



ACADEMIA  
BRASILEIRA  
DE CIÊNCIAS

MCMXVI

**O**

**P**

**E**

**N**

**S**

**C**

**I**

**E**

**n**

**C**

**E**

## OVERVIEW AND GENERAL RECOMMENDATIONS



## OVERVIEW AND GENERAL RECOMMENDATIONS

### Working Group

- Claudia Maria Bauzer Medeiros (coordinator/editor)
- Alberto Henrique Frade Laender (co-editor)
- Abel Packer
- Adalberto Luis Val
- Carlos Henrique de Brito Cruz
- Cristina von Flach Garcia Chavez
- Eduardo César Leão Marques
- Fabio Kon
- Iscia Lopes Cendes
- Marcello A. Barcinski
- Marie-Anne van Sluys
- Ulisses Barres de Almeida

### Administrative Support

- Deborah Santos Lima Sant'Anna
- Marcos Cortesão Barnsley Scheuenstuhl
- Vitor Vieira de Oliveira Souza

### Design & Art

- Pedro Armando Santoro Dantas

NOVEMBER  
2023



### Board of Directors 2022-2025

#### President

- Helena Bonciani Nader

#### Vice-President

- Jailson Bittencourt de Andrade

#### Regional Vice-Presidents

- Adalberto Luis Val - *North*
- Anderson Stevens Leonidas Gomes - *Northeast & Espírito Santo*
- Virgílio Augusto Fernandes Almeida - *Minas Gerais & Central-West*
- Maria Domingues Vargas - *Rio de Janeiro*
- Glaucius Oliva - *São Paulo*
- Ruben George Oliven - *South*

#### Directors

- Alvaro Toubes Prata
- Maria Domingues Vargas
- Mariangela Hungria
- Roberto Lent
- Virgílio Augusto Fernandes Almeida

### Board of Directors 2019-2022

#### President

- Luiz Davidovich

#### Vice-President

- Helena Bonciani Nader

#### Regional Vice-Presidents

- Adalberto Luis Val - *North*
- Jailson Bittencourt de Andrade - *Northeast & Espírito Santo*
- Mauro Martins Teixeira - *Minas Gerais & Central-West*
- Lucia Mendonça Previato - *Rio de Janeiro*
- Oswaldo Luiz Alves\*/Glaucius Oliva - *São Paulo*
- João Batista Calixto - *South*

#### Directors

- Elibio Leopoldo Rech Filho
- Francisco Rafael Martins Laurindo
- Marcia Cristina Bernardes Barbosa
- Ruben George Oliven
- Virgílio Augusto Fernandes Almeida

\*deceased in 2021

Dados Internacionais de Catalogação na Publicação (CIP)  
(Câmara Brasileira do Livro, SP, Brasil)

Open science [livro eletrônico] : overview and  
general recommendations / [coordinator/editor]  
Claudia Maria Bauzer Medeiros. -- 1. ed. --  
Rio de Janeiro : Academia Brasileira de  
Ciências, 2023.  
PDF

Vários autores.  
Bibliografia.  
ISBN 978-65-981763-2-7

1. Ciência e tecnologia 2. Pesquisa científica  
3. Tecnologia I. Medeiros, Claudia Maria Bauzer.

23-182416

CDD-500

Índices para catálogo sistemático:

1. Ciência e tecnologia : Ciências 500

Aline Grazielle Benitez - Bibliotecária - CRB-1/3129

# Index

<b>Executive Summary</b> .....	<b>6</b>
<b>1 Introduction</b> .....	<b>10</b>
1.1 Open Science definitions .....	<b>11</b>
1.2 Open Science in a non-digital context .....	<b>13</b>
<b>2 Open Access publications</b> .....	<b>15</b>
2.1 Early Open Access initiatives: ArXiv .....	<b>15</b>
2.2 The quantitative evolution of Open Access .....	<b>18</b>
2.3 Types of Open Access .....	<b>19</b>
2.4 Another route to Open Access: Preprints .....	<b>21</b>
2.5 Benefits to authors and their institutions .....	<b>23</b>
2.6 Open Access policies .....	<b>25</b>
2.6.1 FAPESP's Open Access policy .....	<b>26</b>
2.6.2 The European Commission's Plan-S policy proposal for Open Access .....	<b>28</b>
2.7 Free for readers, not for authors or their funders/institutions: the costs of Open Access .....	<b>29</b>
<b>3 Open Data</b> .....	<b>33</b>
3.1 What is data? .....	<b>33</b>
3.2 Data life cycles and practices .....	<b>34</b>
3.3 Defining Open Data .....	<b>36</b>
3.4 The role of Data Management Plans .....	<b>37</b>
3.5 Open Data in Brazil .....	<b>38</b>
3.6 Trusted data repositories .....	<b>39</b>
3.7 International open data bodies — RDA and WDS .....	<b>43</b>
<b>4 Open Source Software</b> .....	<b>45</b>
4.1 Free/Open Source Software .....	<b>45</b>
4.2 Software life cycles and practices .....	<b>46</b>
4.3 Open Source and Research Software .....	<b>47</b>
4.4 Examples of Successful Open Source Research Software .....	<b>47</b>

<b>5</b>	<b>Some additional aspects</b> .....	<b>49</b>
5.1	Citizen Science .....	49
5.2	Open Science and Biodiversity research .....	51
5.3	Ethics, Privacy, and Security .....	52
	5.3.1 Data privacy protection legislation .....	53
	5.3.2 Data and algorithm ethics .....	55
5.4	Overall challenges .....	56
	5.4.1 Change in culture and attitude .....	56
	5.4.2 Training and education — educate for openness, train in open digital practices and good science .....	57
	5.4.3 Sustainability and Costs .....	59
<b>6</b>	<b>Recommendations on Open Science in Brazil</b> .....	<b>61</b>
<b>7</b>	<b>References</b> .....	<b>63</b>
<b>8</b>	<b>Acronyms</b> .....	<b>67</b>
<b>9</b>	<b>Glossary</b> .....	<b>69</b>

# Executive Summary

This report aims to give an overview of the Open Science (OS) movement to the Brazilian scientific community as a whole, commenting on some of the associated challenges and presenting an initial set of recommendations on how to launch initiatives towards fostering OS practices within Brazilian academia, and supporting the corresponding change of culture in all sectors. It intends to raise awareness within the Brazilian scientific community of the many facets of OS, starting discussions, and motivating initiatives, thereby bringing new perspectives on best research practices directed towards collaboration in research and the opening of all processes involved in scientific creation. We point out that, although there are studies that consider that OS has been practiced since the 18th century, most of this report concerns digitally mediated Open Science.

There is no standard definition of Open Science, but rather a growing consensus on its importance and goals. It is centered on the notion of advancing research through making scientific knowledge openly available, accessible and reusable for everyone, thereby benefitting science and society. There are two key underlying concepts. The first is advancing knowledge through *scientific collaboration* — and *information sharing* is an intrinsic element within any collaborative effort. Although collaboration and sharing have existed in research for centuries, most Open Science definitions center on achieving them through digital means, i.e., collaboration and sharing mediated by Information and Communication Technologies (ICT). The second key concept is that openness (sharing and dissemination) mediated by ICT is implemented by making all “objects” associated with a research effort publicly available to all, in *reliable repositories*. These objects include, for example, publications, methods, data, code, algorithms, hardware specifications, among others. The goal is to make scientific research accessible to all, so that it is reusable and verifiable.

Though applicable to all kinds of research practiced in any type of environment, public or private, Open Science recommendations and policies mainly refer to publicly funded research. The underlying principle is that *outputs of research financed by public funds are a public asset*. Thus, such outputs must be made openly available to all sectors — the scientific community, businesses, government and ultimately society as a whole — as soon as possible, while respecting ethical and legal constraints such as privacy, security, or intellectual property. This need for openness under constraints has given rise to the expression “*as open as possible, as closed as necessary*”.

Open Science is more than storing data, or papers, or software in reliable repositories. It includes many aspects and stages of research processes, from the minute a given research proposal is conceived, to beyond its completion — so that all outputs can be preserved for future (re)uses by anyone, anywhere, anytime. Let us briefly enumerate some of its main outputs.

*Open Access* refers to making scientific papers available at no cost to readers — though, as discussed in Section 2, there are costs associated with open publications, under different kinds of licenses. The publication ecosystem involves many stakeholders, including publishers, researchers or funders, who can establish open agreements under many kinds of configurations, Article Processing Charges (APC) and constraints, with or without embargo.

*Open Data*, detailed in Section 3, considers all kinds of digital data consumed or produced in a project — e.g., text, spreadsheets, photos, videos, sound. Such data may be primary (directly collected by researchers from main sources) or secondary (resulting from analyses performed in the project). Synthetic data refers to those that are created using computational techniques, for instance when data collection is unfeasible, e.g., due to cost, availability or privacy — such as when a physical phenomenon cannot be repeated or creating artificial medical records. For a more detailed document on issues related to Open Data, see the report published by the Brazilian Academy of Sciences [ABC 2020].

*Open Source Software* (for short, Open Software), presented in Section 4, refers to a collaborative way of developing and sharing software. Major ICT companies in the world use and produce software distributed under open source licenses, either in large packages or small components. Open software can be found at all implementation levels — from instrument drivers to high-level applications and frameworks. Free software refers to code that is freely available to everyone to use and change, its usage being subject to open licenses. Since open software is often made available as code, sometimes software is considered yet another kind of data, but a software is a separate, first-class digital citizen — software embodies the computational processes that are performed on data.

Additional elements include *Open Peer Review*, *Open Hardware*, interoperable scientific infrastructures, open and shared research methodologies, and so on. Open Peer Reviews, discussed in Section 2, consider that peer reviews should be made publicly available, since reviews are a necessary part of the publication process. Interoperable infrastructures, mentioned in Section 1, consist of hardware and software that are set up so that distinct research installations can be interconnected to support sharing and collaboration. Open Hardware is not covered in this report and refers to design specifications and code used to create any physical device from chips to complex computers, and even robots.

An important component of an interoperable infrastructure are *Repositories*, covered mostly in Section 3, which provide support to store, access, use and reuse all digital research assets mentioned here. The creation and maintenance of repositories for OS require the adoption of standards, to ensure interoperability and knowledge transfer to all, and across scientific fields, Institutional, trustable repositories ensure that their contents have the required quality for reuse. An associated concept is that data, software, hardware specifications and others must be created and made available in repositories in a FAIR manner (Findable, Accessible, Interoperable, Reusable). FAIR principles and documentation via metadata are discussed in Section 3.

Open availability of research objects can improve the effectiveness and productivity of the research system, e.g., by reducing duplication costs in collecting and managing data, creating new knowledge through reuse of data or code, and facilitating participation of individuals and institutions beyond regional, national and even cultural or disciplinary borders. Reproducibility is also facilitated, since openness helps independent audit and verification of research findings. In addition, the contents of open repositories are being used in combating fake news, promoting innovation and public understanding of science. Citizen Science, seen in Section 5, is an example of how openness stimulates citizen participation. Many studies find that open access increases citations; in some cases, when publications are accompanied by the underlying open data, citations increase by as much as 25% since data can be cited through its own DOI (Digital Object Identifier) — see Sections 2 through 4. And this leads us to contrast benefits to costs.

Several sections of this report are dedicated to the costs of Open Science. Science is not for free and involves long-lasting commitment and stable financing. Open Science is no different. First of all, *it is science*, practiced in an open manner — thus, the traditional costs to support research do not disappear. There are savings in the long run, for example, through the reuse of data and other assets mentioned in the previous paragraphs. But openness brings about new considerations. Like all scientific endeavors, it still requires the same types of investment, including continuous support and maintenance of all kinds of computational infrastructure (hardware, networks, software), now with interoperability in mind. Open access to publications, for instance, also involves new kinds of costs, such as those to operate journals under this model. As shown in several studies, the benefits of OS surpass and justify the costs. Through fostering collaboration, it fortifies integration at regional, national and international levels, decreasing inequity in research.

Nevertheless, perhaps the most important and perennial cost factor is related to the continuous education and training of people, involving not only technical subjects, such as how to prepare data or software for publication and reuse, but also culture change. Best practices in open research require considering that all steps of a research process must take openness into account — documenting for reuse by unknown others, preparing data or software for sharing, considering long-term preservation. Scientists write papers carefully, so they can be understood beyond their group. By the same token, data and software must be managed and prepared for reuse. Thus, scientists must be offered training on “preparing for openness” that goes beyond the often-mentioned Data Science disciplines.

