. Data reuse: FAIR principles and the research ecosystem. *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. 13, p. 155-165. DOI: 10.22477/9788570131959.cap13.

14. BeFAIRandCARE: FAIR AND CARE PRINCIPLES IN RESEARCH DATA MANAGEMENT

Silvana Aparecida Borsetti Gregorio Vidotti[158]
Emanuelle Torino[159]
Caio Saraiva Coneglian[160]

14.1 INTRODUCTION

The current moment in society can be understood from an important evolution of digital technologies, but specifically from an understanding of the role of data in the decision-making process. In general terms, data collection, treatment, analysis, and availability became essential processes so that all sectors of the economy could survive and advance. In addition, the scientific research process itself was strongly impacted by this trend, and started to use, in a very significant way, data analysis and reuse for conducting research.

Such context lead to the development of studies related, for example, to data science, big data and scientific research and communication, which made it possible to understand the role of data in the current moment. Thus, an important aspect for understanding and analyzing the time we live in is the volume of existing data that, when treated and analyzed, are capable of generating great wealth for all processes involved.

It is essential to point out in this context the open data movement, which is mainly linked to government data. This movement demonstrates how the trend towards valuing data can contribute to the transparency of public authorities, as well as to improve the efficiency of services provided to the population. The open data movement is a trend across the planet and, when aligned with people's interests, it is capable of improving the well-being of a community.

In the scientific field, an increasingly valued trend is the open access movement, which aims to provide transparency and openness to the scientific results achieved in the research process. This movement aims for researchers to clearly demonstrate their research results, including the free availability of data collected and results achieved.

Even more broadly, open science is another movement strongly linked to research data availability that, when analyzed from the perspective of publication, demonstrates the potential of making research data available and shared. With the support of data repositories, data journals and data articles, such movement has led to a

[158] PhD by the Graduate Program in Education (Unesp). Faculty member in the Department and Graduate Program in Information Science (PPGCI - Unesp) at São Paulo State University - Unesp. E-mail: silvana.vidotti@unesp.br

[159] Doctorate student by the Graduate Program in Information (PPGCI-Unesp). Librarian at the Universidade Tecnológica Federal do Paraná (UTFPR). E-mail: emanuelle@utfpr.edu.br

[160] PhD by the Graduate Program in Information Science (PPGCI-Unesp). Universidade de Marília (Unimar). E-mail: caio.coneglian@gmail.com

transformation in the way research is carried out, with an impact on the various stages of the scientific process, including funding agencies that have started to demand data management plan.

Based on the trends presented, an appreciation of the aspects involving data in different areas is identified. From government data to research data, performing data management has become essential. In this sense, it is fundamental that all aspects involving the ability of machines to locate, access and interoperate data are understood and well-defined.

Additionally, there are other aspects that must be observed throughout the research process when working with data from specific people and communities, considering the data life cycle. These communities and the research data generated that involve them need to have a special treatment, which ensures data sovereignty, to maximize the benefits and mitigate the impacts, as well as enable adequate treatment for the availability, access and (re) use. Therefore, this book chapter aims to discuss the FAIR and CARE principles for managing human research data from specific communities.

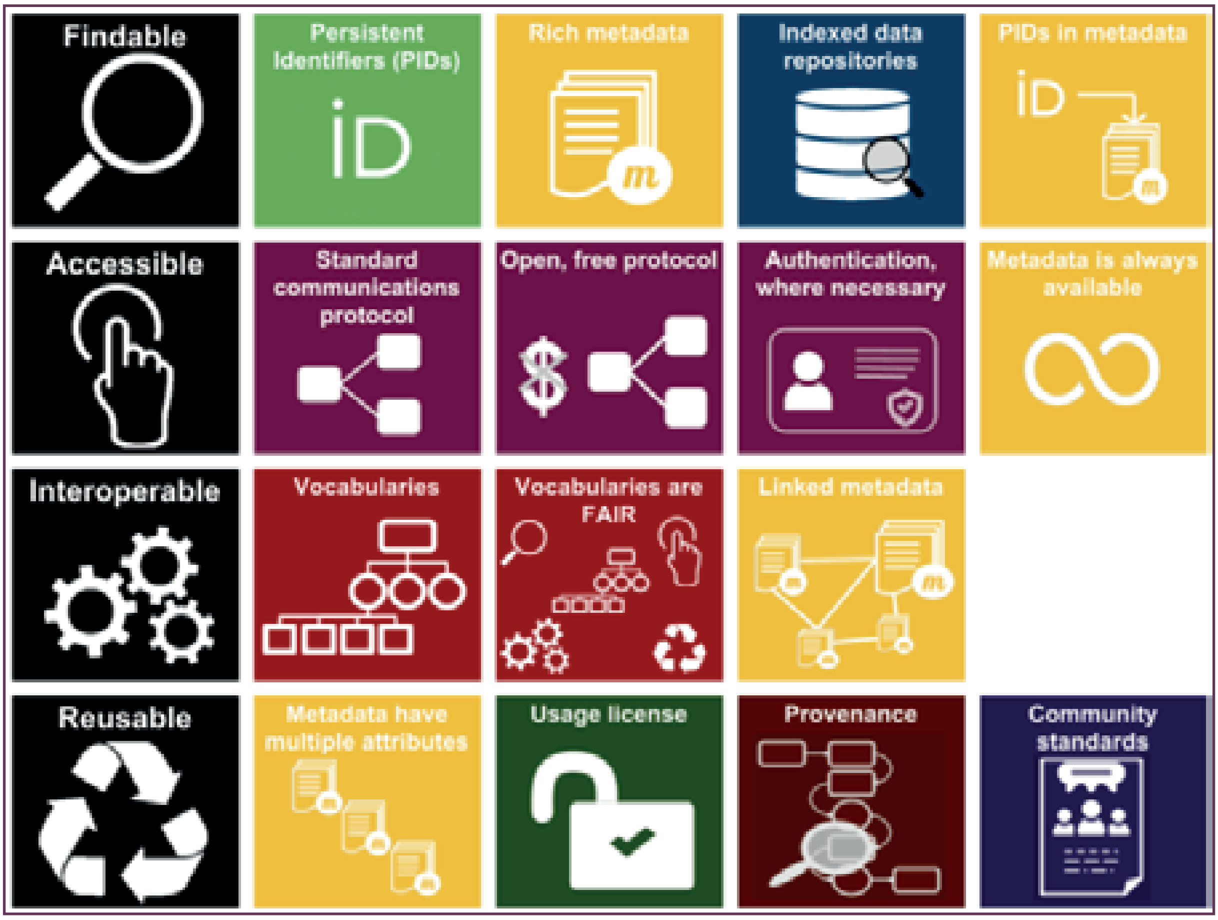
14.2 FAIR PRINCIPLES

FAIR principles (Findable, Accessible, Interoperable and Reusable) were proposed in 2016, as a way of dealing with computational aspects involved in making data available in different contexts, including research data. Such principles were developed understanding a scenario in which, in the same intensity that data become important for the scientific process and the whole Society, there was a difficulty for people and machines to be able to locate and use available data. Thus, "[...] FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals." (Wilkinson *et al*., 2016, p. 1).

In the initial proposal of FAIR, Wilkinson *et al*. (2016) point out the importance of computational agents in the environments that store data and highlight the need for these agents to be able to have information about the databases, in an interoperable way and with easy access. In this way, computational agents can help people to locate and use available data. The authors also report that, in environments with a huge number of databases, people depend on computational agents to relate to such data.

Based on this understanding, and having FAIR principles as reference, Australian National Data Service (2020) present FAIR principles in Figure 1.

Figure 1- FAIR Principles



Source: Australian National Data Service (2020).

Figure 1 highlights characteristics linked to each one of FAIR principles, showing elements that must be considered when making data available or when releasing them.

According to GOFAIR (2020), the Findable (**F**) principle assumes that in order for data to be used and/or reused it must be located, activates, readable and processable by humans and computer applications. In addition, it is necessary to adopt persistent identifiers for the data, describe them exhaustively through enriched metadata and make them available in an indexed infrastructure. The Accessible (**A**) principle reflects the capacity of a data set to be accessed and the specifications to do so, including the use of communication protocols, authentication, levels of access and metadata persistence, even if data is not available anymore. The Interoperable (**I**) principal aims to optimize the communication between different systems and the integration of different data sets. For this purpose, data, and metadata need to be readable and adequate to recognized standards and vocabularies, enhancing the link with other standards and including qualified references. Finally, the Reusable principle (**R**) aims at optimizing data reuse process. Data reuse is about how well data and metadata are described, including information on right of use, data provenance and context, in a way to allow data to be combined and reused by other instances.

According to FORCE21 (2020, local. 1 preamble):

> [...] through the definition and widespread support of a minimal set of community-agreed guiding principles and practices, data providers and data consumers – both machine and human – could more easily discover, access, interoperate, and sensibly re-use, with proper citation, the vast quantities of information being generated by contemporary data-intensive science.

It is noteworthy that FAIR principles apply to the treatment of data, metadata, and infrastructure to maximize the location, access, interoperability, and reuse of data.

14.3 CARE PRINCIPLES

The technological infrastructure and connectivity increase data value, therefore, principles for the processes of collection, storage, and availability are essential. In this matter, and considering indigenous data sovereignty – presented by Stone and Calderon (2019, local. introduction) as "[...] the rights of Indigenous Peoples and nations to govern themselves and the data about them [...]" -, on which only indigenous peoples have primacy to take decisions, according to their interests and values, Global Indigenous Data Alliance established in 2018, at the International Data Week and Research Data Alliance Plenary, CARE Principles for Indigenous Data Governance.

CARE is an acronym for Collective Benefit, Authority to Control, Responsibility, Ethics.

> [...] the 'CARE Principles for Indigenous Data Governance' were developed by the Research Data Alliance (RDA) International Indigenous Data Sovereignty Interest Group. They aimed to empower Indigenous Peoples, by shifting the focus of data governance from consultation to values-based relationships that promote equitable Indigenous participation in processes of data reuse, which will result in more equitable outcomes, as well as preserving relationships built on trust and respect. (Carroll *et al*., 2020b, local. introduction).

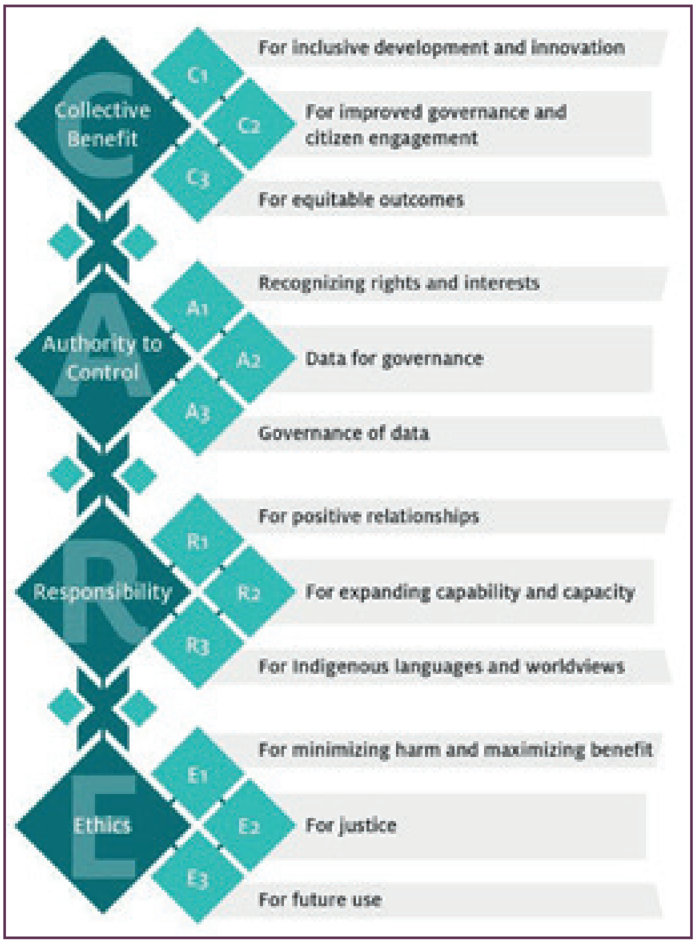
Such initiative is based, according to Global Indigenous Data Alliance (2019), on the Declaration of the United Nation on the Right of Indigenous People (Nações Unidas, 2008, p. 1), which recognizes the auto-governance and authority rights of indigenous peoples on their cultural heritage. The "languages, knowledge, practices, technologies, natural resources and territories" are considered by them as indigenous data, often expressed orally and considered essential for their development and rights.

In this way, it has become necessary for indigenous people to establish principles that make it possible to indiscriminately make decisions about their data, since the global movement around open data, whether governmental or research, does not compromise with the aforementioned interests and that data exchange favored by the movement and structured on principles such as FAIR do not mention ethical and cultural characteristics, and/or characteristics from historical context. According to Carroll *et al*. (2020a), it happens due to the tension of indigenous community in protecting their data and interests and supporting initiatives such as openness, data sharing and machine learning, aiming for researchers, managers, and data users to be "fair and care".

Therefore, CARE principles seek to establish governance over people-drive data. "These principles complement the existing FAIR principles, encouraging open and other data movements to consider both people and purpose in their advocacy and pursuits." (Global Indigenous Data Alliance, 2019, p. 1). Carroll *et al*. (2020a) emphasize that CARE principles were designed to complement FAIR principles, aiming to include indigenous peoples so that they can be implemented together; in addition, they emphasize that the governance of indigenous data encompasses the administration and control over data, which includes the processes of collection, storage, analysis, use, and reuse.

The four CARE principles are structured into 12 sub principles associated to it, according to Figure 2.

Figure 2- CARE Principles for Indigenous Data Governance


Collective Benefit
C1
For inclusive development and innovation
C2
For improved governance and citizen engagement
C3
For equitable outcomes
Authority to Control
A1
Recognizing rights and interests
A2
Data for governance
A3
Governance of data
Responsibility
R1
For positive relationships
R2
For expanding capability and capacity
R3
For Indigenous languages and worldviews
Ethics
E1
For minimizing harm and maximizing benefit
E2
For justice
E3
For future use

Source: Carroll *et al.* (2020, p. 5).

The first principle, Collective Benefit (C), establishes that "Data ecosystems shall be designed and function in ways that enable Indigenous Peoples to derive benefit from the data." (Global Indigenous Data Alliance, 2019, p. 2). For that purpose, **C1** - For inclusive development and innovation – governments and institutions must support data (re)use by indigenous peoples and communities, aiming at innovation, value generation and local development; **C2** - For improved governance and citizen engagement – data enables the involvement of governments, institutions and citizens, provide transparency and assist in planning, evaluating and decision-making, in addition to providing information on the indigenous peoples' interest; **C3** - For equitable outcomes – indigenous data are related to their values and can be extended to society, in a way that all must benefit indigenous peoples.

The second principle Authority to Control (**A**) consists in:

> Indigenous Peoples' rights and interests in Indigenous data must be recognized and their authority to control such data be empowered. Indigenous data governance enables Indigenous Peoples and governing bodies to determine how Indigenous Peoples, as well as Indigenous lands, territories, resources, knowledge and geographical indicators, are represented and identified within data. (Global Indigenous Data Alliance, 2019, p. 3).

In **A1** - Recognizing rights and interests – indigenous peoples must have their rights and interests in their data and knowledge recognized, which is done through free, prior and manifest consent during data collection, including the data uses, data policies and protocols used during collection; **A2** - Data for governance – indigenous peoples must exercise governance on their data, which must be available and accessible to them; **A3** - Governance of data – indigenous people can develop protocols of governance and access to their data, especially those regarding indigenous knowledge.

The third principle, Responsibility (**R**), establishes that:

> Those working with Indigenous data have a responsibility to share how those data are used to support Indigenous Peoples' self-determination and collective benefit. Accountability requires meaningful and openly available evidence of these efforts and the benefits accruing to Indigenous Peoples." (Global Indigenous Data Alliance, 2019, p. 4).

In this matter, **R1** - For positive relationships – indigenous data use is possible when based on relationships of respect, reciprocity and trust, so that the researcher is responsible for ensuring that data collected, their interpretation and use guarantee and respect the dignity of indigenous peoples; **R2** - For expanding capability and capacity – the use of indigenous data requires reciprocal responsibility, competence in data with indigenous people and the development of digital infrastructures that enable data collection, management, security, and governance; **R3** - For indigenous languages and worldviews - such resources must generate data based on languages, experiences, values, principles and world perspectives of indigenous peoples.

The fourth principle, Ethics (**E**), "Indigenous Peoples' rights and wellbeing should be the primary concern at all stages of the data life cycle and across the data ecosystem." (Global Indigenous Data Alliance, 2019, p. 5). consists in: **E1** - For minimizing harm and maximizing benefit – minimize damages that can derive from stigmas or deficits related to the indigenous peoples, guiding data collection, treatment, and use in ethical presets, in line with indigenous ethical structures and with rights established through the United Nations Declaration on the Rights of Indigenous Peoples; **E2** - For justice – use ethical processes to address imbalances in power and resources, as well as the way they affect indigenous and human rights; **E3** - For future use – data governance must consider the potential future use based on ethical bases, values, and principles of indigenous peoples, and also express provenance, purpose, rights of use, including limitations, obligations in use and consent in the metadata.

14.4 #BeFAIRandCARE: DISCUSSIONS AND NOTES

Global Indigenous Data Alliance (GIDA)[161] expresses, through "#BeFAIRandCARE", that FAIR and CARE principles are complementary when considering technologies, purposes, and people in open data movements. Thus, while FAIR mainly emphasizes the computational aspects given their relevance so that computational agents can help humans in the face of the expressive volume and complexity of data; CARE, as principles for the governance of indigenous data, emphasizes, above all, people and purpose, considering the relevance of data for the advance, self-determination, and sovereignty of indigenous peoples.

Regarding CARE, although it was designed for the governance of indigenous data, Carroll *et al*. (2020a) highlight that indigenous peoples, nations, people, and communities are actors in contemporary global societies. Therefore, CARE principles address aspects relevant to different populations, such as social minorities, communities and groups, who want or need to maintain different levels of treatment and responsibility for their data use. And, among these aspects, the authors highlight privacy, use, reuse and management, which can constitute elements for the establishment of standards, policies, and agendas.

When considering the activity of scientific research, it is evident the convergence of different movements that determistonene standards, principles and general practices and their domains, whose peculiarities must be met. Stone and Calderon (2019, local. a journay) reaffirm that "The CARE principles certainly prompt us all to consider the people reflected in data and how our actions with it may impact on them."