Training and education lead us to the “who”, namely, the actors involved in and who benefit from the OS movement. This includes researchers, research staff, IT professionals, librarians, policymakers, funders, academic institutions, publishers, businesses, supra-national entities and, ultimately, all citizens. This is covered in Sections 3 (OS ecosystem) and 5 (Training and Education). Since Open Science requires additional work, reward mechanisms must be devised to recognize those that practice it. This has already become part of researcher evaluation criteria in several European countries, as part of official recommendations proposed by hundreds of research institutions and most major European funders, as of July 2022.

Needless to say, “as open as possible, as closed as necessary” entails ethical and legal issues. The Internet has facilitated collaboration, and new devices allow us to collect and disseminate data in unimaginable volumes, from everywhere (from the center of the Earth to cosmic space) at all times. This, in turn, has created new scholarly research fields, such as Data and Algorithmic Ethics, as well as required establishing new legal frameworks concerning, e.g., security and privacy. Ethics and legal frameworks appear in Section 5, including the emergence of Data Protection Laws, and Ethical Artificial Intelligence.

At the end, this report presents a few recommendations on Open Science to the Brazilian scientific community, so that it can continue to fulfill its initiatives towards the scientific development of Brazil and the world. As such, ABC aims to be a primary actor in the Brazilian Open Science movement, promoting and disseminating best practices and open policies, and assisting and informing government, funders and academic institutions. It is our hope that the focus of ABC on open science education through thematic events and its many international activities will contribute to the success of the movement in Brazil.

# 1 Introduction

In November 2021, UNESCO voted its recommendation on Open Science [Unesco 2021]. The 34-page text is the result of three years of regional discussions and consultations, with contributions from more than 110 countries and experts indicated by academies and research coalitions. This is perhaps the most visible outcome of the *Open Science movement* that is growing all over the world. Through this effort, UNESCO recognizes the importance of this movement, and its implications to the advancement of knowledge.

Before UNESCO, other supranational entities had already issued recommendations on Open Science, in particular, OECD [OECD 2015], providing insights into not only its scientific, but also economic value. OECD's report is mostly geared towards publicly funded research, and concerns open publications and open data, analyzing their effects on research, innovation, academia, businesses and society as a whole. Additional OECD reports, some of which mentioned subsequently here, have since then continued to explore different aspects of the OS movement, with guidelines, recommendations, and legal instruments.

In parallel, the European Parliament has allotted considerable funding to Open Science initiatives (for instance, within its European Open Science Cloud<sup>1</sup> framework, throughout 2027). Canada and the USA officially recognize its importance. The year 2023 has been declared by the American government to be “the year of Open Science” for all institutions within the federal government<sup>2</sup>.

The Brazilian research community has been a strong contributor to many facets of Open Science, recognizing that it offers a wide range of opportunities that can bring about scientific, technological, socioeconomic and cultural impact. For instance, not only are we considered a data-rich country, but we are also renowned by making scientific data openly available. Brazil is also the country where SciELO<sup>3</sup> was born in 1998 — a world pioneer in open publications.

What is, then, Open Science, and what are its challenges and opportunities, and how can the Brazilian Academy of Sciences contribute to this movement? Though there is no fixed definition of the term “Open Science”, it is usually used to denote the set of policies, initiatives and actions to disseminate knowledge, usually through digital means, so that all outputs associated with scientific research become accessible to all, are reusable and support reproducibility. Such outputs include publications, data, algorithms, computational processes, software, hardware design specifications, and methodologies used to conduct a given research project.

The ultimate goal of Open Science is to promote innovation and advancement of knowledge through collaboration and reuse of research outputs, regardless of geographic, temporal, politi-

---

1 <https://eosc-portal.eu>

2 <https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/>

3 <https://www.scielo.br/>

cal, social or cultural barriers. While this collaboration is actively promoted among researchers, Open Science also supports collaboration between researchers and society, thereby accelerating scientific, technological, economic, and social progress.

## **1.1 Open Science definitions**

While many scholars claim that the Open Science movement started in the 1960s, others indicate that its origins date back to the 18th century and the birth of encyclopedias, and scholars' practices of exchanging information via correspondence. Open Science in the digital era expands these earlier concepts of communicating science globally and validating the shared information. Most studies and definitions concern themselves only with the digital world, thereby clearly limiting Open Science practices to the creation, management, sharing, and reuse of non-physical (i.e., digital) objects, using Information and Communication Technologies (ICT) to do so.

A landmark document on the specification of the pillars of Open Science was the 2018 study of the American Academy of Sciences [NAS 2018]. It is one of the earliest publications to enumerate and discuss the full chain of Open Science practices in a scientific research environment. It sums up the movement through a description of its practices — namely, all activities conducting to research output sharing. Under the framework established by this text, to practice Open Science, one has to consciously work towards it, from the first stage of research design, throughout all activities of the research life cycle, so that all outputs are planned and produced with subsequent preservation and sharing in mind. The most frequently cited outputs in the document are open publications, open data, and open software/computational processes.

Another relevant report was produced in 2020 by the InterAcademy Partnership [IAP 2020]. It presents Open Science as being centered on global collaboration — through the publication, in open repositories, of all outputs of research. This scenario is defined in terms of the so-called “Open Science ecosystem”, in which researchers deposit the outputs of their work in public repositories, to enable reuse and collaboration without geographic, temporal or social barriers. Just as no one can say who will read our paper, when and for what purposes, and how its results will be applied, Open Science requires that all outputs of research be prepared for being shared under the same assumptions. Figure 1, reproduced from that report, shows the components of this ecosystem. At its center lies the research community, which collaborates through exchange of research outputs and research activities. This community influences (and is influenced by) open research practices and facilitators. Its implementation and expansion requires not only the appropriate research e-infrastructure (software, hardware, repositories, networks), but also a set of enabling factors, including education and cultural change. Open Science benefits science, technology, innovation and society, and it must be guided by the principles of trust, equity, inclusion, and responsible research conduct.

While the recommendations of The National Academy of Sciences [NAS 2018] and The InterAcademy Partnership [IAP 2020] are written having scientists and research in mind, UNESCO's recommendations [UNESCO 2021] extend the discussion to all actors of the Open Science movement. The basis of the text's recommendations was produced throughout 2020 and most of 2021 by scientific academies and some regional coalitions. While emphasizing the absence of a consensual definition of Open Science, it points out the following: "... Open Science is defined as an inclusive construct that combines various movements and practices aiming to make *multilingual* scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and information sharing for the benefit of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community. It includes all scientific disciplines and aspects of scholarly practices ..." It singles out the following consensual elements that enable open access to scientific knowledge — "open scientific publications, open research data, open source code, and open hardware". The latter refers to the open design specifications of hardware devices (not the physical devices themselves). In this definition, it is worthwhile noticing that the word "multilingual" did not appear in any draft, being introduced in a three-day meeting in May 2021, when the final draft was discussed by representatives from all countries. It is in italics here to help finding it out in the definition.

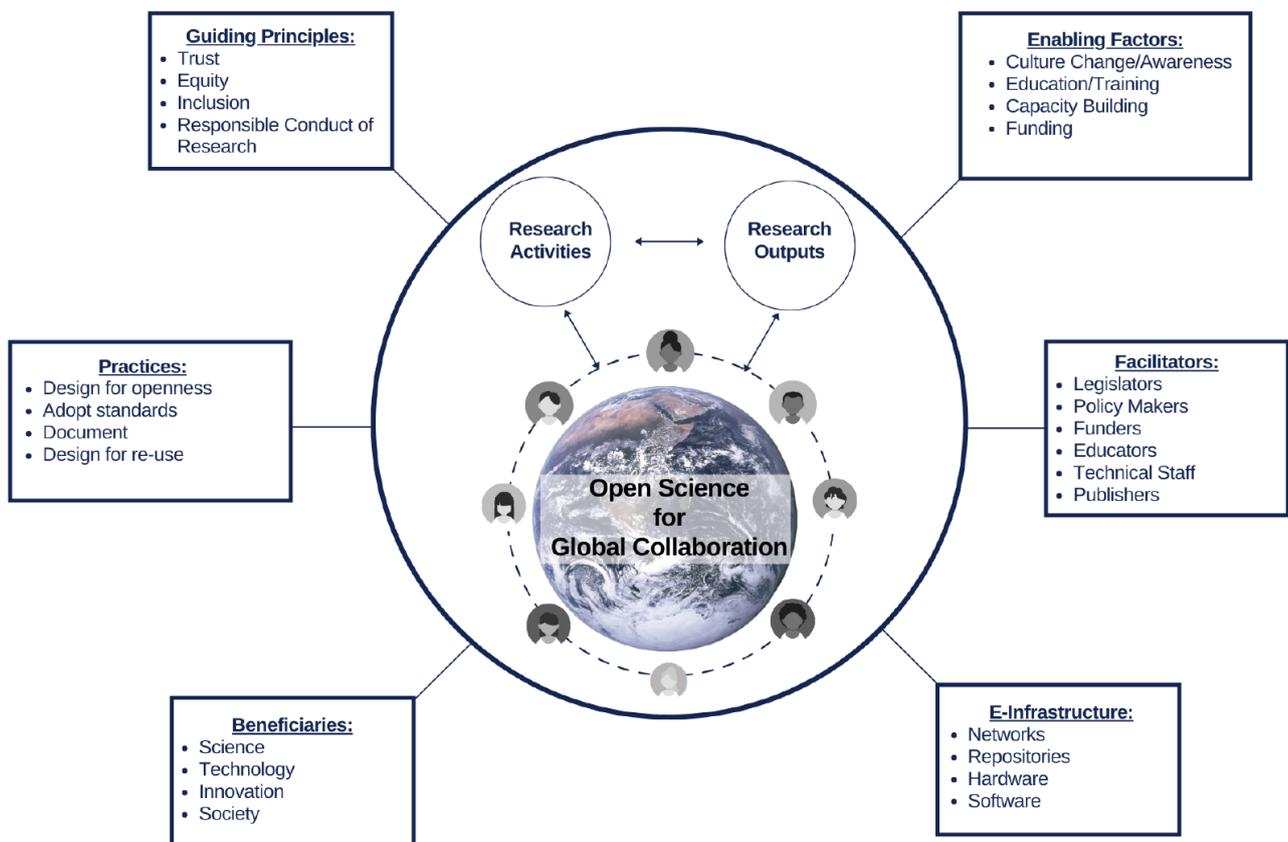


Figure 1. Open Science Ecosystem. Reproduced from [IAP Report 2020]

Be that as it may, the definitions and recommendations agree on some key principles and guidelines, in line with a core of best practices towards Open Science. Key to all definitions is the notion of collaboration, which is extended from the usual context of cooperation within a laboratory, or within a set of research teams, to collaboration through digital objects. These objects are, in turn, created in response to a research question. They may be research outputs generated during a specific research, but they may also be produced by non-scientists, e.g., in citizen science (see Section 5.1 in this report). Figure 2 illustrates this overview of the implementation of Open Science through public repositories, which computationally mediate scientific collaboration.

This report covers three of the four basic elements of the UNESCO report — open publications (Section 2), open data (Section 3), and open software and code (Section 4). For a brief introduction to issues in open hardware, in Portuguese, see [Carro 2021].

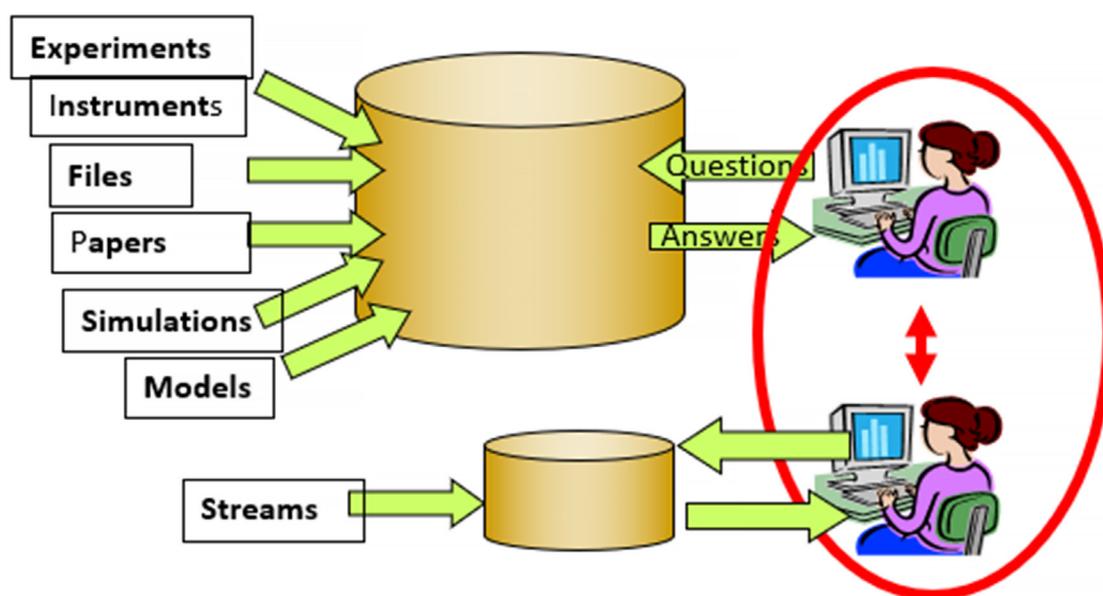


Figure 2. Collaboration through sharing digital objects related to some research. Such objects must be made available through public repositories, which are subject to auditing, and constructed and maintained following FAIR and TRUST principles. This figure is an adaptation of a 1990 slide by Jim Gray and Alex Szalay on e-Science.

## 1.2 Open Science in a non-digital context

This report includes this section for completeness' sake. It is neither our goal to discuss the intricacies and particularities of the practice of Open Science concerning physical samples, nor it is our goal to discuss issues such as collection, curation, sharing, and preservation of non-digital objects that are part of open scientific research. Last but not least, we will not discuss the problems of transforming such physical objects into digital ones, always remembering that one physical sample (e.g., a blood sample, a rock, or a parchment) can be processed to create tens or hundreds of digital objects that describe that physical instance.

Science is a human activity built cooperatively from several simultaneous or hierarchical-dependent insights. Science progressed by the need to take notes, share by communication, and build libraries, museums, herbaria, pharmacopeia, seed banks, mineral collections and other repositories of physical samples. Thus, science resides in openness to be fruitful. The intricate connection of ideas, observation, experimentation, interpretation and dissemination has been successful over the centuries because information was preserved and communicated. As a result of its success, science faces the challenge to make the information generated worldwide easily findable. All of these long-standing repositories of physical scientific objects have their codes of integrity and maintenance of collections that are clearly set for preservation and sample sharing. Information has been shared by mailed letters and packages, or by visits, feasible to a small community. Nowadays, there are hundreds of thousand researchers that want to communicate and share their work. The transition from physical to digital Open Science is a 21st century challenge, not equally distributed across disciplines.

There are several points of view on the origins of Open Science as a means to conduct research. David [2014] analyzes the movement from a historical perspective, tracing the origins to, at least, the collaboration practices of the Age of Enlightenment. Be that as it may, science built on evidence depends on its reproducibility. Open Science presupposes open sharing for collaboration, but also reproducibility. While the digital facet of Open Science makes sharing easier, it also creates new kinds of barrier to sharing, reproducibility and preservation.

Experimentation and observation require careful data collection, recording and analyses. The digital data deluge and data science practices challenge the need for preservation, because of the size of the data files (either collected or produced as a result of simulations), the collection speed and issues of digital decay. The need for summarization, or size reduction to provide a manageable analysis, depends on criteria that also need to be clearly specified.

## 2 Open Access publications

Open Access is a term that denotes publications, though sometimes scientists use “open access” to denote any kind of digital open research output. Open Access became, in the last 20 years, a well-established scholarly publishing model through which scientific literature is made available on the Web to readers, without any financial or other barriers (though there may be cost barriers for authors to publish). Normally, readers can reuse texts following options established by Creative Commons licenses<sup>1</sup>. The Web is the main driver of the Open Access initiatives worldwide.

### 2.1 Early Open Access initiatives: ArXiv, SciELO, The Budapest Declaration and PLOS

One of the first large initiatives related to Open Access came in 1991 when Paul Ginsparg started, at the Los Alamos National Laboratory, an online repository of electronic preprints (then called e-prints) of scientific papers. The repository was renamed ArXiv.org in 1999 and, since 2001, is maintained and operated by Cornell University with support from donors and foundations. As of January 2020, ArXiv hosted 1.6 million e-prints (or pre-prints, using the more up-to-date term) in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, Statistics, Electrical Engineering and Systems Science, and Economics<sup>2</sup>. By July 2021, the number of downloads per month reached 35 million<sup>3</sup>.

Another relevant early Open Access initiative (though, as ArXiv, it did not explicitly mention the expression “Open Access”) was the launching in 1998 of the Scientific Electronic Library Online (SciELO), a Web-based library to index, publish and interoperate nationally peer-review journals from all disciplines. One of the motivations for SciELO was to provide more visibility to research results created in Brazil [Gibbs 1995] (and later, Latin America, Portugal, Spain, and South Africa). SciELO was created as a cooperation between the São Paulo Research Foundation (FAPESP) and the Latin American and Caribbean Center on Health Sciences Information of the Pan American Health Organization / World Health Organization (BIREME)<sup>4</sup>. Established as a FAPESP Program, SciELO is also supported by the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) and the National Council for Scientific and Technological Development (CNPq) as a national research infrastructure program aiming to improve quality, visibility, and impact of journals and the research they publish [Packer and Meneghini 2007].

---

1 About CC Licenses, in Creative Commons. Retrieved on November 16, 2020, from <https://creativecommons.org/about/cclicenses/>

2 [https://arxiv.org/about/reports/2020\\_update](https://arxiv.org/about/reports/2020_update), checked on November 14, 2020

3 [https://arxiv.org/stats/monthly\\_downloads](https://arxiv.org/stats/monthly_downloads), checked on August 8th 2021

4 <https://tonyhey.net.com/2017/03/10/a-global-view-of-open-access-2-the-perspective-from-brazil-and-the-scielo-open-access-portal/>

By 2020, SciELO Brazil already indexed and published a portfolio of about 300 open access peer-reviewed journals. Its publishing model is adopted in 16 other countries whose national collections total over 1200 active journals. Altogether, the SciELO initiatives publish about 50 thousand new documents per year and have accumulated more than 940 thousand documents that serve an average of more than 40 million accesses and downloads per month. The SciELO Brazil repository alone accumulates 425 thousand full-text documents that serve a monthly average of over 25 million accesses and downloads based on data from the first semester of 2020 — according to the COUNTER methodology which excludes robots.

Most journals in SciELO Brazil improved their visibility in terms of citations received. Figure 3 shows the 10-year evolution of the average of Scimago “cites per doc in 2 years” indicator, similar to Impact Factor, for the 222 SciELO journals indexed by Scopus in 2019. There was a systematic improvement of visibility in all major research fields due to the combination of open access and internationalization.

The pioneering and increasing visibility of SciELO contributed to the dissemination of the publishing of Open Access journals, particularly in Latin America, including other networks of publications. In 2003, the Universidad Autónoma del Estado de México established the Red de Revistas Científicas de América Latina y el Caribe, España y Portugal (Redalyc) which, by 2020, aggregated a collection of more than 1300 open access journals<sup>1</sup>. In 2012, the Federated Network of Institutional Repositories of Scientific Publications (La Referencia) was established and since then operates under an agreement between public research institutions in 10 countries. By the end of 2021, the La Referencia indexed about two million articles<sup>2</sup> and an additional million doctoral theses and M.Sc. dissertations. The Brazilian Institute of Information on Science and Technology (IBICT) plays a lead role in the operation and development of La Referencia.

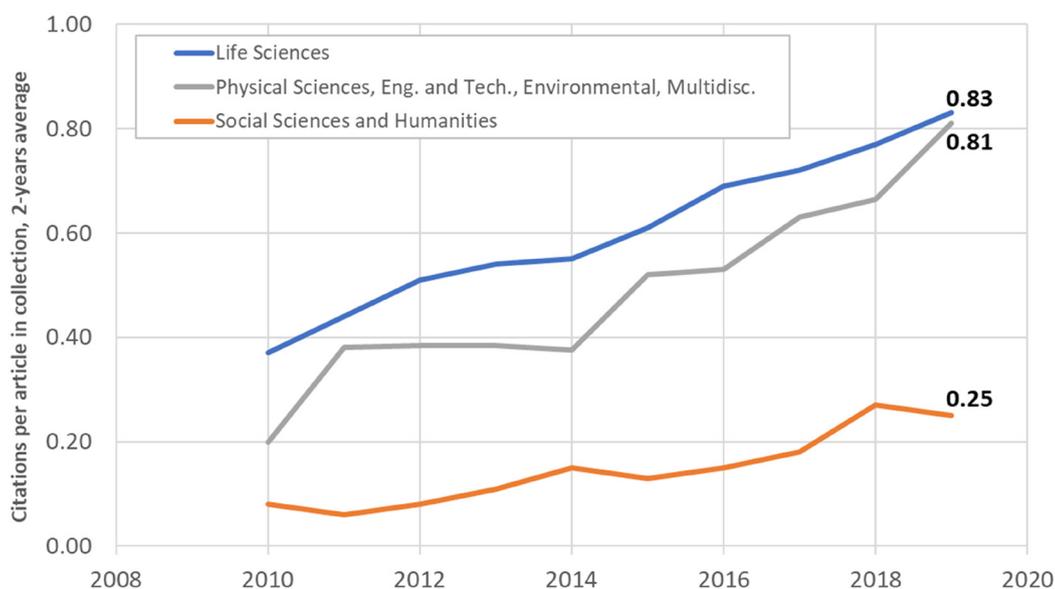


Figure 3. Evolution of the median of the Impact Factor of SciELO Brazil Journals in Scopus by major knowledge fields, 2010-2019 (Source: Scopus database, retrieved on 2020-11-10).

1 <https://www.redalyc.org/>

2 <https://www.lareferencia.info/en/>

Since 2018, SciELO has been promoting the transition of the journals to an Open Science *modus operandi* that includes the adoption of preprints as a formal start of the research article communication, the citation and referencing of all underlying article contents to ease evaluation, reproducibility, reuse, and options to make the peer review process more transparent.

The third early initiative is The Budapest Open Access Initiative (BOAI) declaration. Agreed to in 2002, it is internationally recognized as “one of the major defining events of the open access movement” [Budapest 2002]. It formalized a widely accepted concept of open access to the literature that “scholars give to the world without expectation of payment”, including journal peer-reviewed articles and preprints. The BOAI showed a useful definition of Open Access [Tenant 2016]:

*“By ‘open access’ [peer-reviewed research literature], we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.”*

Some additional relevant Open Access initiatives include the Public Library of Science (PLOS), launched by an open letter organized by Harold Varmus, Patrick Brown, and Michael Eisen in 2001 (which was afterwards subscribed by more than 34,000 scientists<sup>1</sup>), and BioMed Central (BMC), a company created in 2000 using the concept of “free to readers” publishing. PLOS launched its first publication, PLOS Biology, in 2003, while BMC launched its BMC Biology in 2001. Both PLOS and BMC pioneered<sup>2</sup> the introduction of Article Processing Charges (APC), charged to the authors of each article accepted, to fund the costs of operation of the journal. In 2006, PLOS launched PLOS ONE, the first mega journal (the term refers to peer-reviewed academic open access journals that are much larger than a traditional journal by exercising low selectivity among accepted articles).

---

1 <https://plos.org/about/>

2 <http://newsbreaks.infotoday.com/nbreader.asp?ArticleID=17276>

## 2.2 The quantitative evolution of Open Access

The number of publications available in Open Access has been growing steadily worldwide, as shown in Figure 4. In 2019, out of a total of 2,462,631 scientific publications (articles and reviews) available in the Scopus database, 33%, or 813,860 were in Open Access (Gold OA only). We point out that Gold OA refers to publications that were accessible in OA immediately after publication by the journal, with no embargo time. In Brazil, the adherence is above the world average (Figure 5): in 2019 out of 72,676 publications with authors from the country, 46%, or 33,752 were available in Gold Open Access. It would be reasonable to expect that the existence of SciELO contributes to this result.