The need to manage research data since planning through the research data management plan is an action increasingly required from researchers by education and research institutions, funding agencies, repositories and scientific and data journals. And, although it may seem like a bureaucratization of scientific work, it allows the researcher and others involved in a research project to plan and record the processes and decisions made during the investigation to document each step, which allows for an optimized management, considering the life cycle of research and research data.

In most cases, institutions list a set of tools and templates to support the researcher in the elaboration of research data management plan, which are important elements for data management, availability and future use. However, such templates are generic and open, When They could guide the researcher in the planning and execution of the research stages already based on guiding principles, among which we highlight FAIR and CARE. This way, the necessary elements are treated during the research process and ensure that data opening is properly done.

Throughout the research data management process, since planning, collection, availability and reuse, it is essential to pay attention to the relevance of these data treatment so that they can respect technological principles that make them are easily processable by computational applications, and, therefore, be reused in robust infrastructures and by humans. On the other hand, human principles linked to people, purposes, and the consequences

[161] Available in: https://www.gida-global.org/care. Access on: 30 sept, 2024.

that the collection, storage, and availability of data from specific communities can bring to the development of the communities themselves must be respected.

We emphasize, from Carroll *et al*. (2020a), that although CARE principles have been defined for the governance of indigenous data, whose context is recognized as relevant, it is possible to expand them to other specific communities that may also need them so that they have governance and can develop from their data. In this matter, we highlight other social minorities, such as quilombo communities, riverside communities, social settlements, suburbs and LGBTQ+ communities.

Researchers, institutions, funding agencies, governments, and policymakers are encouraged to ensure that the planning and execution of research, as well as the disclosure of its results, are properly carried out, especially when materialized in open research data, in addition to being based on different established and validated principles, among which we highlight FAIR and CARE

Advances in the sense of operationalizing such principles in a computational way are the object of study, involving domains, communities and correlated instances. According to Carroll *et al*. (2020b), Research Data Alliance, through the International Indigenous Data Sovereignty Interest Group and the FAIR Data Maturity Model Work Group, has already started the necessary discussions to operationalize CARE principles together with FAIR principles. From this perspective, they highlight that one of the challenges is the need to apply CARE at all stages of the data cycle. Thus, issues related to the optimization of computational processes for the location, access and (re)use of data, as well as the sovereignty and equitable treatment of data, can be adequately addressed.

However, for this to occur, in addition to the proper treatment of research data, it is necessary for the researcher to adopt daily practices throughout the research process and throughout the data life cycle. This attitude will ensure that the processes take place fairly and carefully.

Therefore, in the context of scientific research and research data, #BeFAIRandCARE can be:

a. FAIR: make data available in an open way and in accordance with FAIR principles that favor location, access, interoperability and use, with:

- adequate and exhaustive use of metadata in data sets and for representation;
- adoption of standardized vocabularies;
- adoption of persistent identifiers;
- selection of proper digital environment for data availability and/or release;
- use of open formats, protocols, and standards;
- determination of use license;

- maintenance of reasonable periods of embargo;
- adequate indication of data provenance;
- data versioning;
- data preservation;
- acknowledgement of ownership when using research data collected by third parties;
- use of data in accordance with the provisions of the use license.

b. CARE: respect CARE principles, considering the hegemony of specific communities, their world perspective, sovereignty, and governance on data, with:

- focus on inclusive development;
- establishment of equitable relationships of trust, reciprocity, and respect;
- compliance with ethical and legal precepts in the collection, treatment, storage, and availability of data;
- data identification;
- protection of rights, interests, values, and culture;
- participation in governance and control over data;
- improvement in data representation;
- training to use data.

Being fairly consists in formalizing data and making it available following good practices and principles so that they can actually be reused by humans and computational applications. And being carefully includes and expands on the previous aspects by properly dealing with research involving human beings, especially specific communities and social minorities. Thus, #BeFAIRandCARE practices must clearly be present in the daily routine of those who work, directly or indirectly, with processes related to the collection, analysis, treatment, storage, and availability of data, especially when they involve human beings.

14.5 FINAL CONSIDERATIONS

FAIR and CARE principles seek to build reliable, fair and responsible data practices, in both management and governance processes, as well as in results and in the quality of available data sets.

It is worth noting that CARE principles are involved in the whole data life cycle, starting in data management plan, passing through the processes of collection, representation, storage, and potential data availability and reuse, respecting the collective benefits, the authority to control, responsibility and ethics. On the other hand, FAIR principles are also linked to the life cycle; however, they focus on technological infrastructure so that data can be findable, accessible, interoperable, and reusable.

Therefore, it is recommended to adapt data management plans to the FAIR and CARE principles, and the adoption by researchers of #BeFAIRandCARE practices throughout the research process and in the data life cycle, especially in research related to people from specific communities for the equitable treatment of data.

In this chapter, the discussion begins by pointing out ways that can be followed in working with research data involving human beings. Such discussions can be in-depth, for example, based on the United Nations (UN) recommendations for human rights, the Sustainable Development Goals (SDGs) and national and international legislation for the management of personal data.

REFERENCES

AUSTRALIAN NATIONAL DATA SERVICE. **FAIR data training and resources**. [*S. l.*]: ARDC, 2024 . Available from: https://ardc.edu.au/resource/fair-data-training-resources/. Access on: 30 sept. 2024.

CARROLL, Stephanie Russo; GARBA, Ibrahim; FIGUEROA-RODRÍGUEZ, Oscar L, *et al*. The CARE Principles for Indigenous Data Governance. **Data Science Journal**, [*S. l.*], v. 19, n. 1, p. 43, nov. 2020a. DOI: http://doi.org/10.5334/dsj-2020-043. Available from: https://datascience.codata.org/articles/10.5334/dsj-2020-043/#:~:text=The%20CARE%20Principles%20are%20a,value%20of%20data%20for%20reuse. Access on: 5 feb. 2021.

CARROLL, Stephanie Russo; HUDSON, Maui; HOLBROOK, Jarita; MATERECHERA, Simeon; ANDERSON, Jane. **Working with the CARE principles**: operationalizing Indigenous data governance. London: Ada Lovelace Institute, nov. 2020b. Available from: https://www.adalovelaceinstitute.org/blog/care-principles-operationalising-indigenous-data-governance/. Access on: 6 feb. 2021.

FORCE21. **Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1.0**. Calfórnia: Force21, 2020. Available from: https://www.force11.org/fairprinciples#Annex1-1. Access on: 7 feb. 2021.

GLOBAL INDIGENOUS DATA ALLIANCE. **CARE Principles for Indigenous Data Governance**. [*S. l.*]: Data Alliance, sept. 2019. Available from: https://static1.squarespace.com/static/5d3799de845604000199cd24/t/

5da9f4479ecab221ce848fb2/1571419335217/CARE+Principles_One+Pagers+FINAL_Oct_17_2019.pdf. Access on: 5 feb. 2021.

GOFAIR. **FAIR principles**. [*S. l.*]: GOFAIR, 2020. Available from: https://www.go-fair.org/fair-principles/. Access on: 7 feb. 2021.

NAÇÕES UNIDAS. **Declaração das Nações Unidas sobre os direitos dos povos indígenas**. Rio de Janeiro: Nações Unidas, mar. 2008. Available from: https://www.un.org/esa/socdev/unpfii/documents/DRIPS_pt.pdf. Access on: 5 feb. 2021.

STONE, Paul; CALDERON, Ania. **[Spotlight] CARE Principles**: unpacking indigenous data governance. [*S. l.*]: opendatacharter, nov. 2019. Available from: https://opendatacharter.medium.com/spotlight-care-principles--f475ec2bf6ec. Access on: 6 feb. 2021.

WILKINSON, M., DUMONTIER, M., AALBERSBERG, I. *et al*. The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, [*S. l.*], v. 3, n. 160018, mar. 2016. DOI: https://doi.org/10.1038/sdata.2016.18. Available from: https://www.nature.com/articles/sdata201618. Access on: 6 feb. 2021.

How to cite this chapter: VIDOTTI, Silvana Aparecida Borsetti Gregorio; TORINO, Emanuelle; CONEGLIAN, Caio Saraiva. BeFAIRandCARE: FAIR and care principles in research data management. *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. 14, p. 166-176. DOI: 10.22477/9788570131959.cap14.

15. AN IMPLEMENTATION MODEL FOR THE INTERNET OF FAIR DATA & SERVICES

Luís Fernando Sayão[162]
Luana Farias Sales[163]

15.1 INTRODUCTION

There is nothing new in saying that the vast and growing amount of data and information that spreads throughout contemporary society is profoundly reshaping its modus vivendi in all its dimensions, activated by techno-social systems: leisure, education, public administration, business, health care, cultural expression and, above all, personal interlocutions. This intensively connected and planetary distributed "infosphere" becomes possible through the vertiginous advance of computer and network Technologies – and their digital materiality – that provide the creation, capture, copying, transmission, sharing and massive storage of information in a massive, easy and low--cost way. (National Research Council, 2015). It is to be expected that these transformations overlap in a forceful way with the processes of construction of Science knowledge.

Regardless of the point of observation, what can be seen is that scientific research – due to this global trend allied to its immanent connections with technical systems – is producing an enormous and growing flow of digital data. Countless sensors installed in the most diverse devices ranging from distant satellites, particle accelerators, automatic DNA sequencers, even in unpretentious medical implants, allow data to be captured in an unprecedented amount in all scientific domains, from the exact sciences to humanities, art, and culture.

In light of these findings, research data management is currently a focus of interest and one of the greatest challenges for research organization. As a development, data management and curation, on a planetary scale, stand out with prominence in the 21st century research scenario, as well as the ubiquity of digital Technologies for data collection, analysis and archiving in almost all disciplinary domains (Mayermik, 2012). Therefore, research institutions, in different gradations, are reconceptualizing data management and identifying it as an integral part of research processes, reconsidering or expanding their data treatment strategies, implementing management and curation platforms, acquiring analysis and developing training programs for their teams.

There are many motivations for the implementation of new modalities of information services that support data management in academic and research environments, among them is the need to support research activities, accelerate scientific progress and innovation through intense national and international sharing and collaboration (Mushi, 2020). However, we can identify reuse and reproducibility as the main objectives of data management

[162] PhD in Information Science by PPGCI IBICT-UFRJ, researcher at CNEN, professor at PPGCI IBICT-UFRJ, luis.sayao@cnen.gov.br

[163] PhD in Information Science PPGCI IBICT-UFRJ, professor at PPGCI IBICT-UFRJ, luanasales@ibict.br

and important parameters, from which other benefits are constituted. There are many motivations for storing and preserving data, but the primary reason is reuse and reproducibility, emphasizes Borgman (2007).

The planning, development and implementation of research data management platforms, due to the number of variables that need to be addressed, are complex and multifaceted problems. They need to be articulated around workflows, specific disciplinary domains, informational, technological, political, ethical and legal parameters, sustainability, and expertise in an Odyssey marked by constant changes, whose sign is heterogeneity.

This complex environment can be a suitable terrain for the adoption of FAIR Principles as a horizon for the implementation of management services that make research data findable, accessible, interoperable, so that they can be reused for the long term, thus creating conditions for the transition from a self-contained research to a more open, networked and cooperative research, which, at the same time, meets disciplinary requirements that benefit communities of specific cultures and constraints. "FAIR Principles are not magic and do not represent a panacea, but they guide the development of infrastructures and tools that make all research objects optimally reusable for machines and people" emphasize Barend Mons *et al.* (2017, p. 55), founders of this movement. However, the alignment and implementation of FAIR principles in a research institution requires financial investments, cultural changes, training and building technical infrastructure (Graaf; Waaijers, 2011), factors that can be put together around the concept of "Platform of research data management". This type of platform has the potential of operationalize the several layers of management and establish an increasing infrastructure of informational, scientific and computational services towards applying FAIR Principles in research objects, which is called FAIRfication process, whether data as such or algorithms, codes, procedures, workflows or other physical or conceptual devices that lead to data.

In the attempt to equate this diversity, the work herein aims to present a generic architecture to support data service platform project by defining, realigning, aggregating and articulating the several conceptual models – guidelines, policies, services, tools, infrastructures, among others – around a layer model that, as building blocks, can be adjusted according to the depth, extent, and philosophy of each institution or discipline. The model aims to build a possible scale for measuring the level of maturity of management service projects. The proposal architecture aims at making data adherent to FAIR principles, opening the prospect for a growing number of applications and services can link and process FAIR data, making real the idea of "Internet of FAIR & Data Services" – IFDS– which unfold into several benefits for the various stakeholders involved.

To outline the elements of the proposed architecture, the analysis of the literature in the area was taken as a methodology, with special emphasis on articles, reports, annuals, and data infrastructure projects developed by researchers and research institutions.

15.2 SOME CONSIDERATIONS ON FAIR PRINCIPLES AND THEIR IMPLEMENTATION

The notion of proper research data management, idealized in a way that it can maximize the opportunities of finding data and the efficient reuse of research results by humans and machines, is not exactly new as it has been present for decades in scientific research domains, especially by semantic web and ontology engineering communities. Along this path, many options of implementation have already been carried out by pioneer communities to associate data management to the notion of "machine actionability". FAIR principles can be considered a synthesis of these previous efforts and emerged from the materialization of a view, from multiple stakeholders, of an infrastructure to support the reuse of data that can be processes by computers (Wilkinson *et al.,* 2016), which was later coined "Internet of FAIR & Data Services" (Jacobsen *et al.,* 2020; Mons *et al.*, 2017).

FAIR Data Principles advocate that all research products should be findable, accessible, interoperable and thus reusable by humans and machines, expressing the researchers' expectations regarding data resources in current science, and offering a guide for data producers and publishers so they can navigate more securely and objectively around the complexity inherent to research data management. The primary focus of the Principles is to ensure that data can be reusable, by both humans and machines, in subsequent research and transversally reinterpreted accelerating interdisciplinarity and innovation, becoming even more valuable; and also maximizing the added value obtained by the developments of academic publications that have digital and network technologies as their substrate (Wilkinson *et al.*, 2016). In this direction, FAIR Principles outline considerations that are part of contemporary publications of research data and are related to the deposit, exploitation, sharing, and reuse of these resources through manual and automated processes. As such, they describe the characteristics that data resources, tools, vocabularies, and infrastructures must have to support discovery and reuse by other stakeholders in subsequent endeavors, now and in the future.

Unlike different initiatives shaped by disciplinary domains that establish specific practices for managing and archiving data, FAIR "describes high-level, concise, domain-independent principles that can be applied to a broad spectrum of research product" (Wilkinson *et al.*, 2016, p. 2), however, it may be "a basis for the development of flexible community [and disciplinary] standards" (Boeckhout; Zielhuis; Bredenoord, 2018, p. 932). Despite this neutrality, well-known standards, such as WC3 Resource Description Framework (RDF) with formal ontologies, are currently frequently applied solutions for interoperability and information and knowledge sharing that meet FAIR requirements, especially at the metadata level (Mons *et al.*, 2017, p. 51).

As a high-level design, the adoption of FAIR Principles precedes implementation choices, which do not recommend any specific technology or solution, which does not constitute a norm, standard, or specification. However, they offer a set of guidelines for management focused on the reuse of digital research resources. The elements of the four FAIR principles are related, yet independent and separable, and can be implemented in any combination and incrementally as the publishing environment evolves towards higher levels of "FAIRness". The importance and degree of implementation of each principle may depend on the priorities and maturity of each community in the use of certain research objects (Hong *et al.*, 2020). These characteristics contribute to the broad adoption of the principles, as specific communities, including those outside the scientific world, can implement their own FAIR solutions, allowing them to be reconfigured over time to follow the evolution of the underlying technologies (Jacobsen *et al.,* 2020). Therefore, it must be recognized that different disciplines require different types of technical solutions to achieve the same benefits of FAIR data.

It is crucially important to note that the application of FAIR principles extrapolates research data in its most conventional sense. In the narrower scope of scientific and methodological practices, FAIR principles should also be extended to algorithms, codes, tools, methodologies and workflows, objects that lead to the acquisition of data and that, if well documented, allow tracking the provenance of these assets. Thus, they need to be identified, described and reuse, like data.

All digital research objects – from data to analytical pipelines – benefit from the application of these objects, as all the components of the research process must be available to ensure “transparency, reproducibility and reusability” (Wilkinson *et al.*, 2016, p. 1).

This characteristic brings FAIR principles closer to the assumptions of open Science, whose considerations need to go beyond conventional publications.

The primary idea of implementing Internet of FAIR Data & Services is not executed by itself. For such, a data management process is needed that can effectively add value over time. The degree of adherence of research products to FAIR principles is linked to the scope and depth of management to which they are submitted. This assumes the need for a multi-layered framework – scientific, technological, informational and governance, which address the numerous ethical, methodological and organizational problems that lodge between the flows of sharing, integrity, reproducibility, research accountability, as well as new needs and opportunities for large-scale analysis and reanalysis (Wilkinson *et al.*, 2016).

15.3 FAIR PRINCIPLES X SERVICE MANAGEMENT

The implementation of FAIR principles is due to the varying degrees of actions applied to data by the set of data management services made available mainly by the various disciplinary platforms. These sets of FAIRification are captured by the models in three categories: informational, computational and scientific. Boeckhout, Zielhuis and Bredenoord (2018) make this relationship obvious in their analysis for the area of genome that, however, can be generalized.

- The **Findability** principle stipulates that data must be easy to find by humans and machines. In this way, data must be **identified, described and recorded or indexed in a clear and unambiguous way so that they can be located, and their content understood by humans and computational explorers.** In terms of services, this means that a unique and persistent **identifier** must be assigned in a data collection; that the main features are systematically specified, ideally using standardized formats; and that it is deposited or indexed in a public device such as an archive or data center or a disciplinary or institutional repository, which emphasizes the need for information services and management infrastructure. Meaningful and machine-actionable metadata is essential for the automatic findability of relevant datasets and services, and therefore is an essential component of the FAIRfication process (Jacobsen *et al.*, 2020).

- **The Accessibility principal** advocates that research objects are preferably accessible through the implementation, when appropriate, of automated data retrieval protocols; it also recommends that data are available according to clear and well-defined procedures. These conditions involve the establishment of authentication and authorization processes that are aligned to the organization policies and to the disciplinary culture, and also to data specificities – for example, sensibility level. As a FAIR mantra, which must be worked on by data management services, we have that metadata must be unconditionally accessible even if data is not, or is no longer available.