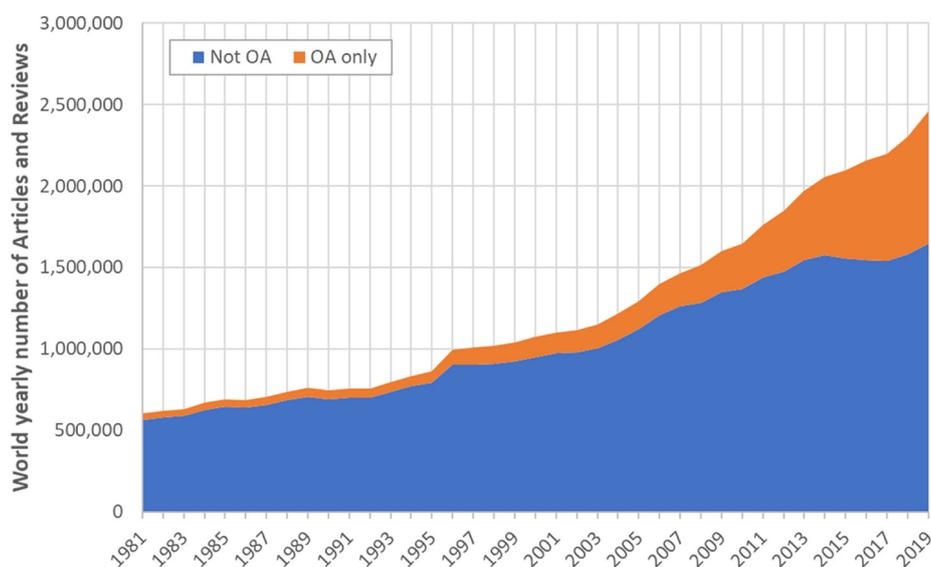


Figure 4. Number (world total) of Open Access (Gold OA only) and Non-Open-Access articles and reviews published yearly, since 1980. In 2019, out of 3,349,988 scientific publications, 1,003,102 were in Open Access (Source: Scopus database, searched on 2020-11-14).

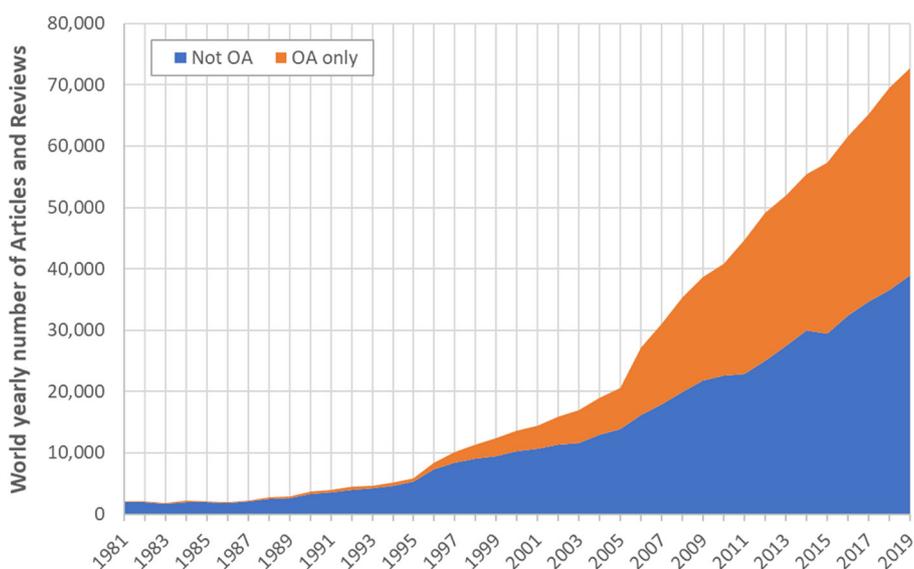


Figure 5. Number of Open Access (Gold OA only) and Non-Open-Access articles and reviews published yearly with authors in Brazil, since 1980. In 2019, out of 86,963 scientific publications, 36,506 were in Open Access (Source: Scopus database, searched on 2020-11-14). The same source shows that Brazil, with 47% of its publications in OA, comes in fourth place.

Scopus database, checked in November 2020, shows Brazil in 4th place in ranking (among 30 countries with more than 20,000 publications per year) in the percentage of Gold OA in the 2018-2019 period.

## 2.3 Types of Open Access

The progress of Open Access scholarly publishing encompasses several business models and mechanisms to cover publishing costs and mark-up. They are best established in peer-reviewed publishing of journal articles, which is widely considered the predominant and most structured way to communicate and evaluate research, a practice that will probably remain for a long time in coexistence with the utilization of other Open Science research communication objects. In general, the open availability of articles is provided by journal publishers and/or by author self-archiving.

Open Access to a publication can be directly provided by the journal publisher (sometimes called gold OA), or indirectly by being uploaded in some version (we will explain the possible versions below. Roughly, they might be an Author's Submitted Manuscript (ASM) or an Author's Accepted Manuscript (AAM)) and made freely available somewhere else on the Web (mostly called green OA) – e.g. an institutional repository of publications, or scientific social networks such as ResearchGate and Academia. Laakso et al affirm that “Both options increase the potential readership of any article to over a billion individuals with Internet access and indirectly speed up the spread of new research ideas” [Laakso 2012]. In parallel, journals may adopt the Hybrid Open Access model, in which subscription-based journals allow authors to turn individual articles into open access articles immediately upon payment of a publication charge.

The classification of the types of Open Access practice using colors (gold, green, pale green, and so on) was once popular. Recently, the main organization that offers a classification (Sherpa Romeo<sup>1</sup>) gave up on the colors as the nuances between diverse publishers became too subtle and varied<sup>2</sup>. Sherpa Romeo is an online resource that aggregates and presents publisher and journal open access policies from around the world. Still, it is useful to know that:

- a) *Gold Open Access* refers to the cases in which the journal makes the Version of Record (VoR — the actual accepted, reviewed, copy desked, and formatted version) immediately available freely in its website and, optionally, in the author's institutional websites or other, under a Creative Commons (CC) license.
- b) *Green Open Access* refers to the cases in which a journal accepts (sometimes with conditions, like a latency time of up to 12 months) that the ASM or AAM version (but, in many cases, not the VoR) of the article be offered freely in an institutional repository. This procedure is sometimes referred to as “self-archiving”.

---

1 Sherpa Romeo <https://v2.sherpa.ac.uk/romeo>

2 <https://v2.sherpa.ac.uk/romeo/about.html>

Sherpa Romeo is an extremely valuable resource that publishes a frequently updated database of the policies for Open Access used by thousands of relevant scientific journals. For example, consulting about the publisher American Physical Society and the journal Physical Review Letters (see Figure 6), one will learn that:

- a) There might be different conditions for the ASM and AAM versions, and two possible paths for the actual published version.
- b) The symbols for the “submitted versions” indicate that there is no time delay (from the date of publication), and no cost, to post this version to the Author’s Homepage, Institutional Repository, or an Institutional Website (certain conditions might apply and can be found by clicking on the “+” icon).

Going into more detail, one will find that the Published Version of an article (PDF form) on PRL can be freely offered on the Author’s homepage or on their Institutional Repository or Institutional Website. As long as the Institutional Repository is not shared with other institutions, a link to the published article is included in the page at the repositories or author page, and the published source is acknowledged with citation. As mentioned above, there are many different possibilities used by publishers. Nevertheless, it is worth knowing that many publishers have policies that lower the restrictions when the authors are subject to certain types of funder or institution mandates to have their articles available in Open Access (see Figure 7).

#### Publisher Policy

Open Access pathways permitted by this journal’s policy are listed below by article version. Click on a pathway for a more detailed view.

Published Version [pathway a]	 None  	 Institutional Repository, Institutional Website, Author’s Homepage	+
Published Version [pathway b]	   None  CC BY 	 Any Repository, Journal Website	+
Accepted Version	 None  	 Institutional Repository, Institutional Website, Author’s Homepage	+
Submitted Version	 None  	 Institutional Repository, Institutional Website, Author’s Homepage	+

Figure 6. The site Romeo Sherpa displays, for each journal, the different possibilities for offering an article in Open Access, and the conditions for that. The figure shows the case for the American Physical Society’s Physical Review Letters. (Source: <https://v2.sherpa.ac.uk/id/publication/13640>, accessed on 2020-11-15).

Published Version [pathway a]	 None    Institutional Repository, Institutional Website, Author's Homepage
 Embargo	No Embargo
 Location	Author's Homepage Institutional Repository Institutional Website
 Conditions	Must link to published article Published source must be acknowledged with citation
 Notes	Institutional repository must not be shared with other institutions

Figure 7. Example of information for the American Physical Society's *Physical Review Letters* treatment of Open Access Possibilities (Source: <https://v2.sherpa.ac.uk/id/publication/13640>, accessed on 2020-11-15).

This diversification of alternatives is relevant, considering the nuanced consequences of a transition from a reader financed system to a system funded by authors (and their funding agencies). If, in general terms, the full adoption of Open Access publications is clearly beneficial, important details may increase inequalities in academic fields. In publication systems financed by readers, academics from poorer and intensively unequal countries tended to have lower (and slower) access to academic reading. Open Access intends to face this problem directly. However, if this means the generalized adoption of APCs, then social inequalities between countries (and within them) may be reinforced. The problem is not solved just by directly providing financing support by research funding institutions, since some countries (or country regions) lack those agencies or have them severely underfinanced. The problem also presents itself in some knowledge production fields that traditionally did not depend on funded research projects up to this point, such as philosophy, regardless of the presence of strong funding agencies. Therefore, the existence of a variety of Open Access alternatives is highly beneficial to science and knowledge production.

## 2.4 Another route to Open Access: Preprints

The first course to accelerate Open Access is led by authors via the deposit of their manuscript in a preprint server before — or in parallel with — the submission to a journal. Preprints are open access manuscripts not yet peer reviewed and published by a journal [Bourne 2017]. They may, of course, have been informally peer-reviewed outside the framework of a journal. By now, the use of preprints is a practice of Open Science viewed as a means to accelerate the dissemination of new knowledge and strengthen transparency. Preprint servers have been around for almost 30 years since the launch of arXiv in 1991, initially for physics and progressively covering astronomy, mathematics, and statistics related disciplines, computer science, and quantitative biology.

In the end of 1993, the Cold Spring Harbor Laboratory launched the bioRxiv preprint service, promoting the wide acceptance of preprints for biology, and in June 2019 the medRxiv service was launched in partnership with Yale University and BMJ for medical, clinical and related health sciences, which acquired special relevance for the communication of manuscripts related to SARS-CoV-2 and COVID-19. In fact, both servers received a total of 10 thousand preprints in the first 10 months of 2020, with medRxiv receiving 80% of the uploads. Commercial publishers have increasingly established associated preprint servers in the peer review process: Springer Nature's In Review — Research Square preprint; Wiley Under Review — Authorea preprint; Elsevier First Look service — SSRN preprint; Taylor & Francis — F1000 Research pre and postprint. The SciELO Program launched SciELO Preprints in April 2020, which is expected to enhance the options of research communication associated with the SciELO journals. All these cited preprint servers are recognized as trustworthy, as they have quality control through moderation to identify research-related manuscripts, assignment of a DOI (Digital Object Identifier) and a Creative Commons (CC) access license to accepted preprints, and support preservation, interoperability with related data sets, and provisions for versioning mechanisms so that manuscripts can be improved before submission to journals. When approved by a journal, the preprint usually points to the published version of the article.

In addition to open access and preprints, publications in Open Science renew the classical research communication flows with additional practices envisaging improved cooperation and transparency, in particular Open Data (see Section 3) and the so-called Open Peer Review (OPR). Both impact the classical way journals communicate research, which in most of the cases rely only on published articles.

“Open Peer Review (OPR)” applies to the evolving types of manuscript review that is being adopted by journals — Figure 8 portrays such types. There is no consensus on a definition of OPR. In fact, a systematic review identified 22 different definitions of OPR [Ross-Helauer 2017]. The goal is to enrich manuscript evaluation by improving communication between authors and reviewers including opening identities, publishing the review reports along the approved articles as a new type of literature, and exposing the version of record to comments. OPR is expected to speed the availability of research results, to address the frequently criticized unreliability and the lack of responsibility of traditional single - or double - anonymized peer review. It is still far from a wide acceptance. The longest and most advanced application of OPR is conducted by F1000 Research<sup>1</sup>, an open platform where authors submit their manuscripts that, if accepted by an initial moderation, are made publicly available in a preprint status identified by a DOI. Next, these “preprints” are submitted to a publicly Web-based open peer review process and, when finally accepted, they are made available as a version of recorded articles with new DOIs. The highly prestigious British Medical Journal leads another remarkable adoption of OPR<sup>2</sup>. Reviews are signed by reviewers and published alongside the authors' article<sup>3,4</sup>. There is some disagreement in the research community about the pros and cons of OPR. Some of these are discussed in an editorial of Nature Neuroscience [OPR 1999], which highlights, in particular, the wariness on the part of editors that reviews might become bland and timid.

---

1 <https://f1000research.com/>

2 <https://bmjopen.bmj.com/pages/reviewerguidelines/BMJ>

3 <https://bmjopen.bmj.com/content/12/2/e054271.info>

4 <https://bmjopen.bmj.com/content/bmjopen/12/2/e054271.reviewer-comments.pdf>



Figure 8 - Types of Peer Review - extracted from <https://plos.org/resource/open-peer-review/>

## 2.5 Benefits to authors and their institutions

Science is a social endeavor, which means that ample communication of results is essential to foster the advancement of knowledge. This is the reason why scientific journals, meetings, conferences and libraries themselves were created. Open Access to the scientific literature may facilitate the communication among researchers, as long as the costs of providing such a kind of access do not prevent scientists from being able to publish, an essential condition for being a scientist, namely, communicating their results openly. As will be seen later in this report, the term “publication” of such results now also includes data and software, among others.

Thus, it is not difficult to imagine that facilitating access to the scientific literature may bring a boost to the advancement of science. It is not easy to design studies that can directly measure this effect, but some efforts exist, often based on meta-analyses of the literature.

The Open Citation Project followed the literature on citations obtained by comparable Open Access and non-Open Access articles. The Scholarly Publishing and Academic Resources Coalition (SPARC Europe) updated their study in 2015 and showed that, considering 70 articles on the subject, 46 demonstrated an advantage in the number of citations for Open Access articles, 17 found no effect and seven found a disadvantage for OA articles<sup>1</sup>.

A specific work [Swan 2010] viewed 31 studies in the literature and found that 27 found advantages for Open Access articles, while seven found no difference between Open Access or subscription articles. Swan was able to identify different behaviors in terms of increased (or diminishing) citations according to the field, as shown in Table 1.

<b>Size of OA citation advantage when found (and where explicitly stated by discipline)</b>	<b>% increase in citations with Open Access</b>
Physics/astronomy	170 to 580
Mathematics	35 to 91
Biology	-5 to 36
Electrical engineering	51
Computer science	157
Political science	86
Philosophy	45
Medicine	300 to 450
Communications studies (IT)	200
Agricultural sciences	200 to 600

Table 1. Variation in the number of citations received by articles published in OA, according to the field of knowledge. (Source: [Swan 2010])

More recently, Sotudeh [Sotudeh 2020] found that:

*“ ... there are significant gaps among the OA–NOA (Open Access–Non-Open Access) pairs dealing with highly similar subjects in all the OA models. However, the OA–NOA pairs with highly similar contents do not adhere to the finding. This means that the OACA (Open Access Citation Advantage) is not an artifact caused by different topics with various citation potentials for the OA and NOA papers. Nor is it associated with (dis)similarity in their publication factors, because the highly similar OA–NOA pairs were also detected to be significantly different, no matter if they are published in the same or different years, journals or document types. The higher citation performance of the OA in comparison with their subject-similar NOA pairs with different publication characteristics signifies that subject similarity is so powerful to counterbalance the effect of publication in different journals, publication time and document types.”*

<sup>1</sup> <https://sparceurope.org/what-we-do/open-access/sparc-europe-open-access-resources/open-access-citation-advantage-service-oaca/>

It is reasonable to think that Open Access also facilitates the translations of research results into formats readable and intelligible by lay people. In Brazil, the Bori<sup>1</sup> agency, following international experiences, is successfully operating a service that translates research for journalists to inform their readers.

## 2.6 Open Access policies

As the idea of publication under Open Access got traction in the world, many research funding and research performing organizations, particularly from developed countries, decided to implement Open Access policies to encourage researchers to choose the available open access options. Most policies for Open Access are not mandatory and include several loopholes to allow consideration of special cases that tend to arise in research (e.g., collaboration among scientists subject to different restrictions, different funding capability, or specific choices of venue for the publications, among others). The main justification for Open Access policies is to increase the visibility of research results and facilitate collaboration. In the case of public organizations, often the justification includes the right of taxpayers to access freely the research they funded. In all cases, funders and universities are wary of interfering too much with the researcher's choice of venue for the sharing of results, as this is considered not only a relevant element of academic freedom, but it is also necessary for the better advancement of science.

The Registry of Open Access Repositories<sup>2</sup> (ROAR) keeps a database of Open Access policies mandated by funders, universities, and research organizations. In July, 2022, the number of policies registered was 1,113 (see Table 2). Table 3 shows some details on the policy adopted in research organizations.

Funder	85
Funder and research organization	57
Multiple research organizations	12
Research organization e.g., university or research institution)	877
Sub-unit of research organization e.g., department, faculty or school)	82

*Table 2. Number of Open Access Policy mandates classified according to the type of institution. (Source: Registry of Open Access Repositories, ROAR, <http://roarmap.eprints.org/>, accessed in July 2022).*

1 Agência Bori - <https://abori.com.br/>

2 <http://roarmap.eprints.org/>

## 2.6.1 FAPESP's Open Access policy

In Brazil, the São Paulo Research Foundation (FAPESP)<sup>1</sup> was the first research funding agency to formally adopt an Open Access policy. Such a policy was approved by its Superior Council in 2008 and reinstated in 2019<sup>2</sup>. At the end of 2021, the policy was updated to establish the maximum acceptable embargo period (12 months) and additional clarifications. The policy states that:

*“... the full texts of articles or other types of scientific communication, originated from research and projects financed by FAPESP, partially or totally, and published in international journals must be deposited in an open access institutional repository of scientific papers, considering the policy of each journal, as soon as the manuscripts are approved for publication or within a period compatible with the restrictions of each journal, with the maximum embargo of 12 months after the publication date.”*

Thus, there are two basic requirements in FAPESP's policy:

- a) It requires that, once the researcher freely chooses the journal to publish the results, maximum use must be made of the opportunities allowed by the journal to make the work openly accessible, either the Author's Submitted Manuscript (ASM), the Author's Accepted Manuscript (AAM) or the Version of Record (VoR) of the journal published article. Thus, it does not interfere with the choice of the researcher about where to publish.
- b) It requires that the institution to which the researcher is affiliated implements an institutional repository of publications, including services to take care of all the tasks of maintaining and feeding the repository, verifying the standards for Open Access, once the researcher informs the DOI of the work and the necessary files. Thus, it does not impose any additional costs to the researcher.

In response to FAPESP's policy, most research universities and some research institutions in the state of São Paulo have implemented institutional repositories for publications. Examples are: USP<sup>3</sup>, Unesp<sup>4</sup>, Unicamp<sup>5</sup>, Unifesp<sup>6</sup> and UFSCar<sup>7</sup>. FAPESP's experience indicates that Open Access usage in Brazil could benefit from:

- research agencies (national, regional, and private) defining clear open access policies;

---

1 <https://www.fapesp.br/en>

2 <https://fapesp.br/12632/portaria-cta-no-012019>

3 <https://repositorio.usp.br/>

4 <https://repositorio.unesp.br/>

5 <http://repositorio.unicamp.br/>

6 <https://repositorio.unifesp.br/>

7 <https://repositorio.ufscar.br/>

- university and research institutions defining clear open access policies including the operation of article repositories;
- universities and research institutions commit their library services to support, assist and guide researchers on the use of Open Access possibilities to the full extent of the academic norms

<p><b>NIH</b> Policy Date: Apr. 2008 <a href="http://publicaccess.nih.gov/">http://publicaccess.nih.gov/</a></p>	<ul style="list-style-type: none"> <li>• Deposit of item: Required</li> <li>• Locus of deposit: Subject repository</li> <li>• Date of deposit: No later than the publication date</li> <li>• Content types specified under the mandate: Peer-reviewed manuscripts</li> <li>• Journal article version to be deposited: Author's final peer-reviewed version</li> <li>• Can deposit be waived?: No</li> <li>• Making deposited item Open Access: Required</li> <li>• Can making the deposited item Open Access be waived?: No</li> <li>• Date deposit to be made Open Access: By the end of policy-permitted embargo</li> </ul>
<p><b>ERC</b> Policy Date: Dec. 2014 <a href="http://erc.europa.eu/sites/default/files/document/file/ERC_Open_Access_Guidelines-revised_2014.pdf">http://erc.europa.eu/sites/default/files/document/file/ERC_Open_Access_Guidelines-revised_2014.pdf</a></p>	<ul style="list-style-type: none"> <li>• Deposit of item: Requested</li> <li>• Locus of deposit: Subject repository</li> <li>• Date of deposit: By the end of policy-specified embargo</li> <li>• Content types specified under the mandate: Peer-reviewed manuscripts, ETDs, Books, Book Sections</li> <li>• Journal article version to be deposited: Author's final peer-reviewed version</li> <li>• Can deposit be waived?: Not Applicable</li> <li>• Making deposited item Open Access: Requested or recommended</li> <li>• Can making the deposited item Open Access be waived?: Not Specified</li> <li>• Date deposit to be made Open Access: By end of policy-permitted embargo</li> </ul>
<p><b>NSF</b> Policy Date: Aug. 2013 <a href="http://www.nsf.gov/news/special_reports/public_access/">http://www.nsf.gov/news/special_reports/public_access/</a></p>	<ul style="list-style-type: none"> <li>• Deposit of item: Required</li> <li>• Locus of deposit: Subject repository</li> <li>• Date of deposit: By the end of policy-specified embargo</li> <li>• Content types specified under the mandate: Peer-reviewed manuscripts, Other</li> <li>• Journal article version to be deposited: Author's final peer-reviewed version</li> <li>• Can deposit be waived?: Not specified</li> <li>• Making deposited item Open Access: Not Mentioned</li> <li>• Can making the deposited item Open Access be waived?: Not Specified</li> <li>• Date deposit to be made Open Access: By the end of policy-permitted embargo</li> </ul>
<p><b>MIT</b> Policy Date: Mar. 2009 <a href="http://libraries.mit.edu/scholarly/mit-open-access/open-access-at-mit/mit-open-access-policy/">http://libraries.mit.edu/scholarly/mit-open-access/open-access-at-mit/mit-open-access-policy/</a></p>	<ul style="list-style-type: none"> <li>• Deposit of item: Required</li> <li>• Locus of deposit: Institutional Repository</li> <li>• Date of deposit: No later than the publication date</li> <li>• Content types specified under the mandate: Peer-reviewed manuscripts</li> <li>• Journal article version to be deposited: Author's final peer-reviewed version</li> <li>• Can deposit be waived?: Yes</li> <li>• Making deposited item Open Access: Required</li> <li>• Can making the deposited item Open Access be waived?: Yes</li> <li>• Date deposit to be made Open Access: Publication date</li> </ul>

Table 3. Examples of policies for Open Access. (Source: Registry of Open Access Repositories, ROAR, <http://roarmap.eprints.org/>, november 2020).

## 2.6.2 The European Commission's Plan S policy proposal for Open Access

Since September 2018, the cOAlition S<sup>1</sup> [Schiltz 2018], which includes an increasing number of research funders (almost all in Europe) under the auspices of the European Commission, has been promoting the challenging Plan-S, which demands that all publications from research they fund starting in 2021 must be available in full open access immediately upon publication<sup>2</sup>. In May 2019, cOAlition S launched the “São Paulo Statement on Open Access”<sup>3</sup> with the African Open Science Platform, AmeliCA, OA2020, and SciELO during the 8th Annual Meeting of the Global Research Council (GRC).

After a public consultation, Plan S' principles were revised in May 2019. There are three routes to Open Access accepted by Plan S:

- a) Open Access publishing venues (Gold OA);
- b) Subscription venues (repository route with zero embargo time);
- c) Transition of subscription venues (transformative arrangements) towards full OA mode — see, for example, [Schimmer et al 2015] for a brief analysis of transformative arrangements.

Plan S is summarized in ten principles<sup>4</sup>:

- 1) Authors or their institutions retain copyright to their publications. All publications must be published under an open license, preferably the Creative Commons Attribution license (CC BY), in order to fulfill the requirements defined by the Berlin Declaration.
- 2) The Funders will develop robust criteria and requirements for the services that high-quality Open Access journals, Open Access platforms, and Open Access repositories must provide.
- 3) In cases where high-quality Open Access journals or platforms do not yet exist, the Funders will, in a coordinated way, provide incentives to establish and support them when appropriate. Support will also be provided for Open Access infrastructures where necessary.
- 4) Where applicable, Open Access publication fees are covered by the Funders or research institutions, not by individual researchers. It is acknowledged that all researchers should be able to publish their work Open Access.