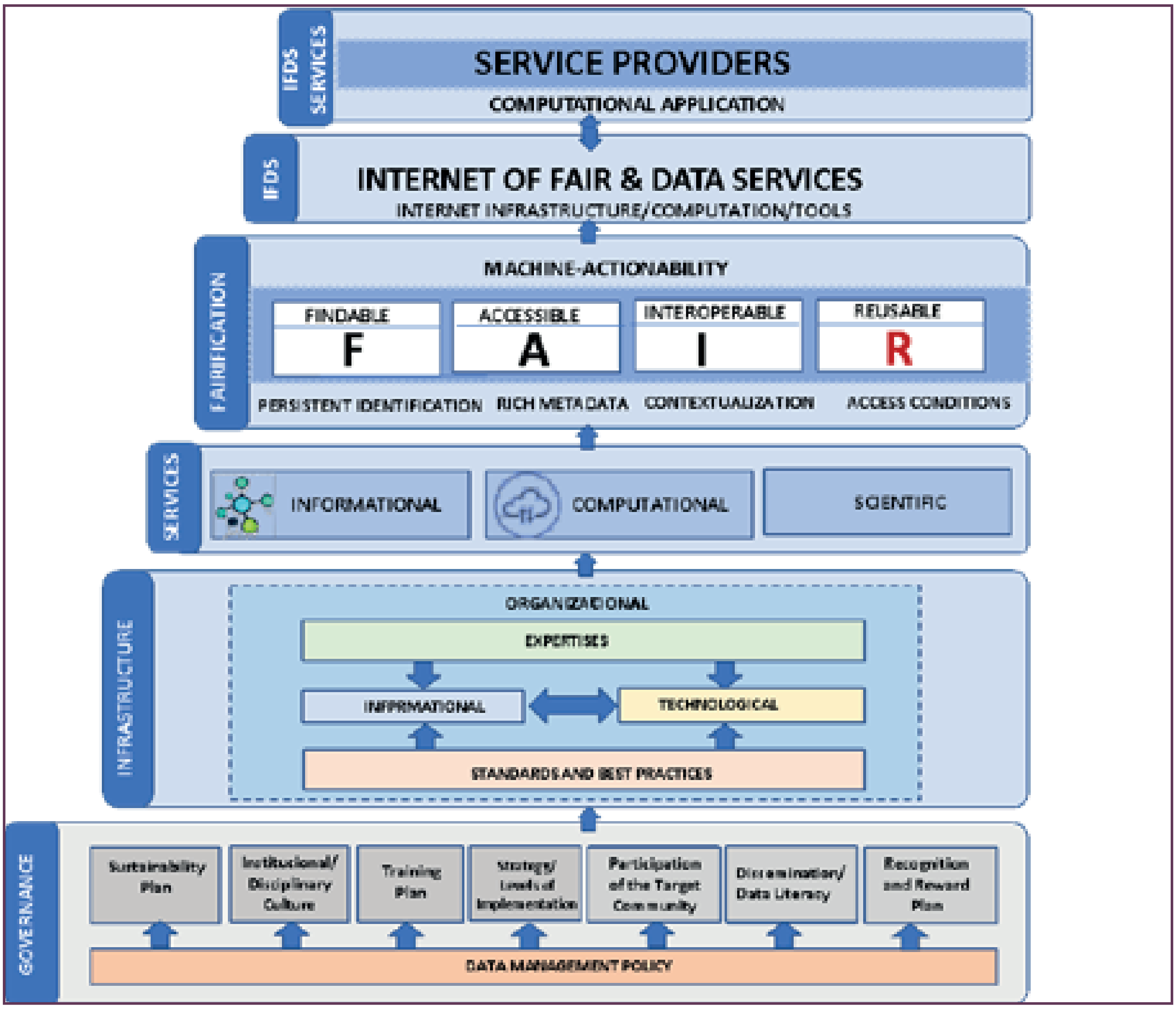
- **The Interoperability** principle is the most difficult to be implemented (Hong *et al.*, 2020), as it is conditioned to a high level of standardization in all its articulation. In general terms, when two or more digital resources are related to the same topic or entity, it must be possible for the machine to merge the information from each one of the resources in a unified and richer view of this topic or entity; similarly, when a digital entity is capable of being processed by an online service, a machine must be capable of automatically detecting this conformity and facilitate the interaction between data and this tool. It requires that the meaning (semantic) of each participant resource – whether data and/or services – is clear (Jacobsen *et al.*, 2020). In this sense, the Interoperability principle assumes that data and metadata are conceptualized, expressed and structured through widely accepted, published, trackable and accessible standards (that is, also FAIR). To achieve this objective, the implementation of this principle comprises a rigorous application of technical and semantic standards at all levels – scientific, computational and informational – such as in terms of variables, protocols, file formats, ontologies, and workflows.

- **Reusability** principle is a consequence of previous principles and reinforces important points advocated by them, such as the detailed description of data characteristics for human beings and computers, including the provenance, according to standards relevant to domain-specific communities. This allows a machine to decide: if a digital resource should be used – that is, if it is relevant for the task discussed; if a digital resource can be used and in what conditions – that is, if the resource meets the reuse conditions; and to whom to give credit in case the resource is reused. The degree of reusability points out the need for licenses appropriate to the conditions of use required.

15.4 DESCRIPTION OF THE PROPOSED MODEL

Research data management has many facets, but none of them can by itself fully explain the intrinsic complexity of its processes. Starting from this point, the model seeks to represent different aspects of research data management without losing sight of the interrelated nature of the dynamics of the activities that unfold in a data-intensive scientific environment whose objective is making data FAIR. As a convenient abstraction of reality that one wants to understand, a model is a cultural creation, a "mindfact", destined to represent a reality, or some of its aspects, to make them qualitative and quantitative describable and, sometimes, observable (Sayão, 2001). From this point on, it was decided to divide the model into six representational layers: 1) governance, where guiding principles of data management services project are discussed; 2) technical infrastructures which also include the necessary categories of expertise; 3) informational, computational and scientific services; 4) the results of the execution of these services manifested by data FAIRfication; 5) which, in turn, is consolidated in a global and shared environment, conceptualized as Internet of FAIR Data & Services (Sales *et al.*, 2020), where 6) Service Providers, through computational applications, offer diverse services. Figure 1 presents a general view of components grouped in layers and their interrelationships, which are discussed below.

Figure 1 – Model of implementation for Internet of FAIR Data & Services



Source: authors.

15.4.1 Research Data Governance : planning, policy, institutionalization, and sustainability

The organization configuration in which data management is performed can vary in relation to several aspects, such as the intensity of support to the management and level of investment applied. Some institutions such as scientific data reference centers and government statistical agencies may be entirely dedicated to data management, having it as their main purpose; in other settings, data management is part of a broader activity that connects to other research activities, such as, in the case of universities (National Research Council, 2015), whose data management activity is a result of their teaching, research and extension functions. However, even in the academic context, there are many ways to plan and execute data management tasks that vary according to objective references such as investment levels, available technical systems, volume, and type of data and how data management is integrated into its workflows and processes; and with more subjective perceptions such as disciplinary culture and academic prestige. In the present model, these parameters are equated by a more administrative level, understood by the term “data governance”. At a more conceptual level, data governance outlines the principles, policies, and strategies that are commonly adopted in an environment that needs a coherent data management; it also outlines the actions, functions, and roles that are required to implement these policies and strategies. Within a research institution, the principles, operationalized by management, govern the entire data lifecycle – from conceptualization to archiving and possible disposal. The data governance process treats data not only in its spatial aspect, but also along its temporal dimension (Solomonides, 2019), this requirement implies an increase in the degree of complexity and scope of governance commitments.

This structuring framework is necessary since digital research data can only be managed and preserved properly over time through a sustained institutional commitment (Mayermik, 2012, p. 1). To some extent, the consolidation of data management services reflects the level of organizational acceptance built into them and the degree of planning the various actions required: current sustainable budget, appropriate data policy, organic connection with target communities, compliance with ethical and legal codes, alignment with institutional strategic objectives and a development strategy that considers the possible paths for each institution. It is also necessary to consider the inevitability of the fact that the technological structures to access, interpret and preserve digital information are continually evolving; anticipating these problems and developing strategies to mitigate them is an activity relevant to governance commitments (National Research Council, 2015). Advanced data services that can appropriately support the entire life cycle of these information assets according to the interests of the various stakeholders involved can be developed based on these pillars. Considering these issues, we propose the following approaches as part of the model:

Institution Data Management Policy **–** Establishes the institution foundations, guidelines, and commitments regarding the management, use, ownership, compliance with ethical and legal ethical codes, adherence to funding agencies policies, national science, technology and innovation policies, to international guidelines and practices and, finally, but of critical importance, to culture, practices, and idiosyncrasies of communities and disciplinary domains: a comprehensive research data management policy must also identify the responsibilities of each of the actors – library, laboratories, information technology, management etc. – since data management involves different sectors of the institution (Mushi, 2020) and the project is considered as part of the institution research activities. It is necessary to emphasize that the process of developing an institutional data management policy requires extensive consultation with all stakeholders and the approval of relevant scientific communities and organizations (Wilson *et al.*, 2011). Policy guidelines must permeate the entire management cycle. "Policies can be an important motivating factor for FAIR data and other research objects (software, workflow, models, protocols etc.). Therefore, it is essential that "bottom-up" Community-based efforts are combined with policies with a "top-down" approach, complete Hong *et al*. (2020).

- *Institutional/disciplinary Culture-the* implantation of research data management services platform must be preceded by an analysis of requirements that considers the institutional, community and disciplinary context and culture and its unique characteristics. This process is expected to help define a more effective portfolio of data management services to support the research practices of the institution and its communities (Mushi, 2020; Coates, 2014; Reed, 2015). It is also important to recognize that some disciplines require different types of technical solutions to obtain the same benefits from FAIR data (Hong *et al.*, 2020).

- *Sustainability Plan* **–** One of the great challenges of a data management infrastructure implementation program is to ensure that each phase of the Project is sustainable as a continuous service over time (Wilson *et al.*, 2011). Once research data management is recognized as necessary for research activities, its costs must be estimated and its funding sources - especially perennial ones - identified. In this way, a project to implement research data management services needs to be associated with a sustainability plan that outlines a possible commitment to the present and the future. Creating and committing to a long-term strategy for services can more clearly reveal the resources needed for continuity of services and the infrastructure needed to do so. This, therefore, may include a succession plan (Mushi, 2020).

- *Dissemination/Literacy in data* **–** For the implementation of a FAIR research environment, it is necessary that the communities involved develop a shared understanding of what is limited by FAIR concept and Principles. In general, researchers and other stakeholders have a low level of perception about the importance of data management practices and the management and sharing requirements of funding agencies and data deposit commitments established with scientific editors, in addition to the ethical and legal issues involved in publishing the data. For example, in relation to the FAIR concept, Hong *et al.* (2020) observe that the researcher does not know what FAIR data is and often thinks it is the same as open data. This indicates that planning and dissemination and awareness actions are needed to elicit these issues. A dissemination program in this direction should include the development of didactic material (booklets and guides), courses, events, workshops, among others.

- *Knowledge/participation of the target Community* **–** As creators and users of research data, the engagement of researchers is crucial in the development of data management services. The provision of any service needs to be based on a close understanding of standards and flows of research that is developed in the institution, its motivations, characteristics, and priorities. Therefore, the precise definition of service requirements needs to be established with the commitment and contribution of the researchers' community; without these considerations, the characteristics of services may not be in accordance with researchers' goals. The community must be accompanied by changes in interest in data, and its participation in the development and choice of sharable standards for practices and for FAIR infrastructures must be recognized and institutionalized. The proximity, interaction, and alignment of communities with national and international organizations that deal directly with FAIR data management such as GO FAIR, RDA, CODATA, DCC and others, should be encouraged.

- *Training plan* **–** To offer complete services in data management, libraries need to have technologically qualified staff or greatly increase technological training for existing staff (Tenopir *et al.*, 2012). Human sustainability is critical to ensure the continuity and consistency of service offerings over time. However, few formal programs in informational studies include data management in their curricula; thus, research data managers are normally trained in service in the specific disciplines where they work (Borgman, 2007, p. 155).

- *Strategy/levels of implementation* **–** The development and implantation of data management infrastructure, in addition to many resources, require time to reach its full maturity and mirror the demands of the scientific communities, which implies the need to establish levels of implementation of infrastructure and services. Research libraries, for example, have, often, proactively sought to meet data management needs for their user communities. This often happens without additional financial support for the development and availability of data services. Therefore, libraries have to start on a simpler scale, building a base on which to develop more sophisticated services (Erway *et al.*, 2016, p. 5), starting with basic services that only require resources from the library itself, until they reach more complex services that require a high level of institutional commitment and more financial, technological and human resources (Kouper *et al.*, 2017).

- *Reward and recognition* **–** Research data management consumes time, resources and requires great dedication from the researcher; however, this effort is rarely noticed by the academic reward system, except When linked to publications in scientific journals. Therefore, to encourage this new task for researchers and highlight its importance, it must be properly recognized and that it is considered in the evaluation, promotion and hiring criteria.

15.4.2 Infrastructures of Research Data

Infrastructure is a broad and multidimensional notion. It can have a technical, legal, organizational connotation and, often, it is essential to also consider social, cultural and political aspects. Indeed, it is so in the science domain: research infrastructure projects are simultaneously a technological issue, a matter of identifying research needs in specific disciplinary areas, and a political issue. This more general perspective applies to institutional research data management infrastructures that need to provide technologies and tools, processes, policies, resources, and training for the various and diverse stages of data management.

Thus, just as institutions must provide basic infrastructure for research – such as laboratories, instrumentation, high-performance computing, networks, reagents and much more – they must also take steps to properly manage data. This assumes a broad spectrum of managerial, technological and informational activities that include information professionals trained to support researchers in the planning and management of their data, in the access of secure to secure storage devices and backups during project development and availability of access and long-term preservation platforms, necessary after the end of the research (Strasser, 2015); it is also essential to have a body of norms, standards and good practices that allow, mainly, a dialogue at different levels of systems and services, both local and global, which can be translated by interoperability.

When we compare traditional academic publishing with data publishing, we verify that the underlying infrastructures of academic publishing create an epistemological bridge between disciplines, having as aggregation point the research libraries that select, collect, organize and make publications of all kinds and all fields. Due to their nature, social institutions work to stabilize particular practices and forms of knowledge. In a certain sense, the institutions are social infrastructures in themselves. Therefore, the technical infrastructure is intertwined with the social infrastructures of the institutions, many times mediated by standards, protocols, documents, and devices that link the social and technical aspects of the infrastructures (Leonardi, 2010). However, there is no infrastructure of this magnitude for data. Some few areas have consolidated mechanisms to release data; others are in the stages of development of standards and practices to add their data and become the most widely accessible. "The lack of infrastructure for data amplifies discontinuity in academic publication" (Borgman, 2007, p. 155).

The infrastructure frameworks used for data management are diverse and fragmented in terms of flows, complexity, application and topology, and organized differently across various disciplines and in different countries (Graaf; Waaijers, 2011). However, the infrastructures increasingly shape standards and data management practices.

Therefore, knowledge about the origin, disciplinary domain, degree of processing, collection system, workflows etc. seem to be of essential importance in the conception of infrastructures for data management (Sayão; Sales, 2020).

In this model proposal, we consider five necessary types of infrastructure: standardization, technological, informational and organizational.

Standards and Best Practices-standards are consensual ways of codifying knowledge that circulates transversally through communities to ensure uniformity and similarity in our products and processes through time and space. They reflect more current knowledge about professional practices and increase interoperability, consistency, preservation, reusability, security, and protection of digital collections. Therefore, ensuring that in a scientific ecosystem, in which infrastructures are globally dispersed, its products are aligned with FAIR Principles, as well as they have a satisfactory degree of quality and excellence and are appropriate to researchers' needs, requires a body of standards and practices widely adopted and shared. Considering this fact, it is proposed that a consensual body of standards and best practices establishes infrastructures that must underlie the data management processes. This is because it is expected that data collections are suitable to be used for a wide variety of purposes – and not only for the purpose for which they were initially collected. To do so, they need to be added to other collections in other systems, shared, accessed, analyzed and archived using a wide spectrum of technologies. This condition makes a body of standards and common practices an essential infrastructure for the management and curation of research data. As the principle and practices of research data management develop, they begin to acquire knowledge as a distinct field of knowledge and to draw the attention of organizations interested in their improvement, such as DCC, Codata, GOFAIR, DataOne, DataCite, among many others. In this regard, standards and procedures commonly adopted for data management are taking part in many disciplines and sectors and are being redefined in other disciplines. As a result, practices improved to ensure digital data quality and durability have been continuously established. (National Research Council, 2015).

- *Technological Infrastructure* – Comprises a broad set of activities, equipment, processes, and expertise that can enable operational technological requirements necessary to data management cyberinfrastructures, such as logical, physical and virtual data organization; devices for high-performance processing, grid computing and storage of local or cloud data collections; local networks, communication, external connections, internet, web services; acquisition/development of scientific codes, workflow software; equipment for data analysis and view; physical, logical and network security strategy.

- *Informational Infrastructure* **–** Comprises persistent representation and identification schemes; descriptive, technical, administrative, preservation and disciplinary metadata; apart from taxonomy, ontologies, classification schemes; databases; it also includes repositories, digital libraries and reliable platforms for long-term archiving.

- *Personnel Infrastructure* **–** The many research institutions develop the most diverse approaches to data management. This assumes support teams made of different professionals (Pinfield; Cox; Smith, 2014). Roles as data stewards and data scientists are emerging in the world of contemporary science and joining the more traditional of researchers, lab technicians, research assistants and analysts; on the other hand, within the scope of specialized libraries and repositories, new stakeholders as librarians and data archivist and curators make the connection between libraries and laboratories and support the management of disciplinary idiosyncrasies of data life cycles (Ball, 2012). However, an essential requirement – especially when it comes to services associated with curation – is the need to know the disciplines and domains in which data are collected, processes and used. Without some familiarity with the problem to be addressed, he disciplinary culture, the goals to be pursued, as well as the methods used, nomenclature and practices of Fields in which digital assets are used, curators will not be able to make the most correct decisions to manage these assets for current and future use (National Research Council, 2015).

- *Organizational Infrastructure* – The infrastructure framework assumes, like governance, an anchoring based on some organizational structure aimed at research, as a university, research institution, or even a company whose projects depend on data management. These organizations offer technologies and tools, processes, policies resources and training to several and diversified stages of data management.

These infrastructure aspects – which enable interrelation of knowledge and practices that are underlying to equipment, installations, methodologies and mainly people – provide several services, tools, and processes that continuously put research objectives in line with FAIR principles. These limits are not always clear, for instance, the repositories are points of aggregation of technologies, standards, informational resources and expertise around archiving of research objects and constitute an essential link to reach Internet of Fair Data & Services, bring various stages of data management life cycle together in their research environments.

15.5 SERVICES FOR DATA FAIRFICATION

To begin with, it is necessary to clarify that we are dealing here with services offered by the various data management platforms to provide research objects with degrees of alignment with FAIR Principles. These services are distinct in nature from the services offered by IFDS to humans and computing agents, for of researchers and other stakeholders. Therefore, services for FAIRfication can be classified as informational, computational and scientific:

- *INFORMATIONAL SERVICES* **–** They comprise the services offered by information professionals within organizations such as scientific libraries and information centers: persistent identification of research objects and researchers; development of representation structures such as metadata schemas, taxonomy, and ontologies; cataloging and indexing of research objects; data release; disclosure; researchers literacy; development of data collections; support for the elaboration of data management plans; long-term archiving/ preservation; linking/contextualization.

- *COMPUTATIONAL SERVICES* – It comprises availability of software tools and computing resources to support the processing analysis and visualization of research data; recommend how data can best be structured and stored, and work, if necessary, with researchers in structuring databases and text marking (Wilson *et al.*, 2011); these services may also include specific training for the research team in the resources offered and, in more advanced situations, offer high-performance processing and grid computing.

- *SCIENTIFIC SERVICES* - They comprise services that are limited to the scientific environment, such as laboratories, and performed by researchers or data stewards specialists with disciplinary knowledge. These are services related to preparing data for wider uses and may include activities such as evaluation, cleaning, normalization, file organization, appointment and, when necessary, anonymization, and other strategies for preserving privacy, disciplinary indexing; code documentation, workflow and processing, data aggregation. Even considering that these services are carried out by researchers themselves, they need considerable computational support.

The services that support FAIRfication processes towards an Internet of Fair Data & Services, have as focal point some essential concepts for the materialization of their assumptions and reuse. They are: machine-actionability; metadata; and access conditions.

- *FAIR IS ABOUT ACTIONALITY BY MACHINE* **-** "Recognition that computers must be able to access data released autonomously, without the help of human operators, is central for FAIR Principles", categorically state Mons *et al.* (2017, p. 51); therefore, "FAIR principles place a privileged emphasis on improving the potential of machines to find and use data, in addition to supporting their reuse by human beings" confirm Wilkinson *et al.* (2016, p. 1). This is clear when one observes that much of the data life cycle, such as indexing, retrieval via API, processing and reliable analysis of sensory data, are computer-assisted and executed procedures, highlighting the concept of "machine-actionable". In general, this concept assumes a continuum of possible states in which a digital object provides increasingly detailed information to a computational data explorer of autonomous action. The "computational stakeholders", as Wilkinson *et al.* (2016), called them, such as application programs and computational agents, are explorers who act on our behalf – human beings - , performing an increasingly relevant role in data retrieval and analysis. In this constant transitioning context, it is necessary, therefore, to consider that human beings are not the only critical interlocutors in the data ecosystem. FAIR principles are also, and primarily, for machines. Considering the primary limitation of human beings to operate at the scope, scale, and speed required by the level of complexity of contemporary research, especially in the scope of eScience, it is evident the need for machines to be able to act autonomously and appropriately when faced with the broad spectrum of types, formats, protocols, and access mechanisms encountered in exploring the global data ecosystem. "One of the great challenges of intensive data science is, therefore, to improve the discovery of knowledge through the assistance of human beings and their computational agents" (Wilkinson *et al.*, 2016, p. 3). This interlocution is of great importance in retrieval, access, integration and for "the types of deep and broad integrative analyzes that constitute most contemporary eScience" (Wilkinson *et al.*, 2016, p. 3).

These configurations and conditions of current Science have a profound impact on the processes of modern data management platforms, and the full or partial adoption of FAIR principles as part of the backbone of these management-technical systems is an important step towards machine actionability, as it enables them to optimize the use of data resources through appropriate technical implementation choices. For example, the digital resource can be used as an agent or subtract in analyses based on machine learning or artificial intelligence.

Finally, it should be noted that not all data can or should comply with the condition of being automatically processed. There are numerous circumstances that making data machine actionable reduces its usefulness – for example, when adequate tools capable of efficiently processing certain formats are lacking s (Mons *et al.*, 2017).

- *FAIR IS ABOUT METADATA* – Making an essential bridge between machine actionability and metadata, Wilkinson *et al.* (2016) clarify that a resource that lies on a continuum of possible machine-actionable state, provides increasingly detailed information to a computational explorer, and it is applied in two main contexts: first, it refers to contextual metadata that involves the digital object, that is, recognizing the digital object; second, when it refers to the content of the digital object strictly speaking (how to process / integrate it?). In this matter, this information – depending on the quantity, structuring and quality – allows an agent who is faced with a digital object not previously found to identify the kind of object in relation to structure and intention; to identify its utility in the context considered; to determine if it can be used according to its license, consent, level of sensibility or limits of use; and take appropriate actions, similar to what a human would do. Therefore, assisting machines to find and explore data through technology applications and standards at the level of data platforms becomes the main priority of a good data management and highlights the importance of the concept of metadata. The metadata standards have an important role in scientific communication flow, whose emphasis extends the methodological and transparency requirements of the scientific report for data management domain. As such, FAIR Principles emphasize the importance of metadata and its standards in data management, focusing on the concept of "metadata" across its 15 guiding principles. "FAIR Principle's key message is that metadata and metadata standards should be articulated and made publicly available to the greatest extent possible" (Boeckhout; Zielhuis; Bredenoord, 2018, p. 932).

- *FAIR IS ABOUT ACCESS UNDER WELL-DEFINED CONDITIONS* - "FAIR is not the same as open", assertively state Jacobsen *et al*. (2020). The "A" in the context of FAIR is understood "Accessible under well-defined conditions", which makes it different from open without restrictions. Mons *et al.* (2017, p. 51) point out that may be legitimate reasons to shield data and services generated with public funds from indiscriminate access. These types of data include: sensitive personal data, data on geolocations of endangered species, on patentable processes, national security, among others. Furthermore, several sectors, such as industrial and medical, for legal, ethical, contractual or competitiveness reasons need appropriate security for their data and require additional authorization and authentication measures, both for human explorers and for computational agents; in practice, the Internet of Fair Data & Services cannot function without these mechanisms (Jacobsen *et al.*, 2020). Although maintaining primary connections with Open Science, FAIR Principles explicitly and deliberately do not address ethical and moral questions about the degree of openness of data, their availability is entirely at the discretion of the data custodian. FAIR Principles only address the need to describe a process – automatic or manual – for accessing discovered data; a requirement to describe extensively and openly the context in which these data were generated.

The principles do not require FAIR data to be "open" or "free"; however, they do require clarity and transparency about the conditions that govern their access and reuse; they also require that data have an accessible and clear license, preferably machine-readable. "Transparent but controlled access to data and services, rather than the generic and ambiguous concept of "open", allows the participation of a wide range of sectors – public and private - [...] around the world", concluded Mons *et al.* (2017, p. 52).

15.6 "FAIRFICATION" TOWARDS IFDS

The fundamental idea of implementing an Internet of Fair Data & Services is not made real by itself. To this end, it is necessary a multidimensional data management process that can effectively add value, over time, to research objects; the level of adherence of research products to FAIR Principles is linked to the scope and depth of management to which they are subjected. This assumes the need for a multi-layered framework – scientific, technological, informational and governance, as presented in the previous sections, which address the numerous ethical, methodological and organizational problems that interpose between the flows of sharing, integrity, reproducibility, provision of research accounts, as well as the new needs and opportunities for large-scale analysis and reanalysis (Wilkinson *et al.*, 2016, p. 1).

To clarify the meanings embedded in the acronym FAIR, Mons *et al.* (2017) offer a FAIRfication scale – here understood as the level of depth and coverage of management that make digital research adherent to FAIR Principles. During this process, at the lowest level of this scale are objects with no potential for reuse, which correspond to unreleased data, or released in unstable environments such as a web page. These objects do not have **machine-resolvable persistent identifiers** that lead to both data elements and corresponding metadata; these, in turn, are not machine-readable. The minimum path towards FAIRfication is to assign a persistent identifier to a dataset.

However, without a set of **machine-readable metadata,** it will be difficult to find the resource, unless its identifier is known in advance. This indicates that the identifier is necessary but insufficient, and that we need to go further. The next step is assigning metadata, which can have two origins: "intrinsic metadata", which is signaled at the time data is captured, usually by automated processes carried out by the instruments or workflow that generated the data, for example, file format, time stamp and location; metadata marked by the researchers who created/collected the data, information professionals and the stakeholders who reused it in the form of, for instance, annotations, which provide provenance and contextualization to the data and increase its degree of FAIRfication. Therefore, the addition of rich metadata – and also FAIR – is an essential step in this journey. Thus, "the persistent identification and aggregation of metadata already has a profound effect on the reuse potential of research objects, since they can be identified and retrieved" (Mons *et al.,* 2017).

However, even if data is technically FAIR, access may be restricted for clear and fair reasons such as contracts, protection of endangered species, legal and ethical issues; that said, we understand that the maximum standard of FAIRfication should happen When data elements themselves are available under well-defined conditions, for open reuse by any other interested party.

Going even further on the FAIRfication scale, Mons *et al.* (2017) propose that when data are linked to other FAIR research objects we will have reached " FAIR Data Internet"; since an increasing number of applications and services can link and process FAIR data, it can be said that the "Internet of Fair Data & Services" will have been achieved, meaning a "global and shared environment focused on data-driven research and innovation" (Sales *et al.*, 2020, p. 3), where all researchers can access, store, analyze and reuse data for research, innovation and educational purposes. Based on the contours of this territory, an ecology of data activated by associated services is established, which, for the different user segments, translates into a continuum of benefits triggered by computational applications

Just like the current internet, which does not have a centralized governance and is based on a minimal but rigorous set of standards and protocols that support an immense variety of implementations, the concept of "Internet of Data Fair & Services" assumes maximum freedom of development for all interested parties. In this sense, the scalable and transparent routing of DATA, TOOLS, and COMPUTATION – which processes (executes) the tools – is the central feature of a desired Internet of Fair Data & Services, where all types of service providers, public and private, can begin prototyping FAIR data and service applications FAIR (Go Fair, [20--?a]).

As an abstraction, the IFDS models itself in the shape of a three-blade propeller that corresponds to the fundamental elements – data, tools, and computation – that are "routed" to find each other at the right time and place and to be used and reused more efficiently. In this context, tools are mainly defined as software services that act on data, such as virtual machines packaged to travel through the IFDS doing distributed analysis of data or even a data repository and computing as the infrastructure that enables action. As in the hourglass model of the internet, the helix axis corresponds to the minimum set of standards and protocols, as the growth of the ISDF is based on the mantra of the GO FAIR network: "Only a necessary minimum set of protocols and standards to support a wide variety of implementation choices for data, tools and computing elements".

IFDS would run more smoothly if the underlying infrastructure operated on a strong, common, and globally interoperable network and on an engine that efficiently routed data to tools, tools to data, and both to the computation needed, as these three elements are increasingly not residing in large super storage systems and HPC facilities, but are distributed throughout the internet (Go Fair, [20--?a]; Go Fair, [20--?b]).

15.7 FINAL CONSIDERATIONS

Contemporary science, data-intensive by nature, requires data management whose scale goes beyond the most conventional measures, and needs to continually put these assets and other research objects ready for reuse – the ultimate goal of management – by human beings and, above all, by service providers, through computational applications, thus expanding their potential for reuse, repurpose and resignification for various segments, including those outside the world of research. The difficulties of humans operating at the scale and speed required by the complexity of data intensive sciences, especially science, reinforce the need for computational explorers to act autonomously and appropriately in the face of a global data ecosystem.

However, to reach this state of continuous supply, a chain of processes is necessary, ranging from the establishment of policies to a high degree of standardization that requires an infrastructural framework whose density depends on the level and depth of management. But what can be seen is that this effort, sometimes entropic and with diffuse objectives, needs organization and a horizon. The application of FAIR principles realigns these efforts and establishes clear objectives for the management of research objects synthesized in its four fundamental principles, dimensioned by its fifteen guiding principles.

In this complex ecology, the model sought to deconstruct the building blocks that make up a generic architecture to reach a level of FARIfication that allows the achievement of the desired IFDS by articulating the various conceptual modules– guidelines, policies, services, tools, infrastructures etc.,– in the form of pieces that can be adjusted according to the depth, scope, and philosophy of each institution or discipline, thus providing a possible scale to support the measurement of the maturity level of management services projects.

Even considering the general approach of the model, it is necessary to consider that in the implementation of FAIR practices and infrastructures, the specific context of the scientific communities and the possibilities of adoption must be observed. The importance of each principle may depend on the priorities and maturity of the community and the generation and use of certain research objects. This condition implies that different disciplines find technical solutions and need different infrastructural and organizational frameworks and management services to achieve the degree of FAIRification required by their communities. But it should be noted that although scientific imperatives are different between disciplines – which still present different types of organization and culture -, which makes them seek their solutions and follow particular strategies towards FAIR data, the difficulties, and challenges, as well as facilities, are generally shared, as there is a common core of interest. Furthermore, when the scope of FAIR principles is expanded to include other research objects, it is necessary to consider that many of these objects belong to a specific disciplinary domain, which reinforces the finding that FAIR guidelines and practices are also discipline-specific.

REFERENCES

BALL, A. **Review of data management lifecycle models**. Bath, UK: University of Bath, 2012.

BOECKHOUT, M.; ZIELHUIS, G. A.; BREDENOORD, A. L. The FAIR guiding principles for data stewardship: fair enough? **European Journal of Human Genetics**, v. 26, n. 7, p. 931-936, 2018. Available on: https://www.nature.com/articles/s41431-018-0160-0.pdf. Access on: 25 apr. 2024.

BORGMAN, C. **Scholarship in the Digital Age**: Information, Infrastructure, and the Internet. London: The MIT Press, 2007.

COATES, H. L. Building Data Services from the Ground Up: Strategies and Resources. **Journal of eScience Librarianship,** v. 3, n. 1, 2014. Available on : https://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1063&context=jeslib. Access on:25 apr. 2024.

ERWAY, R. *et al*. **Building Blocks:** Laying the Foundation for a Research Data Management Program. Dublin: OCLC, 2016. Available on: https://files.eric.ed.gov/fulltext/ED589141.pdf. Access on: 25 apr. 2024.

GO FAIR. **The internet of FAIR Data & Service**. [20--?a]. Available on: https://www.go-fair.org/resources/internet-fair-data-services/. Access on: 25 apr. 2024.

GO FAIR. **GO FAIR Initiative**. [20--?b]. Available on: https://www.go-fair.org/go-fair-initiative/. Access on: 25 apr. 2024.

GRAAF, M. V. D.; WAAIJERS, L. **A surfboard for riding the wave**: Towards a four country action program on research data. Copenhagen: Knowledge Exchange, 2011.

HONG, N. C. *et al*. **Six recommendation to implementation of FAIR Practices**. Bruxelas: European Commission, 2020. Available on: https://ec.europa.eu/info/publications/six-recommendations-implementation-fair--practice_en. Access on: 25 apr. 2024.

JACOBSEN, A. *et al*. FAIR principles: Interpretations and implementation considerations. **Data Intelligence**, n. 2, p. 10–29, 2020. Available on: emhttp://www.inf.ufes.br/~gguizzardi/102-Annika_Jacobsen-1_GRFHSzW.pdf. Access on: 25 apr. 2024.

KOUPER, I. *et al.* Research Data Services Maturity in Academic Libraries. *In*: JOHNSTON, L. R. (eds.). **Curating Research Data**: Practical Strategies for Your Digital Repository. Chicago: Association of College and Research Libraries, 2017. p. 153-170. Available on: https://experts.illinois.edu/en/publications/research-data-services--maturity-in-academic-libraries. Access on: 25 apr. 2024.

LEONARDI, P. M. Digital materiality? How artifacts without matter, matter. **First Monday**, v. 15, n. 6-7, 2010. Available on: https://journals.uic.edu/ojs/index.php/fm/article/view/3036. Access on: 25 apr. 2024.

MAYERMIK, M. S. *et al*. The data conservancy instance: infrastructure and organizational services for research data curation. **D-Lib Magazine**, v. 18, n. 9-10, Sep./Oct., 2012. Available on: http://www.dlib.org/dlib/september12/mayernik/09mayernik.html. Access on: 25 apr. 2024.

MONS, B. *et al.* Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. **Information Services & Use**, v. 37, n. 1, p. 49-56, 2017.

MUSHI, G. E., PIENAAR, H., VAN DEVENTER, M. Identifying and Implementing Relevant Research Data Management Services for the Library at the University of Dodoma, Tanzania. **Data Science Journal**, v. 19, n. 1, p. 1-9, 2020. Available on: https://datascience.codata.org/articles/10.5334/dsj-2020-001/. Access on: 25 apr. 2024.

NATIONAL RESEARCH COUNCIL. **Preparing the workforce for digital curation**. Washington, D.C.: The National Academies Press, 2015.

PINFIELD, S.; COX, A. M.; SMITH, J. Research data management and libraries: Relationships, activities, drivers and influences. **PLoS One**, v. 9, n. 12, p. e114734, 2014. Available on: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0114734. Access on: 25 apr. 2024.

REED, R. B. Diving into data: Planning a research data management event. **Journal of Escience Librarianship**, v. 4, n. 1, 2015. Available on: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517608/. Access on: 25 apr. 2024.

SALES, L. *et al*. GO FAIR Brazil: a challenge for brazilian data science. **Data Intelligence**, v. 2, n. 1-2, p. 238-245, 2020. Available on: https://direct.mit.edu/dint/article/2/1-2/238/10004/GO-FAIR-Brazil-A-Challenge-for-Brazilian-Data. Access on: 25 apr. 2021.

SAYÃO, L. F. Modelos teóricos em ciência da informação-abstração e método científico. **Ciência da informação**, Brasília, v. 30, n. 1, p. 82-91, 2001. Available on: https://revista.ibict.br/ciinf/article/view/941. Access on: 25 apr. 2024.

SAYÃO, L. F.; SALES, L. F. Afinal, o que é dado de pesquisa? **BIBLOS**, v. 34, n. 2, 2020. Available on: https://www.seer.furg.br/biblos/article/view/11875. Access on: 12 apr. 2024.

SOLOMONIDES, A. Research Data Governance, Roles, and Infrastructure. *In*: RICHESSON, R.; ANDREWS, J. (eds.). **Clinical Research Informatics**. Cham: Springer, 2019. p. 291-310.

STRASSER, C. **Research data management**. Baltimore: NISO, 2015. Available on: https://wiki.lib.sun.ac.za/images/2/24/PrimerRDM-2015-0727.pdf. Access on: 25 apr. 2024.

WILKINSON, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v. 3, n. 1, p. 1-9, 2016. Available on: https://www.nature.com/articles/sdata201618.pdf. Access on: 25 apr. 2024.