---

1 <https://www.coalition-s.org/>

2 <https://www.coalition-s.org/why-plan-s/>

3 <https://www.coalition-s.org/sao-paulo-statement-on-open-access>

4 [https://www.coalition-s.org/plan\\_s\\_principles/](https://www.coalition-s.org/plan_s_principles/)

- 5) The Funders support the diversity of business models for Open Access journals and platforms. When Open Access publication fees are applied, they must be commensurate with the publication services delivered and the structure of such fees must be transparent to inform the market and funders of potential standardization and capping of payments of fees.
- 6) The Funders encourage governments, universities, research organizations, libraries, academies, and learned societies to align their strategies, policies, and practices, notably to ensure transparency.
- 7) The above principles shall apply to all types of scholarly publications, but it is understood that the timeline to achieve Open Access for monographs and book chapters will be longer and requires a separate and due process.
- 8) The Funders do not support the “hybrid” model of publishing. However, as a transitional pathway towards full Open Access within a clearly defined timeframe, and only as part of transformative arrangements, Funders may contribute to financially supporting such arrangements.
- 9) The Funders will monitor compliance and sanction non-compliant beneficiaries/grantees.
- 10) The Funders commit that when assessing research outputs during funding decisions they will value the intrinsic merit of the work and not consider the publication channel, its impact factor (or other journal metrics), or the publisher.

It is important to note that Plan S and other OA approaches entail a cost burden to researchers (or their funders), which can be especially detrimental (it can be estimated in the million Euros range) to those who do not have access or support to cover APC prices. Plan S mentions an intention to develop guidelines for discounting or cost waving by publishers to benefit authors from low and middle-income countries, however it is not clear if or how this might happen.

## ***2.7 Free for readers, not for authors or their funders/institutions: the costs of Open Access***

Publishing scientific articles is carried out in very different settings, varying from those of the big publishers of hundreds and even thousands of journals, to small publishers with a small number of journals, or research units that run only one journal. From pure academic to a major industrial and commercial activity, the publishing of journals has more than three centuries of methodological and technological development. The transition of contents to digital support and online publication remodeled the field both technologically and economically, but most of the main publishing functions remain in the flow of article production: reception and evaluation, processing of the manuscripts submitted by the authors (editor’s analysis, obtaining reviews, sending reviews to authors with eventual communication with them, receiving the manuscripts back, analyzing the edits). When approved, manuscripts are copy-edited in general by interact-

ing closely with authors, typeset and formatted in XML for processing, as well as in HTML and PDF for screen-displaying and eventually for printing; the final version is stored in a way that is preserved and ready to answer user requests to be displayed and downloaded; the articles are also compliant with web interoperability; there may be marketing activities, help desk and other functions. These functions require physical and information infrastructure, personnel, and the hiring of external services.

The actions and procedures described above reflect in costs that vary from a few hundred to several thousand dollars per manuscript, as illustrated below:

- a) [Pavan and Barbosa 2018] found that, for the year 2016, the cost of APCs for articles (in journals charging non-zero APCs) with authors in Brazil was US\$ 1,039.
- b) The University of Cambridge reports<sup>1</sup> that, from 2013 to 2018, the average APC expenditure for 3,804 articles funded by the RCUK “block” grant<sup>2</sup> for the Office of Scholarly Publishing was £ 2,291 (US\$ 3,062) per article.
- c) The University College London (UCL) informs<sup>3</sup> that in the twelve-month period, from August 2016 to July 2017, a total of £3.3 million was paid (using funds from UCL’s RCUK, COAF and institutional open access funds) for 1,946 APCs, resulting in an average value of £ 1,704 (US\$ 2,275). This value is smaller than for the U. of Cambridge, mentioned in item (b) above, because here the institutional funds can only be used for APCs in Gold OA journals, which are normally smaller than those for hybrid journals.
- d) The OpenAPC database<sup>4</sup> informs that in 2019 the average APC cost for a set of 136 institutions was € 1,850 (US\$ 2,235). Considering only Gold OA, the average APC was € 1,642 (US\$ 1,984), while the APC for Hybrid journals was € 2,530 (US\$ 3,057).

---

1 Arthur Smith, “Cambridge Open Access Spend 2013-2018”, Unlocking Research (blog by the University of Cambridge Office of Scholarly Communication), October 22, 2018. <https://unlockingresearch-blog.lib.cam.ac.uk/?p=2219>, accessed on December 08, 2020.

2 This “block” grant covers only APC charges for publications resulting from RCUK funding.

3 Catherine Sharp, “UCL’s APC spend: an analysis”, OPEN@UCL BLOG, November 22, 2017. <https://blogs.ucl.ac.uk/open-access/2017/11/22/ucls-apc-spend-an-analysis/#comments>, accessed on December 08, 2020.

4 [https://treemaps.intact-project.org/apcdata/openapc/#institution/period=2019&is\\_hybrid=All](https://treemaps.intact-project.org/apcdata/openapc/#institution/period=2019&is_hybrid=All)

	APC (US\$)
<a href="#">SciELO<sup>1</sup></a>	300
<a href="#">OpenAPC database</a>	2,377
Collection of OA articles, Brazil, 2016	1,039
Univ. Of Cambridge, 2013-2018	3,062
Univ. College London, 08/2016-07/2017	2,275
<a href="#">OpenAPC database<sup>2</sup> (All cases, 2019)</a>	2,377
<a href="#">OpenAPC database<sup>3</sup> (Only Gold OA, 2019)</a>	1,984
<a href="#">OpenAPC database<sup>4</sup> (Only hybrid OA, 2019)</a>	3,052

Table 4. Article Page Charge (APC) cost for some collections of scientific articles.

How, then, to estimate the costs for flipping to OA the publications with authors in Brazil, without any time embargo? Table 4 summarizes the APC costs for some collections. This cost depends basically on two characteristics that may change for different types of publication: (i) the distribution of the publications between Gold OA and Hybrid OA (Hybrid tends to cost more); (ii) the visibility of the journals chosen by the authors to best communicate their results (there is a tendency for higher impact journals to charge higher APC) [Van Noorden 2013].

We choose to use the data from OpenAPC (last three lines in Table 4) to estimate the costs that would be incurred if one would want to flip all publications with authors in Brazil to Open Access. We start from the number of publications with authors in Brazil in 2018 and 2019: data from Scopus (see Figure 4) shows a total of 142,237 articles and reviews, of which 66,763 are Open Access. Thus, the task would be to flip to OA a total of 75,474 articles (published in two years), giving an average of 37,737 publications per year.

	Qty public.	Value (US\$)		
		Average	Gold OA	Hybrid OA
Publications 2018-2019	142,237			
Of which, OA Gold	66,763			
Remaining to be flipped	75,474			
Per year	37,737	Average	Gold OA	Hybrid OA
Open APC value for APC (US\$)	37,737	2,235	1,984	3,057
Cost to Brazil (in US\$ per year)	37,737	84,352,018	74,868,116	115,357,085

Table 5. Estimate for the cost of moving Brazil's publications which are not OA to OA.

1 Packer, AL (2021). Programa SciELO/FAPESP de Ciência Aberta – estado de avanço em 2021 [Power Point slides] <https://scielo.figshare.com/ndownloader/files/31698779>

2 [https://treemaps.intact-project.org/apcdata/openapc/#institution/period=2019&is\\_hybrid=All](https://treemaps.intact-project.org/apcdata/openapc/#institution/period=2019&is_hybrid=All)

3 [https://treemaps.intact-project.org/apcdata/openapc/#institution/period=2019&is\\_hybrid=FALSE](https://treemaps.intact-project.org/apcdata/openapc/#institution/period=2019&is_hybrid=FALSE)

4 [https://treemaps.intact-project.org/apcdata/openapc/#institution/period=2019&is\\_hybrid=TRUE](https://treemaps.intact-project.org/apcdata/openapc/#institution/period=2019&is_hybrid=TRUE)

Table 5 shows the memory for calculating a cost estimate. The yearly cost is estimated between approximately USD\$ 75 million and UDS\$ 115 million, depending on the balance of the collection between Gold OA and Hybrid OA. We remark that other characteristics of the collection of publications affect the cost, such as international visibility of the journals chosen as venues. It should also be noted that this is the additional cost, compared to the situation in 2019, since for that year almost 48% of the publications with authors in Brazil were already OA — and their costs, when necessary, have been covered from several sources of funding.

The point of this estimate is to alert that moving towards OA without any embargo for publications on a national scale will bring additional costs to the research system, especially considering that the subscriptions for journals would need to have their costs covered, so that researchers can access their contents. And while there is a large collection of articles published in subscription journals, there is a cost to read the articles and a cost to have the articles published. The embargo time delay is the key point here. As shown above, it is possible to have OA policies that have zero additional cost, as long as it is accepted to have the time delays resulting from the embargo conditions established by publishers (presently, in most cases embargo times are shorter than 12 months, and even shorter when there are mandates defined by funders or research institutions).

## 3 Open Data

In 2020, the Brazilian Academy of Sciences published a report on Open Data [ABC 2020], which presents a broad panorama of the state-of-the-art, challenges, opportunities and directions. This section on Open Data complements that report by presenting additional issues, but is by no means exhaustive.

Already in 2012, there was enough experience and evidence in the advantages of sharing open data in science that allowed The Royal Society in the UK to publish a report on “Science as an open enterprise” [Boulton et al 2012]. This 100-page report concentrates on open data and data sharing as a means to advance scientific discovery, and the use of computational and communication technologies as “new ways of doing science”. It emphasizes the importance of linking publications to the underlying (open) data that was used to conduct the corresponding research, and how data have become an integral part of research efforts.

More recently, in 2021, the OECD Council updated a legal instrument of 2006 with recommendations concerning Access to Research Data from Public Funding [OECD 2021]. This updated policy covers among others, data, metadata, algorithms and software (see Section 4). It provides guidance in seven areas — data governance for trust; technical standards and practices; incentives and rewards; responsibility, ownership and stewardship; sustainable infrastructures; human capital; and international cooperation for access to research data. Many of these topics are discussed in this section.

### 3.1 What is data?

There are many definitions of “data”. We adopt Merriam Webster’s definition, which considers that “data is anything we can capture via any sensing device or organ” (and thus human senses). This being said, we recall once more that we are concerned only with digital data. As such, even though we recognize non-digital objects (e.g., paper notebooks, biological samples, archeological findings) as research data, this section deals exclusively with digital renderings of such objects.

Indeed, one given physical object can correspond to many digital objects, each of which represents a particular digital view thereof. For instance, a rock can be represented digitally by a set of photos, by a table describing its physical and chemical properties, or by a computer specification that will allow reproducing that rock in a 3D printer. By the same token, a paper field notebook (e.g., with notes written by linguists or anthropologists) can be digitized and transformed into a set of photos, or into pdf files, or hypertext in which notes are linked to related digital objects. Also, the notebook’s main facts can also be transported into an Excel spreadsheet. Thus, the rock (or any physical object) can be “virtualized” into a set of distinct digital representations of that single object, each of which may require distinct storage standards, metadata documentation, and preservation policies.

## 3.2 Data life cycles and practices

Good data practices require that those digital representations be appropriately managed throughout the so-called “data life cycle” — namely, all stages through which any data item passes from the time it is “born” (i.e., created/collected) to the moment it “dies” (e.g., deleted or forgotten). Digital death of a data item, and how to avoid it, refers to the disappearance of that object — for instance, it may still exist but is no longer findable because all references to it are no longer valid, or it may disappear because the medium in which it is stored is destroyed through digital decay. The data life cycle is composed of a sequence of stages, each of which is associated with a set of good management practices, including appropriate documentation. Roughly speaking, there are two conceptualizations of the data life cycle: the cycle as perceived by researchers and as perceived by *data librarians* and *stewards* (the professionals who interact with researchers to train and help them to document and format their data for publication, and subsequently perform some data management actions leading to archival and preservation). Each of the stages in this life cycle involves distinct data management practices and different skills.

Figure 9 illustrates these two conceptualizations. Researchers are specifically concerned with the lifecycle stages related to their activities, whereas librarians and stewards consider data as objects to document, maintain and preserve. Data librarianship may also involve ensuring authenticity, e.g., “chain of custody” (that is related to, but goes beyond, ensuring reproducibility). Scientists mostly deal with the so-called “active data”, which are directly connected with their ongoing research, as opposed to archival data, which are not always available online, but can be retrieved from archives, also called “data vaults”. As such, the data life cycle has roughly the following stages:

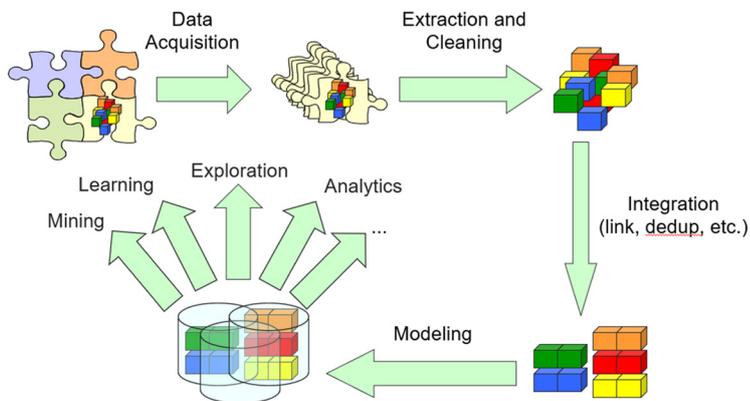
- For researchers: data collection, data curation, data storage, data analysis and visualization;
- For librarians and stewards: data collection, data curation, data cataloging/documentation, data storage, data preservation/archiving.