WILSON, J. A. J. *et al*. An institutional approach to developing research data management infrastructure. **The International Journal of Digital Curation**, v. 6, n. 2, 2011. Available on: http://ijdc.net/index.php/ijdc/article/view/198. Access on: 25 apr. 2024.

How to cite this chapter: SAYÃO, Luís Fernando; SALES, Luana Farias. An implementation model for the internet of FAIR data & services. *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. 15, p. 177-194. DOI: 10.22477/9788570131959.cap15.

16. GOFAIR BRAZIL HEALTH NURSING NETWORK: WHERE ARE WE, AND WHERE DO WE WANT TO GO?

Eliza Macedo[164]
Patrícia Henning[165]
Maria Simone de Menezes Alencar[166]
Sônia Souza[167]

16.1 INTRODUCTION

Discussions about the production of knowledge and scientific work, as well as reflections on relations among science, technology, data, and information are themes that have been constantly evolving since the last century. Nowadays, these themes have been adopting new configurations in the form of evaluation, management, dissemination, and storage molding themselves to the trends of the digital world, aimed at common, collective and collaborative knowledge.

It is in this context that a new paradigm has been taking place in the scientific field, corroborating open scientific practices, which integrate technological and human actors, aimed at the collective, encouraging the sharing, reuse and digital preservation of data. There is, however, the need to rethink new guidelines and policies that can better meet the demands of this new reality focused on Open Science, considered an international phenomenon, the result of the trends of democratization of knowledge.

Among all Open Science practices, open research data are considered inputs of scientific work that have gained greater prominence and importance today, due to the need and urgency of sharing them, as soon as they are generated, in strategic areas such as health, aiming at their reuse whenever possible. This possibility provides greater speed, transparency, and agility to research, leveraging the production of knowledge and science.

However, due to the complexity of data and. The specificities of each field of knowledge, the need for data contextualization and organization, as well as the detailing of their provenance, increases to ensure their long-term preservation and reuse in other research. It is within this context that FAIR principles emerge, internationally regarded as guiding good practices in managing research data and initiatives aimed at their dissemination and implementation.

[164] Ph.D. in Nursing and Biosciences, Associate Professor IV at the Alfredo Pinto School of Nursing. Federal University of the State of Rio de Janeiro (UNIRIO), eliza.macedo@unirio.br;

[165] PhD in Communication and Health Information (PPGICS/Fiocruz), Visiting Professor at the Graduate Program in Nursing (PPGENF) at the Federal University of the State of Rio de Janeiro (UNIRIO), henningpatricia@gmail.com;

[166] PhD in the area of Management and Technological Innovation at the UFRJ School of Chemistry, Permanent Professor of the Professional Master's Degree in Librarianship (PPGB) and of the Doctorate in Nursing and Biosciences (PPGENFBIO) at the Federal University of the State of Rio de Janeiro (UNIRIO), simone.alencar@unirio.br;

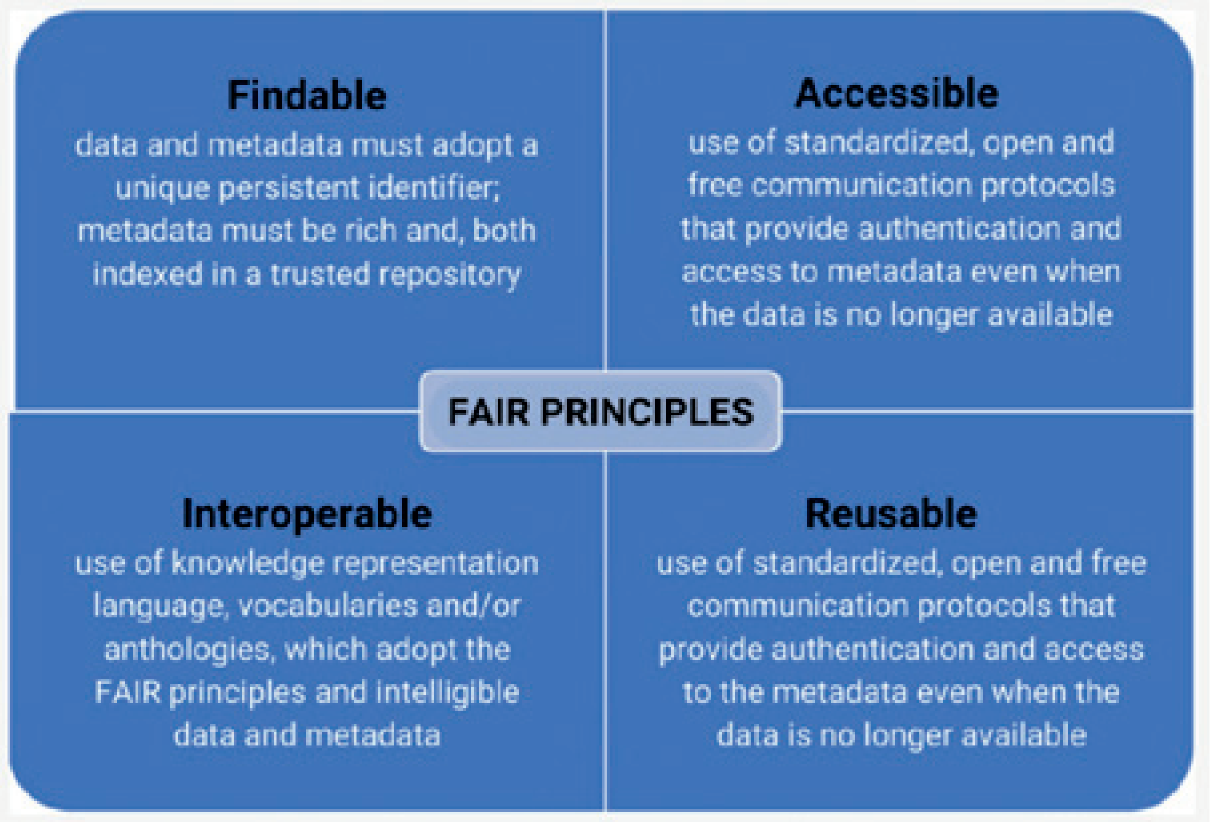
[167] PhD in Nursing. Associate Professor IV at the Alfredo Pinto School of Nursing. Federal University of the State of Rio de Janeiro (UNIRIO), sonia.souza@unirio.br.

This report aims to situate GO FAIR Brazil Health Nursing, describe its implementation trajectory so far, and articulate the mechanisms of action of this international initiative in the Nursing field.   Therefore, a brief introduction of FAIR principles, GO FAIR international initiative, GO FAIR office in Brazil and GO FAIR Brazil Health Network, which constitute the content of the next sections, is necessary.

16.2 FAIR PRINCIPLES AND GO FAIR INTERNATION INITIATIVE

In January 2014, the first manifestations of issues related to data management emerged when a group of experts, scientific editors, representatives from academia, research funding agencies and the industrial field gathered in a workshop entitled *Jointly designing a data FAIRPORT,* at the Lorentz Centre, in Leiden, the Netherlands. This meeting was marked by the high level of discussion around the creation of a global infrastructure that could support the publication, discovery, sharing, and reuse of research data. As a result of this meeting, a set of guidelines for good data management practices called “FAIR Principles” was designed. Such principles are actually an acronym for Findable, Accessible, Interoperable and Reusable (FAIR) and were only officially published in 2016, in *Scientific Data* magazine, in the article by Wilkinson *et al*. (2016) entitled *The FAIR Guiding Principles for scientific data management and stewardship.* Figure 1 briefly presents the FAIR principles.

Figure 1 – FAIR Principles



Findable
data and metadata must adopt a unique persistent identifier; metadata must be rich and, both indexed in a trusted repository

Accessible
use of standardized, open and free communication protocols that provide authentication and access to metadata even when the data is no longer available

FAIR PRINCIPLES

Interoperable
use of knowledge representation language, vocabularies and/or anthologies, which adopt the FAIR principles and intelligible data and metadata

Reusable
use of standardized, open and free communication protocols that provide authentication and access to the metadata even when the data is no longer available

Source: Wilkinson *et al.* (2016), adapted by the authors.

It is possible to observe that these principles, by themselves, do not bring Much clarification regarding their implementation. They only describe a set of desired attributes for good practices in the management and treatment of digital resources. Mons *et al.* (2017, p. 50) clarify that they "deliberately do not specify technical requirements, but rather a set of guidelines for an increasing continuous reuse, through different implementations".

Considering that scientific research is being conducted by increasingly complex data, requiring not Only better treatment and organization, but also greater capacity of machines, appropriate software, better trained human resources and high financial resources to deal with such reality, the *Global Open FAIR* (GO FAIR)[168] emerges in 2017 with the aim of disseminating FAIR principles and services and providing basic guidelines for their implementation in this data science scenario.

This initiative adopts a bottom-up approach as methodology of implementation, that is, it encourages the creation of independent and autonomous networks willfully created by the Scientific Community, following its philosophy and guidelines.

16.3 GO FAIR BRAZIL HEALTH NURSING NETWORK: WHERE ARE WE?

The first GO FAIR Brazil initiative meeting took place on September 25, 2018, in São Paulo, during the 20 years of the *Scientific Electronic Library Online* (SciELO) where the Brazilian Institute for Scientific and Technological Information (IBICT) was established as the coordinator of the GO FAIR office in Brazil. Several stakeholders representing Brazilian universities and research institutes were present, as well as representatives of GO FAIR International (Sales *et al.*, 2020).

However, its official launch to the scientific community took place the following month, December 10, 2018, during an event promoted by the Ministry of Science, Technology, Innovation, and Communication (MCTIC). GO FAIR Brazil[169] is responsible for disseminating, supporting and coordinating the activities for the adoption of strategies to implement FAIR principles, respecting the specificities of the different fields of knowledge, throughout the national territory.

GOFAIR Brazil Health Network[170] is the first Brazilian implementation network, considered today the most active, being responsible for elaborating strategies for the adoption of FAIR principles in health domains. Its coordination is under the responsibility of the Institute of Communications and Scientific and Technological Information in Health (ICICT), of the Oswaldo Cruz Foundation (Fiocruz) and has the participation of several institutions in the fields of Public Health, Sanitary Surveillance, Information, and Communication in Health, History of the Cultural Heritage of Science and Health, Oncology, Nursing, and Professional Education in Health.

In the nursing field, the first steps for the composition of GO FAIR Brazil Health Nursing subnetwork began to be traced, considered one of the branches of the GO FAIR Brazil Health network, in March 2019, When researchers

[168] Available from: https://www.go-fair.org/. Access on: 3 dec. 2020.

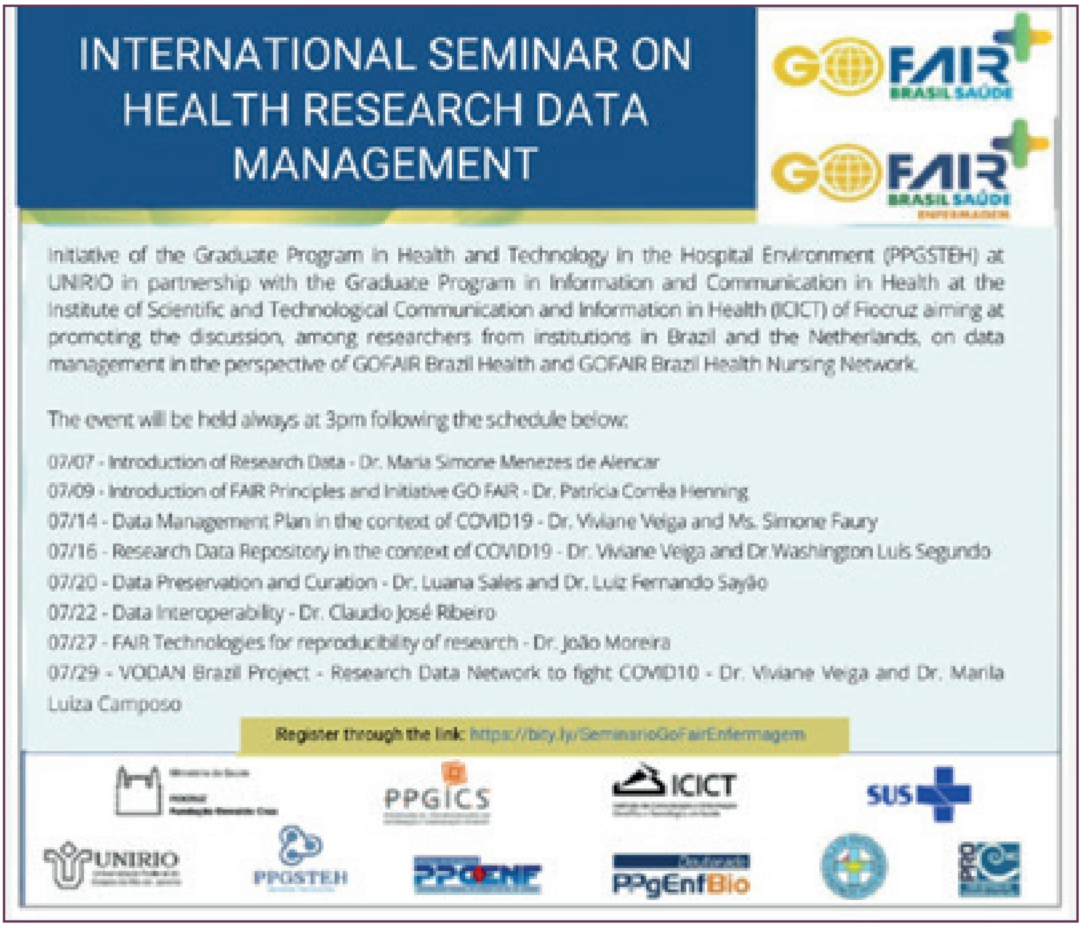
[169] Available from: https://www.go-fair.org/go-fair-initiative/go-fair-offices/go-fair-brazil-office/. Access on: 3 dec. 2020.

[170] Available from: https://portal.fiocruz.br/go-fair-brasil-saude. Access on: 3 dec. 2020.

from Alfredo Pinto Nursing School (EEAP), from the Graduate Program in Health and Technology in the Hospital Environment (PPGSTEH) (professional Master's program), from the Graduate Program in Nursing (PPGENF) (academic Master's program)and from the Graduate Program in Nursing and Biosciences (PPGENFBio) (PhD) met for the first time with representatives of the GO FAIR International initiative, GOFAIR Brazil and GO FAIR Brazil Health Network, to agree on the creation of GO FAIR Brazil Health Nursing Network (Henning, 2019).

Since then, several actions have been developed to strengthen and implement it. First, the "International Seminar on Health Research Data Management - GO FAIR Brazil Health and GO FAIR Brazil Health Nursing", which took place throughout the month of June 2020, was offered to the Brazilian nursing community and was attended by more than 250 participants. The scientific program of this event was conceived using the modality of webinars and delivered by speakers from institutions in Brazil and the Netherlands, with recognized intellectual production and trajectory on the subject.

Figure 2 – International Seminar publicity flyer



INTERNATIONAL SEMINAR ON HEALTH RESEARCH DATA MANAGEMENT

GO FAIR
BRASIL SAÚDE
GO FAIR
BRASIL SAÚDE
ENFERMAGEM

Initiative of the Graduate Program in Health and Technology in the Hospital Environment (PPGSTEH) at UNIRIO in partnership with the Graduate Program in Information and Communication in Health at the Institute of Scientific and Technological Communication and Information in Health (ICICT) of Fiocruz aiming at promoting the discussion, among researchers from institutions in Brazil and the Netherlands, on data management in the perspective of GOFAIR Brazil Health and GOFAIR Brazil Health Nursing Network.

The event will be held always at 3pm following the schedule below:

07/07 - Introduction of Research Data - Dr. Maria Simone Menezes de Alencar
07/09 - Introduction of FAIR Principles and Initiative GO FAIR - Dr. Patrícia Corrêa Henning
07/14 - Data Management Plan in the context of COVID19 - Dr. Viviane Veiga and Ms. Simone Faury
07/16 - Research Data Repository in the context of COVID19 - Dr. Viviane Veiga and Dr.Washington Luís Segundo
07/20 - Data Preservation and Curation - Dr. Luana Sales and Dr. Luiz Fernando Sayão
07/22 - Data Interoperability - Dr. Claudio José Ribeiro
07/27 - FAIR Technologies for reproducibility of research - Dr. João Moreira
07/29 - VODAN Brazil Project - Research Data Network to fight COVID10 - Dr. Viviane Veiga and Dr. Maria Luiza Camposo

Register through the link: https://bity.ly/SeminarioGoFairEnfermagem

Source: UNIRIO (2020).[171]

[171] Available from: https://www.unirio.br/cchs/ppgsteh/eventos/seminario-internacional-sobre-gestao-de-dados-de-pesquisa-em-saude-1. Access on: 3 dec. 2020.

The program content of this Seminar was planned to cover all the practices related to research data management and FAIR principles, distributed in eight modules. Module 1: Introduction to Research Data[172]; Module 2: Introduction to FAIR Principles and the GOFAIR Initiative[173]; Module 3: Data Management Plan in the context of COVID19[174]; Module 4: Research data repositories in the context of COVID19[175]; Module 5: Data Preservation and Curation[176]; Module 6: Data Interoperability[177]; Module 7: FAIR Technologies for research reproducibility[178]; Module 8: VODAN Brazil Project– Research data network to fight COVID19.[179] All videos of the presentations are available at Fiocruz ARCA Repository and on YouTube pages of Alfredo Pinto Nursing School.[180]

It is important to note that GO FAIR Brazil Health Nursing Network is coordinated by the Graduate Program in Health and Technology in the Hospital Environment (PPGSTEH), in collegiate management with Alfredo Pinto Nursing School (EEAP), by the Graduate Program in Nursing and Bioscience (PPGENFBIO) and by the Graduate Program in Nursing (PPGENF), at the Federal University of the State of Rio de Janeiro (UNIRIO).

After collegiate deliberation with the coordinators of the Graduate programs in Nursing at UNIRIO, a group of professors was formed to integrate the programs with the aim of monitoring and implementing the activities for implementing GO FAIR Brazil Health Nursing Network. The group is formed by the following professors: Dr. Eliza Macedo, Dr. Patrícia Henning, Dr. Maria Simone de Menezes Alencar, Dr. Sônia Souza, Dr. Danielle Galdino, Dr. Taís Vernaglia and Dr. Inês Meneses. Among the actions of the group, we can mention: meetings for strategic implementation planning; meeting with a representative of FIOCRUZ for the procedures of the Technical Cooperation agreement; registration of GO FAIR Brazil Health Nursing Research Project at UNIRIO Research Department; elaboration of a project aimed at undergraduates interested in the subject, to act as scholarship holders, already sent to the Dean of Student Affairs, which offers academic incentive; and registration of the actions at the Dean of extension, both at UNIRIO.