Sound open data good practices are subject to the following general principles:

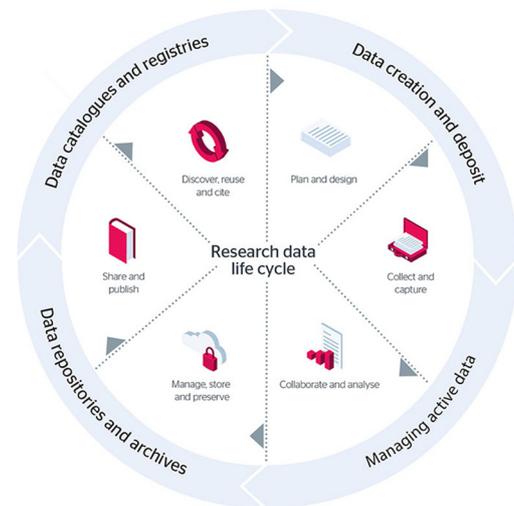
- a) Open Science by Design [NAS 2018]. Open science practices require planning one’s research for sharing its results. Thus, all stages of the data life cycle must be planned beforehand when designing a research project, to consider appropriate documentation, traceability and reproducibility. In many institutions in Europe, Australia and North America, data librarians are trained to help researchers in preparing research outputs for openness
- b) Specifying and maintaining a Data Management Plan (DMP). This is a written text that documents the plan for managing data throughout its life cycle — see Subsection 3.4
- c) Documenting data via metadata records. *Metadata*, or “data about data”, provides basic information on a file, to allow it to be cataloged and found using software. Typically, metadata records provide basic information on a file’s contents to support findability (e.g., keywords, textual description) and, depending on the domain, additional facts on how the data were collected, when, where, and by whom. Notice that scientific articles and software also require metadata for

**The Big Data Pipeline. Source: Altigran Soares da Silva, Keynote speech at the 2018 Brazilian Workshop on Social Network Analysis and Mining**

**Nem Sempre se vê Mágica no Absurdo - Engenharia de Dados e Ciência de Dados**



**Research Data LifeCycle - source JISC website, (<https://www.jisc.ac.uk/>), captured in June 2018**



*Figure 9. Research data lifecycle as portrayed by computer scientists that conduct research on data management (left) and by data librarians and repository curators (right). Sources: 9A - The Big Data Pipeline. Altigran Soares da Silva, Keynote speech at the 2018 Brazilian Workshop on Social Network Analysis and Mining. Nem Sempre se vê Mágica no Absurdo - Engenharia de Dados e Ciência de Dados. 9B - Research Data LifeCycle - JISC website (<https://www.jisc.ac.uk/>), captured in June 2018*

the same purposes — such as title, authors' names, abstract, or keywords. As such, the creation of metadata records is essential to ensure documentation of any digital artifact in Open Science. There is a plethora of domain specific metadata standards, and very few consensual ones. The Metadata Standards Catalog<sup>1</sup>, maintained by a working group of the Research Data Alliance, is an example of the variety and complexity of such standards.

d) Preserving and archiving data in trusted institutional repositories. Data are valuable research assets, and archival means that they will be adequately maintained and preserved for long periods of time, for future reuse. This demands that institutions not only provide the appropriate e-infrastructure (namely, hardware and software), but also the technical staff that will ensure that repositories are appropriately created and maintained. To this purpose, some repositories undergo periodical certification by third parties. For instance, the CoreTrustSeal certification verifies compliance with 16 requirements, such as level of curation performed, accessibility, security, governance mechanisms, and compliance with international standards [CoreTrustSeal 2020]. See section 3.6 for a discussion on repositories, and on a site that has an extensive list of over 3,000 repositories and archives that follow some or all of these recommendations and can be thus queried to find an appropriate repository for data deposit if institutions do not make them available.

<sup>1</sup> <https://www.rd-alliance.org/groups/metadata-standards-catalog-working-group.html>

### 3.3 Defining Open Data

While all agree that open access to scientific literature facilitates communication among researchers, a relatively new concept is that open data also facilitates this communication, since scientists often communicate through data exchange. Given the discussion in the previous section, Open Science grapples with the needs of openly sharing data but at the same time ensuring that specific constraints are met — such as privacy of individuals, or species protection — see section 5.3 (Ethics, Privacy, and Security) for overall concerns, not specific to data. As such, the notion of “open data” is often stated as “as open as possible, as closed as necessary” [Landi et al 2020]. This, in turn, means that even when data cannot be openly shared, *metadata records must be open and available to all*. This raises three specific issues: the notion of FAIR data, metadata standards, and DOI for data.

**FAIR data** — The term “FAIR” in the context of good data management practices was coined in 2016 [Wilkinson et al 2016] to denote that data must be “Findable”, “Accessible”, “Interoperable” and “Reusable”. Findability is usually implemented by two mechanisms: metadata records (which must themselves be FAIR) and a unique identifier (also known as PID, or Persistent ID) that is permanently associated with a file. PIDs play the same role as DOIs for scientific articles. Accessibility requires that, once found, data can be accessed, though under limitations imposed by access constraints. Interoperability and reusability mean that data must be stored in standard, non-proprietary formats, so that they can be referenced by and reused by arbitrary documents and software, themselves subject to standards. In particular, *reusability* means that a given data set can be reused for purposes different from the ones it was intended for.

**Metadata standards** — There are many standards for documenting digital data, which depend on the scientific domain, and the groups involved in a project. For example, astronomical data require a standard different from biodiversity observations, or archeological records. It is up to the researcher to choose the standard that best suits her/his data. Nevertheless, all standards require a common set of compulsory metadata information, such as the description of a dataset, where it can be found/accessed, associated keywords, among others. Other information depends on the discipline — for instance, many standards require the geographic coordinates of where the original data was collected or the characteristics and type of sensing instrument. Metadata information can also associate the digital world with the original physical sample. To this purpose, many domains are establishing catalogs of physical object identifiers (for instance, chemical compounds, or rocks).

**DOI** — Just like publications, datasets are now citable, and even indexed as full-fledged publications by, e.g., Web of Science or Google Scholar. Like identifiers of publications, data DOI are persistent identifiers used to uniquely identify a data set. The rules for creating a DOI for datasets are the same as for publications, consisting of a string of digits that specify, e.g., where the dataset was registered and stored. DataCite<sup>1</sup>, a global non-profit organization, is one of the most widely used Data Registration Agencies, providing DOIs for research data and other research

---

1 <https://datacite.org>

outputs. Such DOIs are intrinsically linked to FAIR principles, supporting Findability. DOIs are a specific case of Persistent IDs (PID), which are being proposed and implemented to uniquely identify not only digital objects, but also physical ones, such as physical samples or instruments [Plomp 2020]. Thanks to this, some journals now support linking descriptions of materials to their PIDs, which in turn allow the reader to, e.g., see a chemical compound structure, or directly access that compound in international catalogs. Enhanced data PIDs allow the identifier to embed data versioning or privacy and security checks.

### **3.4 The role of Data Management Plans**

A data management plan (DMP) is now considered to be indissociable from any research proposal and part of good research practices. It answers three basic questions: (1) which data will a project collect, produce and share; (2) how these data will be managed during the project; and (3) once the project is finished, how, when, where and for how long these data will be shared, under constraints such as intellectual property, guided by ethics, privacy and others.

Data management plans often change as the research evolves. Their use is twofold — document data practices throughout the research life cycle and support preparation for sharing. Data management plans are applied to digital data. Physical data can be part of collection activities mentioned in such plans, but subsequent activities are restricted to the digital realm. The underlying hypothesis is that the maintenance of physical samples is under the responsibility of scientists and curators of physical collections, whereas digital data management is very much a collective endeavor.

Data management plans have become a compulsory part of research proposals in most of Europe, North America and Oceania, as well as in some Asian countries and South Africa. One of the most widely online systems to help researchers prepare a DMP is DMPTool<sup>1</sup> and its European equivalent DMPonline<sup>2</sup>. These tools have spawned many similar online tools that guide researchers through all steps necessary in writing a plan, which can be tailored to specific funders, research domains, and even funding lines. In Brazil, DMPs are compulsory at FAPESP for submitting any research proposal, and their compliance is checked by reviewers when analyzing project reports. As a consequence, the DMPTool site, maintained at the University of California with NSF funding, reports that researchers based at the state of São Paulo are the second largest community to use the site, after those from the US.

---

1 <http://dmptool.org>

2 <http://dmponline.dcc.ac.uk>

## 3.5 Open Data in Brazil

The Brazilian Academy of Sciences published a report on Open Data policies and practices, including a discussion of some Brazilian initiatives [ABC 2020]. Among those, one can point out activities undertaken by institutions such as Embrapa or FIOCRUZ that are engaging in a set of long-term policies towards data sharing that include the establishment of open data repositories, with specific embargo and publication rules. IBICT — the Brazilian Institute of Information in Science and Technology (Instituto Brasileiro de Informação em Ciência e Tecnologia) — has launched, through its OASIS<sup>1</sup> open access portal, a metadata facility that is harvesting metadata from open research data resources throughout the country.

Perhaps the largest and most comprehensive open data initiative in Brazil is the one undertaken at the State of São Paulo, under the auspices of FAPESP. Launched in 2017, it covers three facets — the funder, research institutions, and researchers themselves. Under the principle that the results of research funded by public money are a public asset, FAPESP established norms for open access (see section 2) and open data, including compulsory Data Management Plans upon submission of any research projects, in all modalities. This, in turn, has prompted all 7 public academic institutions in the state to design and create a network of open research data repositories. Officially inaugurated in December 2019, this network took three years to be constructed, under a loose federation model, in which each institution manages its own data and exports all metadata to a single interface. In the process, each institution established its own open data governance and policies, e.g., for data deposit, use of standards and others. Researchers from these institutions, in turn, have the institutional support and infrastructure to share their data. These repositories are being progressively used, as researchers get trained by their institutions and familiarize themselves with the Open Data-Open Science movement. Thanks to the availability of this network, FAPESP was able to launch in July 2020 the COVID19 DataSharing/BR<sup>2</sup> open data repository, which was being fed with clinical and demographic data by several Brazilian health institutions. The design and launching of this COVID repository took only 15 days because the repository network itself had been designed for openness — a good example of “open by design” network.

By mid-2022, Brazilian universities outside São Paulo state were already beginning to create their open data repositories, usually under the governance of committees that have representatives from researchers, ICT professionals, and librarians. All realize this is a long-term initiative, that will require training and culture change. Also in mid-2022, CNPq (the Brazilian National Council for Scientific and Technological Development) launched the Lattes Data repository<sup>3</sup>. Still in a preliminary version, its long-term goal is to store all data produced by research financed by CNPq in an open and curated manner.

---

1 <https://oasisbr.ibict.br/>

2 <https://repositoriodatasharingfapesp.uspdigital.usp.br/>

3 <https://lattesdata.cnpq.br/>

## 3.6 Trusted data repositories

Data sharing requires preparing the data for reuse by others. This, in turn, covers several processing and documentation stages, as well as intervention of distinct actors. Sharing demands checking data quality (and thus the so-called curation), using appropriate data formats and adequate documentation (of which metadata are an essential part). Moreover, to practice Open Science, formats must be open and non-proprietary. In some specific cases, data may be made available together with the software needed to open and/or process them. Many (if not all) of these data preparation actions require participation of domain experts who will ensure compliance with quality standards (for the specific scientific domain or project context) and define domain-specific metadata elements. Such actions may also be assisted by data librarians of an institution.

Preparation and documentation are necessary for sharing, since researchers must make sure that others can understand the data. However, this is not enough. An important question is where to make data available. This is where repositories and preservation play a major role. Sharing data for research purposes requires planning where the data will be made available, to maximize its visibility and appropriate maintenance. This means that data must be stored in institutional repositories that are maintained using sound information technology practices and processes — and this includes, among others, ensuring data integrity, security and preservation.