Two months after the International Seminar on Health Research Data Management, GO FAIR Brazil Health Nursing Network was launched, on September 22, 2020, during the celebrations of the 130th anniversary of Alfredo Pinto Nursing School, at UNIRIO. The launch was opened with the participation of the coordinators of GO FAIR Brazil, GO FAIR Brazil Health and GO FAIR Brazil Health Nursing, and with the dissemination of their manifesto of adhesion to the Network. A number of 144 professionals with varied profiles attended, from researchers, professors, graduate and undergraduates, librarians, archivists, administrative technicians to health professionals

[172] Available from: https://www.arca.fiocruz.br/handle/icict/45046. Access on: 3 dec. 2020.

[173] Available from: https://www.arca.fiocruz.br/handle/icict/45049. Access on: 3 dec. 2020.

[174] Available from: https://www.arca.fiocruz.br/handle/icict/45050. Access on: 3 dec. 2020.

[175] Available from: https://www.arca.fiocruz.br/handle/icict/45052. Access on: 3 dec. 2020.

[176] Available from: https://www.arca.fiocruz.br/handle/icict/45058. Access on: 3 dec. 2020.

[177] Available from: https://www.arca.fiocruz.br/handle/icict/45059. Access on: 3 dec. 2020.

[178] Available from: https://www.arca.fiocruz.br/handle/icict/45060. Access on: 3 dec. 2020.

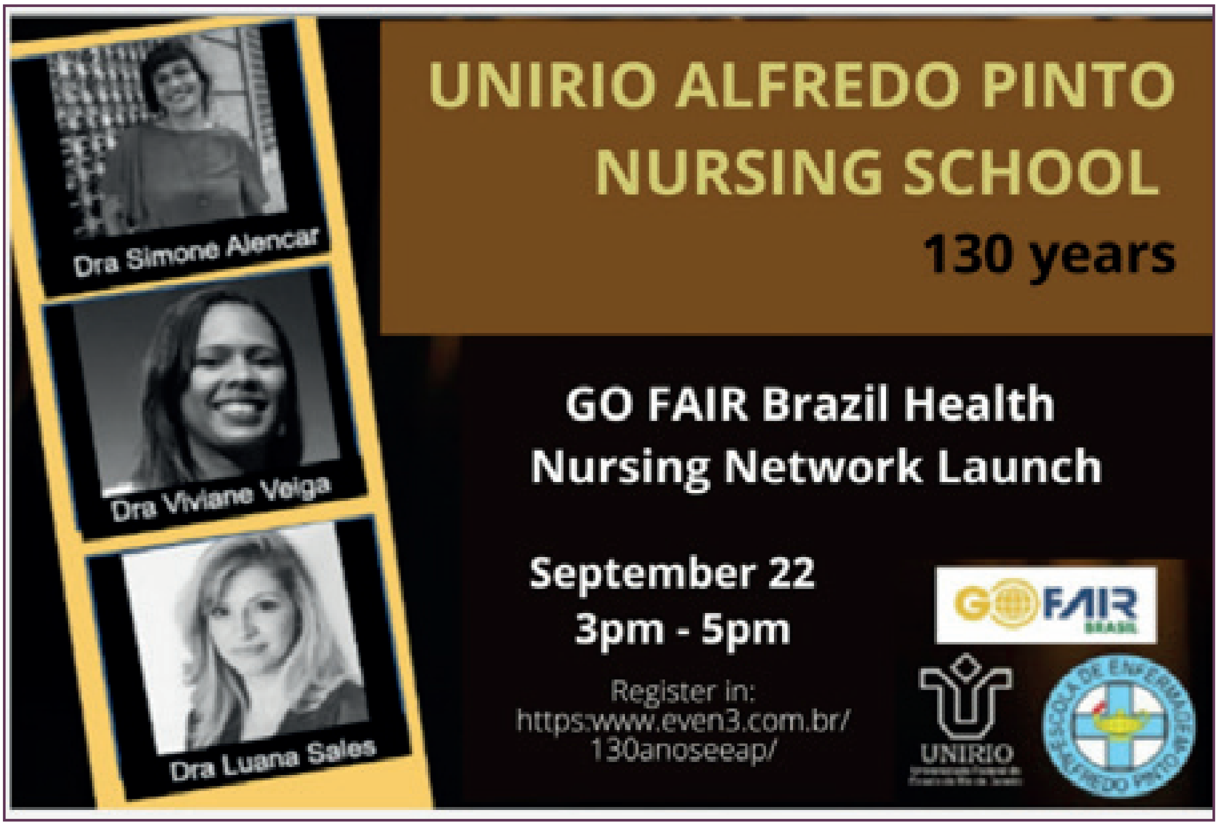
[179] Available from: https://www.arca.fiocruz.br/handle/icict/45061. Access on: 3 dec. 2020.

[180] Available from: https://www.youtube.com/channel/UCH-mOJCskxwnHQweoPkNV-g. Access on: 3 dec. 2020.

from universities, research institutes and hospitals. At the end of the presentations, the participants were invited to fill out a registration form and sign the Open Manifesto for joining GO FAIR Brazil Health Nursing Network [181].

Those interested in participating in the Network only needed to sign the Manifesto and be willing to work collaboratively. The coordinators will contact them and insert those interested in the planning dynamics and future actions of the Network.

Figure 3 – GO FAIR Brazil Health Nursing Network Launch Flyer



UNIRIO ALFREDO PINTO NURSING SCHOOL

130 years

GO FAIR Brazil Health Nursing Network Launch

September 22
3pm - 5pm

Register in:
https:www.even3.com.br/
130anoseeap/

Dra Simone Alencar

Dra Viviane Veiga

Dra Luana Sales

GO FAIR BRASIL
UNIRIO
ESCOLA DE ENFERMAGEM ALFREDO PINTO

Source: Fiocruz (2020).[182]

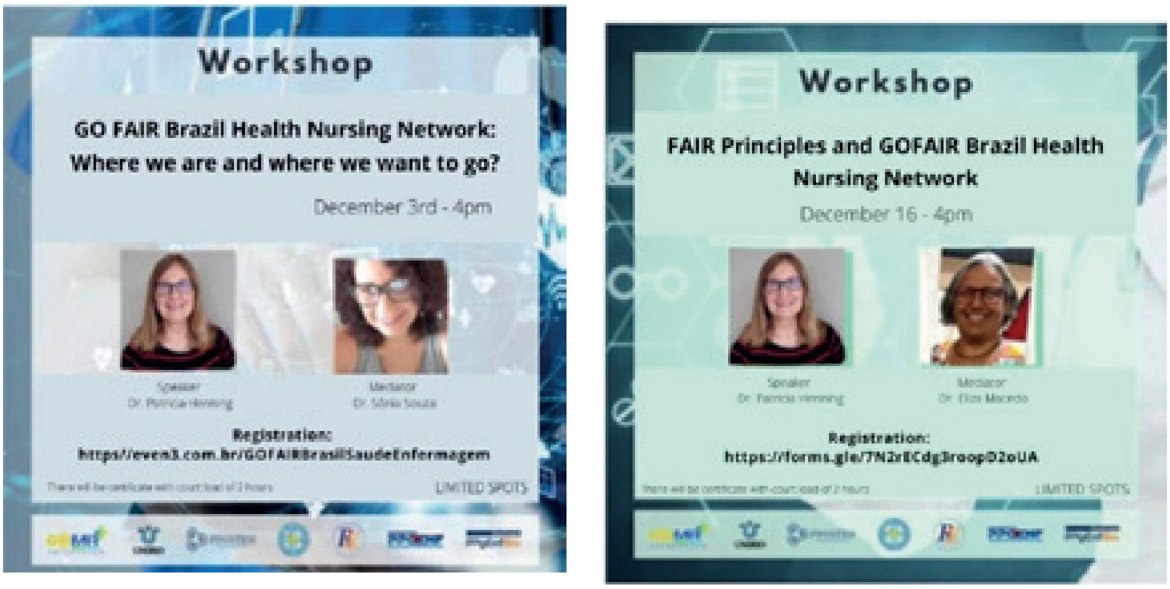
Continuing with the action, the Graduate Program in Health and Technology in the Hospital Environment (PPGSTEH), aiming to seek new adhesions to GO FAIR Brazil Health Nursing Network, offered to the Nursing scientific community, on December 03 and 12, 2020, the workshop "GO FAIR Brazil Nursing Network: where we are and where we want to go"[183]. This event aimed to present to the nursing scientific community and to those interested in working in the health field in general the actions that have been developed within the scope of GO FAIR Brazil Health Nursing Network, aimed at nursing research data management, within the scope of Open Science.

[181] Available from: https://bit.ly/GOFAIRENFERMAGEM. Access on: 3 dec. 2020.

[182] Available from: https://portal.fiocruz.br/noticia/seminario-virtual-marca-o-lancamento-da-rede-go-fair-brasil-saude-enfermagem. Access on: 3 dec. 2020.

[183] Available from: http://www.unirio.br/news/workshop-ira-discutir-gestao-de-dados-de-pesquisa-em-enfermagem. Access on: 3 dec. 2020.

Figure 4 – Promotional flyers of GO FAIR Brazil Health Nursing Workshop



Workshop

GO FAIR Brazil Health Nursing Network:
Where we are and where we want to go?

December 3rd - 4pm

Speaker
Dr. Patricia Henning

Mediator
Dr. Sônia Souza

Registration:
https://even3.com.br/GOFAIRBrasilSaudeEnfermagem

There will be certificate with court load of 2 hours

LIMITED SPOTS

Workshop

FAIR Principles and GOFAIR Brazil Health
Nursing Network

December 16 - 4pm

Speaker
Dr. Patricia Henning

Mediator
Dr. Eliza Macedo

Registration:
https://forms.gle/7N2rECdg3roopD2oUA

There will be certificate with court load of 2 hours

LIMITED SPOTS

Source: UNIRIO (2020).[184]

As described in its Manifesto, this Network proposes to work on strengthening and disseminating FAIR principles in the field of nursing, in an articulated and collaborative way with its members. Currently, the first guidelines for the Network actions are being drawn up, which will be organized through specific work groups with the participation and voluntary action of the Brazilian Nursing community.

16.4 GO FAIR BRAZIL HEALTH NURSING NETWORK: WHERE ARE WE GOING?

GO FAIR Brazil Health Nursing Network seeks its development and consolidation through the strengthening and promotion of management, Nursing research data sharing and reuse, within the Brazilian Graduate Nursing Programs, which are the main generators of Nursing research data. For that purpose, the goals are:

a. Promote research, in the nursing field, focused on specific metadata, standards of technological and semantic interoperability of data such as use of controlled vocabularies and ontologies in the field; data management plan templates; national and international research data repositories, which can reliably store nursing research data; application of use licenses, in accordance with Brazilian regulatory frameworks;

b. Develop methodologies aimed at FAIR products and services practices, which meet disciplinary and operational needs of the nursing field;

[184] Available from: http://www.unirio.br/news/workshop-ira-discutir-gestao-de-dados-de-pesquisa-em-enfermagem. Access on: 3 dec. 2020.

c. Promote meetings, courses, workshops, and seminars aiming at boosting and disseminating FAIR principles among the members of GO FAIR Brazil Health Nursing Network;

d. Create voluntary and collaborative work groups, together with members, who will develop training actions, technical and technological development actions and political actions aimed at expanding the Network;

e. Work in an articulated and collaborative way with GO FAIR Brazil Health Network, together with Fiocruz.

These goals, designed during the meeting of the Network coordination, are medium and long-term, with the possibility of new goals being designed throughout the process as new demands begin to emerge for the sustainability of the Network. Biweekly meetings are being scheduled for the start of activities in 2021 with the aim of bringing members together in the development of activities.

It is considered that there is still a long way to go to achieve the much desired FAIR data management. For this purpose, Graduate Nursing Programs must, in parallel, integrate research data management content into their curricula. In addition, they should also follow suggestions proposed by Raszewski *et al*. (2020, p. 7), who claim that by creating an infrastructure of policies, practices, and curricula that address data management, they will train researchers prepared to meet the data competence expectations of health systems, primary care clinics, in the community and academia.

The interest of Nursing communities in implementing FAIR principles is evident with the large participation of health professionals in various activities developed in a short period of time, such as the International Seminar on Health Data Management and the GO FAIR Brazil Health Nursing Network Workshop. This indicates a promising future in the name of good management of nursing research data, focused on studies, infrastructure development and participation in national and international forums.

16.5 FINAL CONSIDERATIONS

It is known that several types of data are used as research input, from governmental, administrative, private companies, scientific data, as well as those in the health field, where Nursing stands out. In the scope of Open Science, research data must be processed and opened as soon as possible, respecting the disciplinary and legal specificities. In addition to the obstacles inherent in opening research data, such as the lack of understanding of the legal definitions of intellectual property and the lack of standardization in the definitions and configurations of research data, the health field, given its specificities, has one more challenge: the necessary protection of sensitive data, which is a sore point for researchers who are not yet familiar with the subject, requiring adequate software and training.

In this scenario full of doubts and opportunities in Building a new culture that addresses all stages of data life cycle, *GO*FAIR Brazil Health Nursing Network seeks to overcome such obstacles, appropriating the “FAIR Data Management” theme, aiming to develop a set of skills to guide students and researchers from the Graduate Programs in Nursing in the creation of new content with a focus on data and later, develop infrastructure to support the development of data management plans and storage in repositories of appropriate data

By realizing the importance of this scenario that expands and strengthens, the Graduate Program in Health and Technology in the Hospital Environment (PPGSTEH), in collegiate management with Alfredo Pinto Nursing School, the Graduate Program in Nursing (PPGENF) (academic Master's program) and the Graduate Program in Nursing and Bioscience (PPGENFBio) (PhD) begin, through the coordination of GO FAIR Brazil Health Nursing Network, the first steps towards the insertion of research data in nursing in the process of sharing and reusing data in line with FAIR principles, which will promote culture and bring impact and visibility to data produced by the Nursing field, articulating local action with global thinking.

REFERENCES

HENNING, Patrícia. Gestão de Dados de Pesquisa: uma demanda necessária para a geração de novos conhecimentos. **Revista Online de Pesquisa:** Cuidar é Fundamental, Rio de Janeiro, v. 11, n. 3, p. 1-2, 2019. DOI:10.9789/2175-5361. Available from: http://www.seer.unirio.br/index.php/cuidadofundamental/article/view/8939/pdf. Access on: 3 dec. 2020.

MONS, Barend; NEYLON, Cameron; VELTEROP, Jan; DUMONTIER, Michel; SANTOS, Luiz Olavo Bonino da Silva; WILKINSON, Mark. Cloudy, increasingly FAIR: revisiting the FAIR Data guiding principles for the European Open Science Cloud. **Information Services & Use**, [*S. l.*], v. 37, n. 1, p. 49-66, 2017. Available from: https://content.iospress.com/articles/information-services-and-use/isu824. Access on: 3 dec. 2020.

RASZEWSKI, Rebecca; GOBEN, Abigail H.; BERGREN, Martha Dewey; JONES, Krista; RYAN, Catherine; STEFFEN, Alana D.; VONDERHEID, Susan C. A survey of current practices in data management education in nursing doctoral programs. **Journal of Professional Nursing**, [*S. l.*], v. 27, n. 1, p. 155-162, 2020. DOI: https://doi.org/10.1016/j.profnurs.2020.06.003. Available from: https://www.sciencedirect.com/science/article/pii/S8755722320301204?via%3Dihub. Access on: 3 dec. 2020.

SALES, Luana; HENNING, Patrícia; VEIGA, Viviane Veiga; COSTA, Maira Murrieta; SAYÃO, Luís Fernando; SANTOS, Luiz Olavo Bonino da Silva; PIRES, Luís Ferreira. GO FAIR Brazil: A challenge for brazilian data science. **Data Intelligence**, [*S. l.*], v. 2, n. 1-2, p. 238-245, 2020. DOI: https://doi.org/10.1162/dint_a_00046. Available from: https://direct.mit.edu/dint/article/2/1-2/238/10004/GO-FAIR-Brazil-A-Challenge-for-Brazilian-Data. Access on: 3 dec. 2020.

WILKINSON, Mark D; DUMONTIer, Michel; AALBERSBERG, IJsbrand Jan *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, [*S. l.*], v. 3, n. 1, p. 1-9, 2016. Available from: https://www.nature.com/articles/sdata201618. Access on: 3 dec. 2020.

This Project has the support of FAPERJ, through the promotion of the Laboratory of Open Science and Research Data for the support of Innovation - LabINOVA.

How to cite this chapter: MACEDO, Eliza; HENNING, Patrícia; ALENCAR, Maria Simone de Menezes; SOUZA, Sônia. GOFAIR Brazil Health Nursing Network: where are we, and where do we want to go?. *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. 16, p. 195-204. DOI: 10.22477/9788570131959.cap16.

17. VODAN BR - a platform for supporting COVID-19 data following FAIR principles

Maria Luiza Machado Campos[185]
Vania Borges [186]
Giseli Rabello Lopes [187]
Maria Claudia Cavalcanti[188]
João Moreira[189]
Sergio Manuel Serra da Cruz[190]

17.1 INTRODUCTION

The COVID-19 pandemic has made clear the importance of having the results of scientific research more easily available for immediate and wide reuse. Several groups already involved in these themes were mobilized, seeking to discuss and speed up the definition and building of supporting infrastructures capable of facing this challenge. In particular, participants from the Research Data Alliance (RDA)[191], World Data Systems (WDS)[192], GO FAIR network[193] and the Committee on Data (CODATA)[194], linked to the International Science Council (ISC)[195], launched a call for action called Data Together (Data Together, [2020?]), which accelerated and promoted cooperation between different ongoing initiatives.

The implementation network GO FAIR *Virus Outbreak Data Network* (VODAN) was designed to initiate a "community of communities" to quickly design and build an infrastructure for a network of interoperable and shareable international data, and to offer support in the search for evidence-based responses to viral outbreak cases (Go Fair, 2020). As it is an initiative of the GO FAIR consortium, data, and services generated must meet FAIR principles (Mons, 2020), which provide guidelines to make data findable, accessible, interoperable and reusable. In the case of VODAN network, the starting point is clinical data of patients with COVID-19, carrying out, in the first phase,

[185] PhD, Graduation Program in IT– PPGI/UFRJ, mluiza.campos@gmail.com

[186] Doctoral Student, Graduation Program in IT – PPGI/UFRJ, vjborges30@gmail.com

[187] DSc, Graduation Program in IT – PPGI/UFRJ, giseli@dcc.ufrj.br

[188] DSc,Military Institute of Engineering– IME, yoko@ime.eb.br

[189] PhD, University of Twente - UTwente, j.luizrebelomoreira@utwente.nl

[190] DSc, Graduation Program in IT – PPGI/UFRJ, serra@ppgi.ufrj.br

[191] Available on: https://www.rd-alliance.org/. Access on: 20 Sept. 2024.

[192] Available on: https://www.worlddatasystem.org/. Access on: 20 Sept. 2024.

[193] GO FAIR – initiative that encourages the availability of FAIR data and services (Findable, Accessible, Interoperable and Reusable) for scientific research projects. Available on: https://www.go-fair.org. Access on: 20 Sept. 2024.

[194] Available on: https://codata.org. Access on: 20 Sept. 2024.

[195] Available on: https://council.science/. Access on: 20 Sept. 2024.

the transformation and treatment of these data, according to the Clinic Research Form (CRF), developed and standardized by the World Health Organization (WHO).

The CRF-WHO is a clinical research protocol developed with the help of specialists to obtain relevant information in epidemic and pandemic cases. This form is dived in three modules: the first aims at collecting patient admission data; the second aims at follow-up data during hospitalization; and the third aims at treatment outcome data, whether by discharge, hospital transfer or death.

According to Manifesto Vodan (2020), the original proposal consists in developing a solution that allows health professionals to record data observed in the format established by CRF-WHO, storing them in repositories or data banks. Later, metadata of these repositories must be available in a FAIR Data Point (FAIR DP). A FAIR DP is a component of the FAIR data support infrastructure through which software agents can have access to descriptors that allow them to find and visit data locally and execute queries on them (Mons, 2020). Local data curator will give the permission or not for the query/analysis to be performed. This structure allows that the patient's information to remain protected in databases of health facilities, respecting the legislation for health data in each country.

In Brazil, VODAN BR project[196] began concurrently with the advance of the pandemic in the country, during the first months of 2020, as part of GO FAIR Brazil Health[197], linked to Oswaldo Cruz Foundation (Fiocruz), in multi-institutional partnership with the Federal University of Rio de Janeiro (UFRJ) and the Federal University of the State of Rio de Janeiro (UniRio), among other institutions. The development of the infrastructure is under the responsibility of GRECO Research Group[198], from UFRJ, and its pilot test partners are Gaffreé Guinle Federal Hospital in Rio de Janeiro, and São José Municipal Hospital, in Duque de Caxias. Data are collected from their original systems and treated to be in line with the standard established, that is, the WHO questionnaire format, aiming at their later availability as well as their metadata, meeting FAIR principles, Semantic Web standards and following licensing and anonymization criteria established.

This chapter aims to present an overview of computational assets being developed to support VODAN BR project. This scalable, distributed and generic infrastructure aims to meet an intensive data collection with high heterogeneity, making them available in platforms that offer data and metadata interoperable and processable by software agents, supporting the discovery of other resources that can be associated with them. Thus, it is possible to obtain greater agility in the discovery and generation of knowledge from a more effective reuse of research results.

The next sections are structured as follows: section 2 addresses the VODAN implementation network and the technologies it supports; section 3 presents VODAN BR platform, describing the process and infrastructure being developed; and section 4 presents the conclusion and future possibilities identified.

[196] Available on: https://vodanbr.github.io/. Access on: 20 Sept. 2024.

[197] Available on: https://portal.fiocruz.br/go-fair-brasil-saude. Access on: 20 Sept. 2024.

[198] Available on: http://dgp.cnpq.br/dgp/espelhogrupo/634046. Access on: 20 Sept. 2024.

17.2 VODAN IMPLEMENTATION NETWORK AND ASSOCIATED TECHNOLOGIES

Although the volume of information available on the Web since the beginning of the pandemic has grown far above expectations, it is observed that, for the most part, they refer to total people infected, hospitalized, recovered and deaths. In addition to these aggregated data, data referring to clinical picture, the treatment of patients and their outcome constitute an important support for more detailed studies in clinical research. However, in general, it is observed that despite being extremely valuable to the scientific community, these data are not generally accessible.

Two main problems immediately emerge from dealing with data at this level of detail. The first is the confidentiality of medical records, which can be circumvented by providing anonymized and structured data to meet the demands of clinical investigations or specifically defined licensing. Another problem, more technical and more difficult to solve, resides in the use, by a large part of the Hospital Units (HU), of software for electronic medical records without greater structuring for data entry, with many free text fields, which make analysis and later extraction difficult.

Added to these problems are the challenges of developing and implementing an infrastructure that supports FAIR data release. Although the proposal of the principles is already some Years old, providing a conceptual basis and guidelines that quickly became popular, there are still few technological alternatives that have already been tried together. Certainly, the use of Semantic Web approaches and standards constitutes a solid contribution to the solutions being prospected and developed, but complementary mechanisms and technologies are still necessary. The next two subsections describe VODAN network in more detail, as well as some main technological resources and solutions that support it.

17.2.1 VODAN Implementation Network

VODAN implementation network emerged in early 2020, as a joint effort to implement, experiment and expedite solutions (some already being independently tried in other domains), to support FAIR data release and exploration in the context of research associated with COVID-19 and to other future viral outbreaks. The network proposes an effort for the so-called FAIRfication[199] of COVID-19 data, even after the fact, employing the CRF-WHO model to establish the standardization of information (Satti *et al*., 2020). The Data FAIRfication process promotes the application of FAIR principles to data and metadata, as well as to the infrastructure that supports them. For that purpose, in general, the process contemplates two stages of: (i) collection of non-FAIR data; (ii) analysis of data collected; (iii) definition of a semantic model for the dataset that allows describing the meaning of entities and their relations; with accuracy and with no ambiguity; (iv) definition of metadata associated to data collected, including, among others, provenance, distribution and location, types of access; (v) treatment to make data potentially interlinkable with other sources, with assignment of persistent identifiers, annotation based on controlled vocabularies and/or ontologies, employing technologies and standards of Semantic Web and Linked Data (Heath; Bizer, 2011); (vi) definition of metadata associated to data (and their treatment so they are also FAIR); data and their metadata release.

[199] FAIRification of data – process for turning non-FAIR data in FAIR data. Available on: https://www.go-fair.org/fair-principles/fairification-process/. Access on: 20 Sept. 2024.

After the FAIRfication process, a set of data and metadata adhering to the FAIR principles is obtained. This well-structured data and metadata can be explored through mechanisms that use machine learning techniques and other artificial intelligence (AI) approaches to discover significant patterns in epidemic outbreaks, supporting decisions and actions to face them. As presented in (Satti *et al*., 2020), it is vital to ensure that the data, metadata, and vocabularies used are FAIR, in the original sense of the acronym, but also in the sense of "*Federated, AI- Ready*", that is, federated data for AI.

At the end of the developments associated with the VODAN initiative, the establishment of a federated network of epidemiological FAIR DPs is expected, promoting FAIR services and data, accessible to researchers, for studies on the COVID-19 pandemic and other epidemics that may arise in the future.

VODAN Africa&Asia network[200] was the first initiative of implementation of VODAN network. It is funded by the Philips Foundation[201] and aims at promoting the access distributed to CRF data from Africa and Asia, to support the fight against the COVID-19 pandemic, assisting Universities and Hospitals in Uganda, Ethiopia, Nigeria, Kenya, Tunisia and Zimbabwe, among other countries. This initiative directed its activities towards training researchers and data designers in the creation of FAIR DPs. The trainings guided the participants on the FAIR principles and on the process of building the FAIR DPs, ensuring that data and metadata released are linked and can be available and processed by software agents. As a result, on July 22, 2020[202], the world's first FAIR DP was made available in Uganda. Since then, other FAIR DPs have been activated, based on data and metadata from partners HUs.

Differently from VODAN Africa&Asia network, VODAN BR project opted for a smaller scope, aiming to develop a supporting environment initially focused on data from two partner hospitals, adjusting them to CRF -WHO and creating a computational infrastructure for its dissemination through a first FAIR DP in Brazil. Subsequently, other hospitals will be contemplated, which can already make use of the results of the pilot experience conducted, and the infrastructure developed and tested.

17.2.2 Semantic Web and FAIR Principles

The Semantic Web proposes that data on the Web is defined and connected in a way to be interpreted by both human beings and machines, promoting their sharing and reusing by applications, companies, and community. For that purpose, the proposal of representation of connected data establishes a set of standards and best practices for releasing and interconnecting data structured on the Web, based on data annotation in controlled vocabularies and ontologies, making the identification of new connections among items from different sources easier, aiming to form a global data space, the so-called Data Web (Heath; Bizer, 2011).

[200] Available on: https://www.vodan-totafrica.info/. Access on: 20 Sept. 2024.

[201] Available on: http://www.digitaljournal.com/pr/4626217. Access on: 20 Sept. 2024.

[202] Available on: https://kiu.ac.ug/special-news-page.php?i=covid-19-computer-readable-observational-data-installed-at--kampala-international-university_1595432235. Access on: 20 Sept. 2024.

FAIR principles, initially aimed at managing research data, add to what is already proposed for the Semantic Web, with the objective of making digital objects findable, accessible, interoperable and reusable. In essence, these principles add to the standards established by W3C[203] the importance of using metadata to facilitate the discovery and understanding of data, especially by machines (software agents). It should be noted that FAIR principles do not establish standards or supporting technologies, but rather guide the creation of FAIR data and metadata.

Metadata standards and content annotations are established to promote the common understanding of data meaning, guaranteeing the right interpretation and its adequate use. In order for this metadata to be machine-interpretable, they need to be findable and structured. Machine-actionable metadata, essential to FAIR principles, has led members of GO FAIR and RDA, start in 2018, to foster discussion on *Metadata for Machine* (M4M), in a series of events[204] to assess the state of the art and encourage the creation and reuse of metadata components and metadata templates for machine processing. In VODAN implementation, M4M has been involved in the standardization of metadata referring to catalogs and datasets that will be made available via FAIR DPs, as well as in a series of services associated with them.

In any case, it is not trivial to unequivocally explain a shared semantics about these digital assets, and ontologies play a fundamental role in this. Currently, ontologies are considered in areas of computing (Studer; Benjamins; Fensel, 1998), two of which are: (i) in the area of conceptual modeling, where, through the process of ontological analysis, models well-grounded on top-level ontologies are built; and (ii) in the area of Web Semantic, where both lightweight ontologies, in the line of vocabularies, taxonomies and thesauri, as well as robust ontologies are used, preferably following well-founded models and represented in expressive languages, which can be explored by inference mechanisms, to generate more knowledge.

Considering the definition of ontology as *"... a formal and explicit specification of a shared conceptualization"* (Santos, [2020?]), it follows that : an ontology is considered formal because it is machine interpretable; it is explicit because it presents specifications of concepts, properties, relations, functions, restrictions, and axioms very well-defined; it is a conceptualization for defining and abstract model and a vision of a phenomenon of the world that one wants to represent; and it is shared because it is consensual knowledge among those who work with the domain or applications in question.

The approach to ensure formalism and flexibility for the creation and availability of data and metadata uses RDF (Resource Description Framework) language[205], including RDFS (RDF Schema) in this context [206]. The OWL (Web Ontology Language) language[207], developed for the creation of robust ontologies, also uses this pattern.

[203] Available on: W3C Semantic Web Activity https://www.w3.org/2013/data/. Access on: 20 Sept. 2024.

[204] Available on: https://www.go-fair.org/resources/go-fair-workshop-series/metadata-for-machines-workshops/. Access on: 20 Sept. 2024.

[205] Available on: https://www.w3.org/wiki/RDF. Access on: 20 Sept. 2024.

[206] Available on: https://www.w3.org/TR/rdf-schema/. Access on: 20 Sept. 2024.

[207] Available on: https://www.w3.org/OWL/. Access on: 20 Sept. 2024.

The formalism of the RDF specification is associated with the structural pattern used to describe and store data. This pattern is defined by triples consisting of the following elements: <*subject*> <*predicate*> <*object*>. Each triple constitutes a declaration, the basic unit of RDF, it is a set of declarations that describe a web resource. Each resource, in turn, has a unique identifier called Universal Resource Identifier (URI). The Uniform Resource Locators(URLs) associated to URIs are dereferenced, that is, they can be accessed through browsers, providing information about the resource. This unique identifier allows the reuse of resources between different data sources, streamlining implementations, providing interoperability and facilitating integrations

By describing data and its metadata, RDF allows flexibility in the construction and evolution of schemas not available in the usually used Database Management System (DBMS), such as those based on relational technologies. The set of statements represented by RDF constitute an RDF Knowledge Graph.

17.3 VODAN BR PROJECT AND THE PERSPECTIVE OF FAIR DATA AND METADATA MANAGEMENT

The VODAN BR project established a set of premises to be respected during its implementation phases. These premises guide the activities related to data and metadata management, aiming to establish a structure capable of being quickly adjusted, which significantly reduces the need for changes in applications/tools with each evolution and version of CRF or of the terminological instruments of reference. Among the established premises, it is worth mentioning:

- create an infrastructure capable of implementing and making available a digital CRF (application), centered on users of health services, which is capable of responding to epidemic episodes of this pandemic;

- store information established in the CRF-WHO, anonymously, considering possible versions of the current CRF for inclusion, alteration, or exclusion of form elements;

- enable the creation of CRFs or the inclusion of specific additional questions. This need was presented considering the different types of survey forms used in Brazil, which, in addition to the elements established by the CRF-WHO, are concerned with specific information, relevant to research in the country, such as, for example, participation in campaigns of vaccination and date of last dose;

- promote a conceptual modeling that allows the alignment of the elements of the forms to the ontologies (semantic models), helping the process of data FAIRfication;

- provide a flexible infrastructure, modular, scalable and agile infrastructure, to support software and database adaptations;

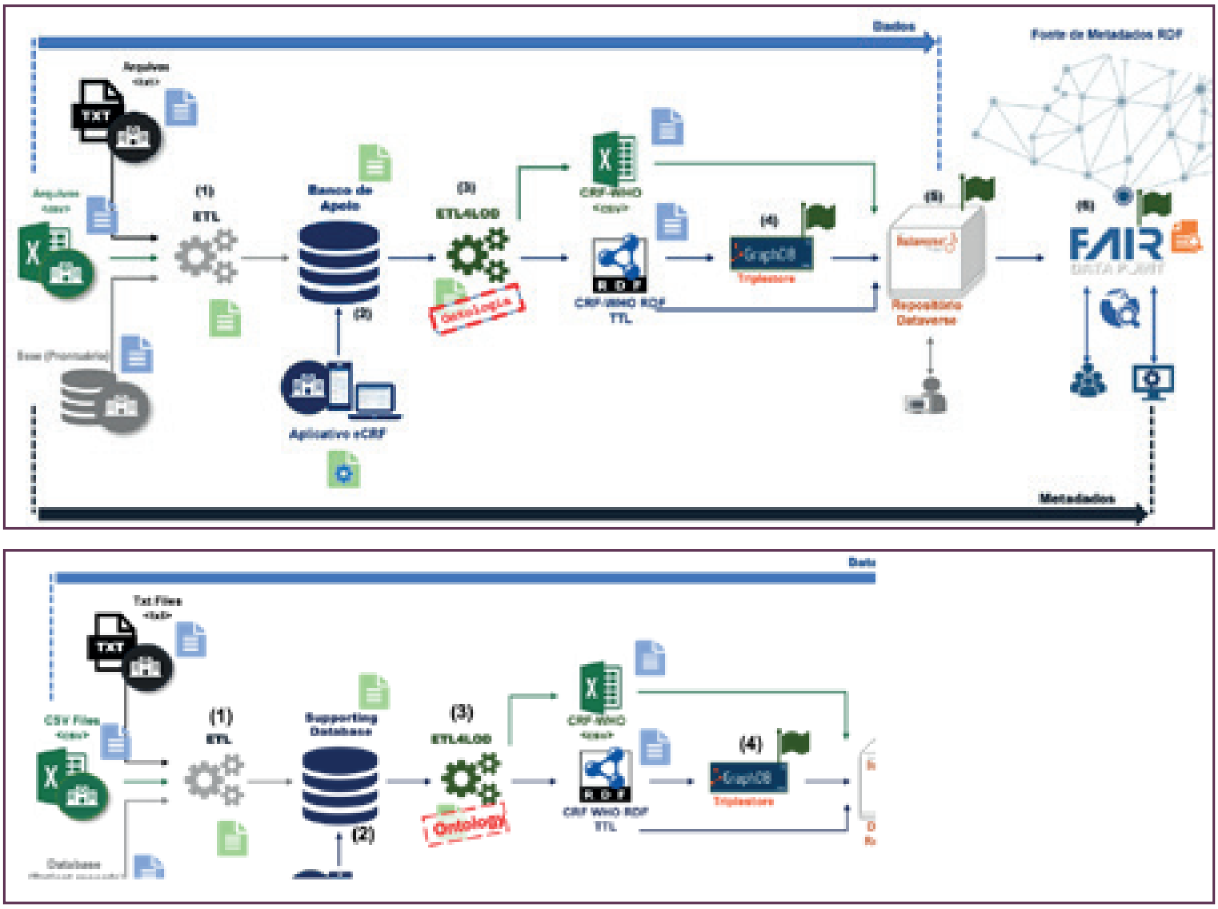
- transform the collected data, that is, “non-FAIR data” into linked data, mapping them to machine-readable formats, using RDF, making them available in datasets and releasing their metadata, also in RDF, in a FAIR DP;

- publicly make available a FAIR DP set to meet the conditions agreed with the participants, enabling access to data through controlled queries and not through traditional downloads.

Respecting these premises, a platform was designed for data processing and services that range from the availability of clinical research data by HUs to metadata release FAIR DP. The platform, represented in Figure 1, has as main requirements to be modular, distributed, scalable, and flexible. Modular, because the planned activities are organized in the form of modules that interact in a chained way, with the result of a module the input of the subsequent module. Scalable and distributed because the idea is that a supporting database is made available in each HU, as well as triple store repositories and/or databases will host data structured according to CRF-WHO in its different distributions or formats. In this way, as more hospital participate of the Project, more computational infrastructure will be added, causing a natural horizontal scaling. In addition, it is a flexible platform, as the heterogeneous data produced by the HUs are treated and transformed into a RDF graph representation, which is one of the formats that facilitates data interconnection.

Initially, as shown in Figure 1, (1) the platform captures data that can be in different formats, such as txt, cvs, or even in the format used in each HU, and via an Extraction-Transformation-Loading (ETL), performs the debugging and transformation of data, storing them in a supporting database (2) which can also directly receive data through a mobile application (eCRF) specially developed. The data stored in the supporting database then undergoes a transformation to connected data, (3), being annotated in vocabularies and ontologies, to meet the interoperability principle. They are then loaded into a graph database (4), in the role of a triple store, or, in the form of an RDF dataset, made available for download in a repository (5). The associated metadata also undergoes a processing and transformation process (3) being loaded and made available in a FAIR DP (6).

Figure 1 – Representation of VODAN BR Platform



Source: Designed by authors.

As established in VODAN network, the datasets must be "visited" by algorithms, respecting the access established by HUs. The metadata associated, contemplating, for example, information on the origin of existing data, types of distribution and the access policies, will be available and accessible in FAIR DP.

Of the elements that make up the platform, the following can be distinguished, due to their relevance in the project and the attention and challenges in the treatment of data: (i) the mechanisms for capturing data, contemplating different requirements and systems of the HUs; (ii) the supporting database, responsible for storing data from these heterogeneous data sources; (iii) the tool to support treatment, transformation, and annotation for inter-connected data and metadata; (iv) alternatives for releasing data; and (v) the creation and feeding of FAIR DP, part of the international VODAN general access point federation.

The tasks performed and the technological choices for these 5 elements of the VODAN BR platform are described in the following subsections.

17.3.1 Data collection

The project envisages three different ways of collecting data from clinical trials of patients with COVID-19:

- by using the application (eCRF) created for recording information, according to the CRF – WHO;
- through an ETL tool for anonymized data uploads from files in txt or csv formats made available by HUs;
- through ETL processes connecting data bank to data bank, with the purpose of transferring information from the patient records to supporting data banks, in the format established by CRF-WHO.

The collection from existing digital records represents an additional challenge. Despite the use of the supporting database by HUs, as a transition bank for CRF-WHO format, and all the facilities it offers, one of the main problems in the analysis and extraction of clinical data for research stems from the flexibility of existing medical records systems that enable textual fields for recording certain aspects of the treatment. The lack of standardization in this record and the large volume of these unstructured data (which includes each procedure performed on the patient, including medications and lab tests), make the collection and transformation process difficult, requiring the support of a health professional for its interpretation and recording. This problem is not new and has been a constant in studies on the interoperability of health treatment data (Santos, [2020?]; Cruz; Campos; Mattoso, 2009). Another important aspect considered was the diversity of information on data provenance (Cruz; Campos; Mattoso, 2009) to be managed.

17.3.2 Creation and Maintenance of Supporting Data banks

Due to the heterogeneity of the data sources and the data per se, it was decided to develop a supporting base for the treatment and formatting of data, aiming to adapt them to the CRF-WHO structure and to support and accelerate the transformation process for connected data.

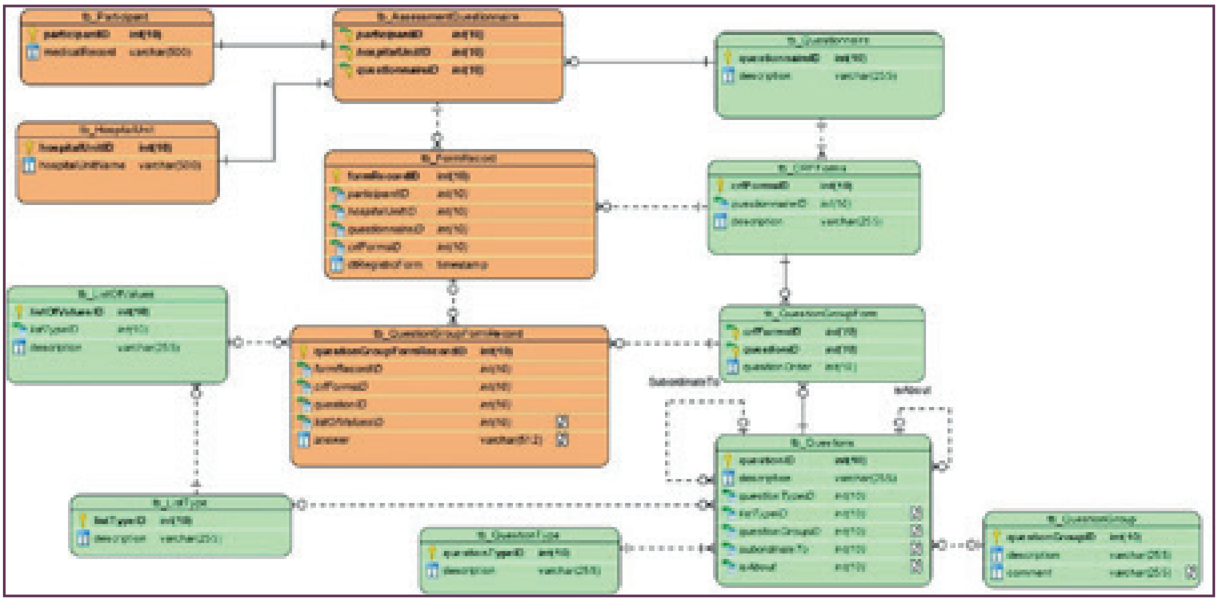
We emphasize that, although part of the data sources come from the medical records of patients with similar systems to a certain extent, it was chosen to follow a model that adheres to the questionnaire for data collection from the CRF-WHO. This decision was critical, as it allowed the establishment of a structure of questions and answers associated with the forms/modules oriented to the service and, consequently, to the collection of information. The form represents a survey, in our case a clinical data survey. It consists of a set of questions, grouped into well-defined categories, collected by a health agent, in this case at a HU, considering observations made about an element of interest, the patient. This research requires a spatio-temporal view, having for that purpose: an admission module, destined to the questions currently the patient arrives; a follow-up module, intended for questions about the patient during hospitalization; and an outcome module, with an overview of the treatment provided and the patient's final situation.

The use of questions with mostly standardized answers helps the discovery of vocabularies and the adoption of Semantic Web, such as ontologies that can be used to define a semantic model.

Another important aspect refers to the combination of the hierarchical structure Module/Group/Question and Subordinate Question that embeds an organization of knowledge by categories, allowing to define different views for analysis. An example of a possible analysis would be the evaluation of cases (from admission to outcome), considering the comorbidities identified at the moment of admission and the medications administered during hospitalization. The result of this analysis could help in the process of guiding drugs indicated or not in a treatment, given the patient's comorbidity.

For the modeling and implementation of this supporting database, it was decided to use a modeling based on relational technology, due to the ease of maintenance and interaction with the mechanisms and applications for data uploading and manipulation. A partial view of this database schema is presented in Figure 2, where the entities in green represent the hierarchical structure of the CRF-WHO and those in orange represent the record of patient information collected by hospital units.

Figure 2 – Extract of the Logical Model of the supporting data bank– CRF-WHO View



Source: Modelled by authors.

In this model, information on CRF-WHO and its versions are registered in the *tb Questionnaire* table; the three modules associated to CRF are registered in the *tb_CRFForms* table; the questions are stored in the *tb_Questions* table, where the type of question is associated (*tb_QuestionType*), the group to which it belongs (*tb_QuestionGroup*), if any, and the type of list of standardized answers (*tb_ListType*), in case the question demands a standardized answer.

The information of the research participants (patients) is entered in the *tb_Participant* table and the information of the Hospital Unit in the *tb_HospitalUnit* table. The table *tb_AssessmentQuestionnaire* records the opening of a CRF-WHO for a participant. For each record in this table, the modules enabled for the patient are launched in the *tb_FormRecord* table. The table *tb_QuestionGroupFormRecord* stores, per module, the questions and answers obtained from the evaluation of each patient, considering textual and standardized answers (*tb_ListOfValues*).

17.3.3 Transformation for Connected Data

After the treatment and the upload of data from each source (HU) in its supporting data bank, the next step refers to a process of transformation for a semantic model, following the paradigm of connected data.

For the VODAN project, the FAIR Data Team made available a model of semantic data that represents the fast version of the CRF-WHO. This semantic model (or ontology) was denominated WHO-COVID-CRF[208] and, in addition to the representation of CRF-WHO, it established a set of entities in the health field domain, to which the questions in the form are related. These entities refer to other existing and well-documented ontologies, providing quality and additional information to guide the users in filling out the form.

As it was developed oriented to the CFR-WHO form, the analysis of this semantic model identified a series of similarities that made it possible to extend the modeling of the supporting database, including ontology information referring to identification, structuring and valuation, to speed up the data FAIRfication process.

The structuring of WHO-COVID-CRF ontology allowed the use of its information for the initial upload of the tables that represent the questionnaire, with very few adjustments. Through this upload, the alignment of the information in the tables referring to the questionnaire with the ontology was implemented, allowing the creation, by the data administrator, of views that present the questionnaire and its ontological information, as well as views that help the stage transformation to linked data performed later.

In order to make data connected, the tool ETL4LOD[209] was used. This tool was initially developed through a partnership between URFJ and UFES universities, in the *LinkedDataBR*[210] Project, aiming at building an infrastructure to support open data release using Semantic Web standards and technologies. An ETL4LOD consists in a set of plugins, developed in JAVA, which extends the *Pentaho Data Integration* functionalities, an ETL tool widely used, providing transformation of data from different sources for connected data.

In the same way that the tool has been adapted for the treatment of data, it also includes the treatment of metadata, to support the FAIRfication process as a whole.

It should be noted that the adopted modeling allows ontologies of interest that may arise to be incorporated into the database, serving to make additional annotations that will contribute to reducing the ambiguity of the data and metadata treated.

[208] Available on: WHO-COVID-CRF: https://github.com/FAIRDataTeam/WHO-COVID-CRF. Access on: 20 Sept. 2024.

[209] Available on: ETL4LOD: available in https://github.com/johncurcio/ETL4LODPlus. Access on: 20 Sept. 2024.

[210] Available on: https://memoria.rnp.br/_arquivo/gt/2010/GT-LinkedDataBR_fase1.pdf. Access on: 20 Sept. 2024.

17.3.4 Data release

Following VODAN network guidelines, survey data must be made available in the connected data format, using RDF standard. Following trends in the research data management and its availability in institutional or thematic repositories, one of our alternatives for data release was the use of a repository platform. In VODAN BR, we chose Dataverse[211], as it is the platform previously selected by the coordinating institution of GO FAIR Health Brazil, Oswaldo Cruz Foundation, for releasing research data.

Data verse is an open-code data repository, developed by the Institute of Quantitative Social Sciences of Havard (IQSS), to store, share, release, cite, explore and analyze research data. The repository hosts several virtual archives called data verses. Each data verse contains sets of datasets, and each dataset contains metadata and descriptive data archives (including documentation and ode that accompany data). As a method of organization, a data verse can also contain other data verses.

To increase the reuse of data, in addition to datasets in the RDF standard, the project established other two formats of distribution: the first, in a triple store supported by a DBMS graph using the GraphDB[212] tool, and the second, in .csv format, for a more traditional use of data.

GraphDB is a DBMS for data banks in graph structure, also used as triple store RDF, which provides an agile structure for release and consumption of connected data. This consumption is performed through SPARQL[213] language (SPARQL Protocol and RDF Query Language), a language for semantic queries with a protocol for accessing data in RDF. Therefore, in an initial proposal, each participating hospital can have their data available in different formats and platforms of distribution, according to their convenience and licensing it defines.

17.3.5 Publication in FAIR DP VODAN BR

As previously mentioned, a FAIR DP is an infrastructure to store and access data that aims to: (i) allow data holders to expose their datasets in accordance with FAIR principles; (ii) make discovery of information on FAIR DP easier by data consumers, in a FAIR DPs network; (iii) establish mechanisms that manage consumers' access, according to licenses and restrictions imposed to data by its managers; (iv) provide data holders with indicators of access on (meta)data available; and (v) provide interaction of data for humans through the Graphical User Interface – GUI, and for software agents, using Application Programming Interface – API (Santos *et al.*, 2016).

To promote standardization for the VODAN FAIR DPs, the reference metadata was established and structures in the RDF model by the FAIR Data Team. This standardization defines a set of rich metadata that describe infor-

[211] Available on: The Dataverse Project – Available in https://dataverse.org. Access on: 20 Sept. 2024.

[212] Available on: GraphDB. Available in: http://graphdb.ontotext.com/. Access on: 20 Sept. 2024.

[213] Available on: SPARQL 1.1 Query Language. https://www.w3.org/TR/sparql11-query/. Access on: 20 Sept. 2024.

mation such as, for example, structural and internal coherence of data, licenses and reference sources, access conditions, context, and provenance (Santos, [2020?]).

Another important aspect considered was the diversity of information from data sources (Cruz *et al*., 2020) to be collected, managed and made available in FAIR DP. In short, data provenance plays a significant role in scientific or even commercial projects. It can be defined as a historical documentation of an artifact (object, data, or dataset) generated by an agent-driven procedure (person, process or computational system). It enables scholars to understand and be able to assess the importance and context of the creation, application, or reuse of that artifact more accurately. Provenance is a type of metadata that enhances the quality assurance and veracity of data or datasets. It assists in managing project data as well as supporting reproducibility and reliability.

The provenance of data in the VODAN BR project could be useful for researchers and health professionals who seek to understand the effects of the pandemic. For example, provenance information can be incorporated at the record level by assigning descriptors as part of the data transformation process (e.g., whether a diagnosis was entered by a physician or derived from a version of the form, or whether the data comes from the ELT process). These details are important because datasets that include records from multiple sources that are indistinguishable in general-purpose databases end up generating very different profiles and analyses.

Following the guidelines of VODAN implementation network and the tutorials developed by VODAN Africa&Asia, the VODAN BR FAIR DP will release the metadata referring to the repositories and their datasets, describing in detail, the data sources and their items. It will thus become part of the FAIR DPs VODAN federation, which aims to facilitate the dissemination/release of metadata about COVID-19 data, promoting access to this data by software agents and humans (Santos *et al*., 2016).

17.4 FINAL CONSIDERATIONS

The development of a computational asset in the form of a platform for the availability of research data regarding viral outbreaks in the middle of a pandemic is a great challenge, both in terms of computational aspects and public health aspects. As presented throughout this chapter, VODAN BR Project has been working continuously to implement its platform, maintaining an overview of the data, from the moment of its capture to the availability of metadata associated with repositories and datasets FAIR DPs.

The experience in the project reinforces the importance of FAIR DP in the infrastructure, not only as an essential element for the federation of access points and search and reuse mechanisms for FAIR data, but also for supporting sensitive research data that requires some degree of confidentiality, as is the case with patients' data. In this aspect, through the FAIR DPs, access to metadata referring to research data is made available, giving them visibility and accessibility. However, effective access to data respects well-defined conditions, promoting "data as open as possible and as close as necessary" (Wilkinson *et al*., 2016).

Among the lessons learned regarding the platform established for VODAN BR project, the lack of tools that help the FAIRfication process as a whole was observed. Some solutions adopted can be automated, improving the process and making the platform more stable to meet new challenges. An example that can be automated occurs in the process of publishing distributions of a dataset in the *Dataverse* repository and its associated metadata in FAIR DP.

Finally, we have end goals similar to those of the VODAN Africa&Asia network. The challenges being experienced during all phases of the Project provide different and complementary visions, providing a wealth of experiences that must be observed and analyzed, for the establishment of good practices to be taken to other FAIR implementation networks.

ACKNOWLEDGEMENTS

This work has been elaborated through multiple efforts. The authors would like to thank the VODAN-BR team, the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES) – Financing Code 001, 88887.613048/2021-00, CNPq – Financing Code 315399/2018-0 and 306115/2021-2, FAPERJ, Gaffreé Guinle Federal Hospital in Rio de Janeiro, São José Municipal Hospital in Duque de Caxias, professor Mauro Martin from ESDI-UERJ, and, in particular, the group of undergraduate and graduate students from PPGI/UFRJ program and other volunteer students who have been dedicating to the project since April 2020.

REFERENCES

CRUZ, S. M. S.; CAMPOS, M. L. M.; MATTOSO, M. Towards a taxonomy of provenance in scientific workflow management systems. *In*: INTERNATIONAL CONFERENCE ON WEB SERVICES, 2009, Los Angeles. **Anais** [...], 2009. DOI: 10.1109/SERVICES-I.2009.18. Available on: https://ieeexplore.ieee.org/document/5190667. Access on: 01 dec. 2023.

DATA TOGETHER COVID-19 Appeal and Actions. [S.l.: s.n.], [2020?]. Available on: https://www.go-fair.org/wp-content/uploads/2020/03/Data-Together-COVID-19-Statement-FINAL.pdf. Access on: 01 Dec. 2023.

GO FAIR. **Declaration**: Virus Outbreak Data Network (VODAN) GO FAIR Implementation Network, 2020. Available on: https://www.go-fair.org/wp-content/uploads/2020/03/VODAN-IN-Manifesto.pdf. Access on: 01 dec. 2023.

HEATH, T.; BIZER, C. **Linked Data**: Evolving the Web into a Global Data Space. Germany: Springer, 2011.

MONS, B. The VODAN IN: support of a FAIR-based infrastructure for COVID-19. **Eur J Hum Genet** v. 28, pp. 724–727, 2020. DOI: https://doi.org/10.1038/s41431-020-0635-7. Available on: https://www.nature.com/articles/s41431-020-0635-7. Access on: 01 dec. 2023.

SANTOS, L. O. B. S. *et al*. **FAIR Data Points Supporting Big Data Interoperability, Enterprise Interoperability in the Digitized and Networked Factory of the Future**,: Lisbon: ISTE Press, 2016.

SANTOS, L. O. B. S. **FAIR Data Point design specification**, [2020?]. Available on: https://github.com/FAIRDataTeam/FAIRDataPoint-Spec. Access on: 01 dec. 2023.

SATTI, F. *et al*. Semantic Bridge for Resolving Healthcare Data Interoperability *In:* INTERNATIONAL CONFERENCE ON INFORMATION NETWORKING, 2020, Barcelona, **Anais**..., 2020 p. 86-91.

STUDER, R., BENJAMINS, R., FENSEL, D. Knowledge engineering: Principles and methods. **Data & Knowledge Engineering**, v. 25, n. 1–2, p. 161–198, 2018.

WILKINSON M. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Sci Data**. v. 3 n. 160018, 2016. Available on: https://pubmed.ncbi.nlm.nih.gov/26978244/. Access on: 01 dec. 2020.

How to cite this chapter: CAMPOS, Maria Luiza Machado; BORGES, Vania; LOPES, Giseli Rabello; CAVALCANTI, Maria Claudia; MOREIRA, João; CRUZ, Sergio Manuel Serra da. VODAN BR - a platform for supporting COVID-19 data following 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. 17, p. 205-219. DOI: 10.22477/9788570131959.cap17.

ISBN: 978-85-7013-195-9

9 788570 131959