Integrity and security in data repositories means that access is controlled, and only people with the adequate rights can update the data (while at the same time allowing ample reading and download permissions). This is not only a matter of setting adequate passwords; the storage space occupied by the data should be configured to increase multiple accesses and optimize network usage, so that all those that want to get the data can do this in a timely manner. Integrity also involves periodical backups, adequate indexing and often using dedicated storage software. Figure 10, reproduced from [Gibney and Van Noorden 2013], illustrates the need for data preservation. They report on a study that shows that 80% of the data used in articles is no longer available after 20 years, with serious consequences for, e.g., reproducibility.

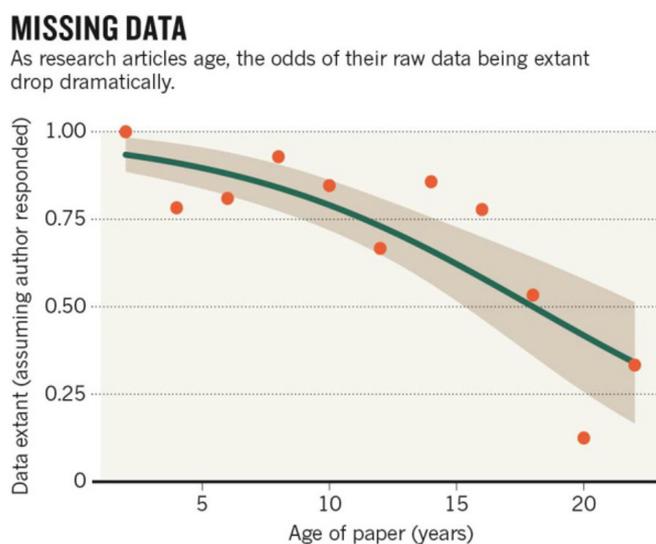


Figure 10. Data used in scientific papers disappear with time because of the lack of appropriate preservation practices — reproduced from [Gibney and Van Noorden 2013]

Preservation means that the data will continue to be available after a project ends. Standard practices require a minimum of 10 years' availability, but that depends on the data, research domain, and available resources. Since storage technology evolves, and storage support decays (e.g., becomes "electronic dust") backups must also be performed to ensure that the media will still be readable in the long run. Besides backup, preservation strategies may involve the so-called "mirror installations", in which exact copies of the data are kept in case of permanent hardware failures.

Thus, while scientists may be actively involved in preparation for sharing, they seldom have the time or resources to afford to take care of integrity, security and preservation tasks. Examples of inappropriate choices for making data publicly available are the researcher's own equipment, the researchers' laboratory or group (unless it maintains an appropriate repository), the researchers' cloud or disk space in the institution or web pages of the project that point to datasets. Appropriate repositories include, among others, those maintained by the researcher's institution (the so-called institutional repositories), by coalitions of institutions, by governments, or funders. It is not so much that researchers cannot be trained to "professionally" store and maintain their data — rather, this is a support activity, essential to Open Science, and that represents full time involvement of staff. The issue of open data, open repositories, and different policies in data archival as well as their costs are subject of several OECD studies (such as [OECD 2017, OECD 2020]).

Figure 11, reproduced from a report from the US National Academies of Science and Engineering [NAS 2020], gives an overview of the three main states for data storage handling within a research environment. This figure is used throughout that report to help estimate data storage, preservation, and archival costs. Though conceived for the Life Sciences, this framework can be extended to digital data management in any research context, always keeping in mind that research activities require frequent data transformations. Transitions between states are bidirectional and can occur at any time and order.

State 1 portrays the "primary research and data management environment, where data are captured and analyzed [NAS 2020, Chapter 2]. Here, researchers are the main responsible for data management, and thus users belong to a more restricted group, and access is restricted to the group. State 2 portrays activities necessary to deposit the data in a repository for sharing, and involves among others, concerns with documentation, security and privacy. This state deals with a so-called "active repository", in which data are easily accessible, with frequent publication activities from multiple users, potentially from many domains. State 3 involves long-term archival, where data are not expected to be frequently updated, and immediate online access is not a priority. It is often the case that active repositories and archival repositories are physically independent.

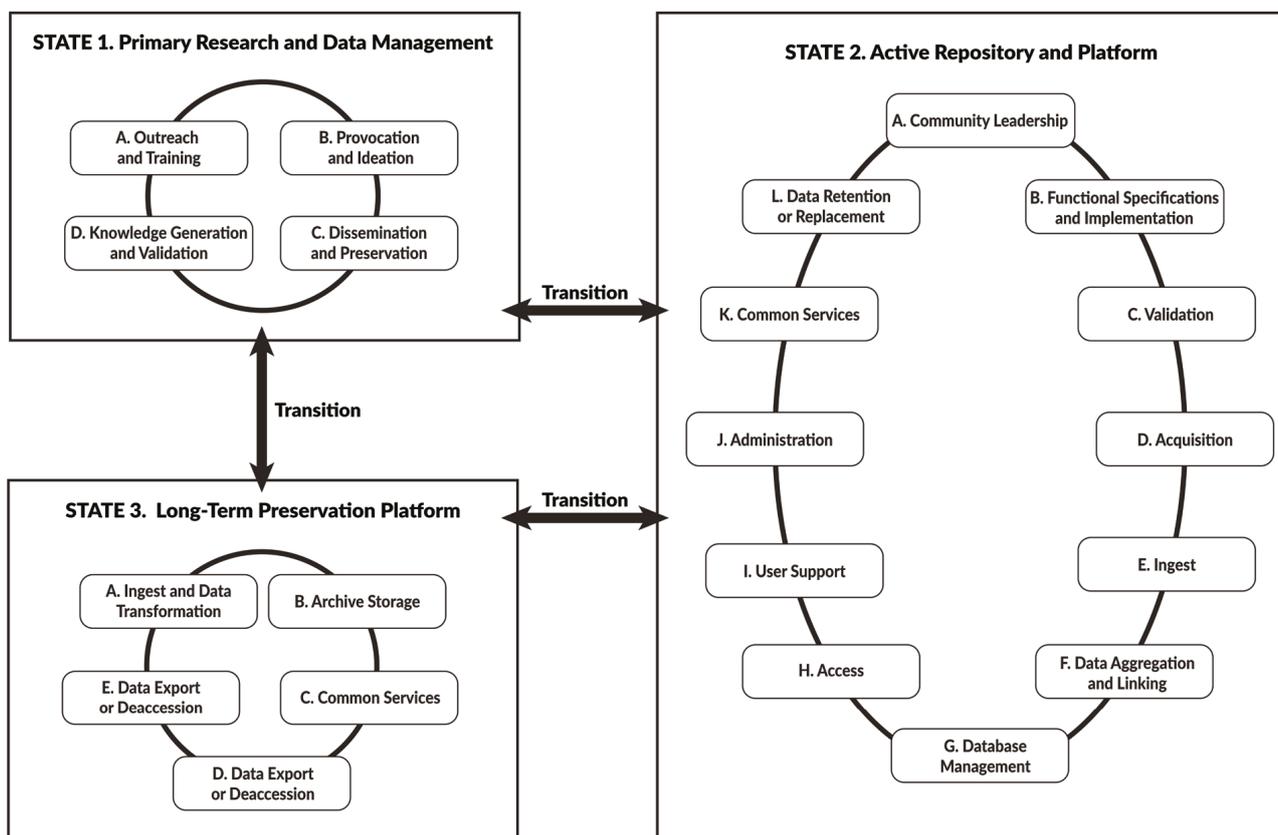


Figure 11. Conceptual diagram extracted from [NAS 2020] showing the three data states and bidirectional transitions among states. Each state portrays a given stage in a research environment, with distinct data management responsibilities and procedures, as well as different hardware and software considerations.

Given these considerations, in which repository should a researcher publish research data for active and archival purposes? It depends on the research domain, and the institutions that participate in the research. Repositories can be domain-specific or generalist (accepting data from all domains) and may be maintained by one institution or a coalition of institutions.

Domain-specific repositories: In many research fields, e.g., -omics or some domains in biodiversity or environmental science, there are consensual repositories in which all researchers deposit their data, and which accommodate both States 2 and 3 of Figure 11. In many cases, researchers must deposit their data in such repositories following international documentation, formatting and metadata standards. GenBank<sup>1</sup>, maintained by the NIH US funding agency is an example of authoritative data repositories in -omics, in which researchers all over the world can deposit their data. By the same token, ICPSR — Inter-university Consortium for Political and Social Research<sup>2</sup>, hosted at the University of Michigan, is dedicated to qualitative and quantitative data in the social sciences and maintained by a coalition of 750 institutions. Examples of domain-specific institutional repositories in which only their researchers can deposit data include, e.g., the UCLA Social Sciences Data Archive<sup>3</sup>.

1 <https://www.ncbi.nlm.nih.gov/genbank/>

2 <https://www.icpsr.umich.edu/web/pages>

3 [https://dataverse.harvard.edu/dataverse/ssda\\_ucla](https://dataverse.harvard.edu/dataverse/ssda_ucla)

Generalist repositories accept data from all domains. Examples are the Open Science Framework<sup>1</sup>, FigShare<sup>2</sup>, and Zenodo<sup>3</sup> that allow researchers from all over the world to use their resources to publish their data. These three repositories are free of charge up to a certain level of storage, but researchers can also pay to use their facilities if these levels are surpassed.

Perhaps the most useful source of information on where to deposit data is Re3data [Pampel et al 2013], a global registry of research data repositories that covers a list of thousands of repositories from different academic disciplines. Re3data<sup>4</sup> is maintained by a few German research institutions and by the University of Purdue; it is recommended by many publishers as a good source of information for authors to find an appropriate repository to permanently store their research data. Each entry is verified by the registry curator, in Purdue, who checks the repository's policies, metadata standards, data sharing mandates and data preservation procedures. Only after these items are considered acceptable, the repository is included in the registry, together with a considerable amount of additional information, including instructions on how to contact its managers for data deposit. The growth and evolution of Re3data's entries is a good indicator of the dissemination of Open Data practices all over the world. Started in 2012 with less than 300 entries, its thousands of repositories are now housed by institutions or research groups in more than 100 countries. By the end of 2022, it contained almost 2,500 domain-specific repositories. To register a repository in Re3data, its stewards must apply for it, providing all kinds of information that is transformed in metadata to index the repository — such as data publication licenses supported, preservation policy, curation processes, and responsibility for its institutional trustworthiness. Only verified repositories are authorized to be indexed by Re3data.

In Brazil, several institutions have already organized repositories for public data sharing, each under specific policies. Both EMBRAPA and FIOCRUZ are examples of these initiatives. The universities that are part of the state of São Paulo's public research data network have created generic data repositories for their research staff, in particular USP, UNICAMP, UNE-SP, and UFSCAR. Throughout 2021, many other Brazilian universities and research centers started designing and deploying their research data repositories. In the same year, SciELO launched its data repository program (SciELO Data).

This brings us to the subject of repository certification and TRUST principles [Dawei et al 2020]. Perhaps the most important guiding principle towards data sharing infrastructures is not a technical one — rather, sharing presupposes trust (that, e.g., the shared object will not be misused, or corrupted, and that appropriate credit will be given to those responsible for creating and maintaining it). TRUST is an acronym of 4 principles towards sound repository construction and management — Transparency, Responsibility, User focus, and Sustainability.

---

1 <https://osf.io>

2 <https://figshare.com>

3 <http://zenodo.org>

4 <https://re3data.org>

Under the TRUST principles, repositories must be created with openness in mind, and being under a top-level management committee, composed of scientists, administrative and technical staff, ideally from multiple institutions. In the coordination of such repositories, there must be transparency in the rules (e.g., which files should be deposited) and policies, with focus on the end-users (researchers who deposit the data and people who reuse the data).

### **3.7 International open data bodies – RDA and WDS**

By the first decade of the 21st century, funders and research institutions had started to acknowledge the value of Open Data and repositories. Countries such as Australia or the UK created government-funded institutions whose goal was to organize open data management practices, and open repositories, in their respective research and academic institutions. Such institutions have trained thousands of researchers and research staff (e.g., data librarians and curators, support staff) in the best practices of open data management. Also, some research funders (e.g., the National Science Foundation<sup>1</sup> in the USA) started to create specific data policies that had to be met by submissions, in particular compulsory Data Management Plans. For more details on the history of such initiatives, please check the report to the Brazilian Academy of Sciences [ABC 2020].

In parallel, two large multinational open data coalitions appeared — the RDA (the Research Data Alliance)<sup>2</sup> and the WDS (the World Data System)<sup>3</sup>. The first is a grassroots organization created in 2013 “with the goal of building the social and technical infrastructure to enable open sharing and re-use of data.” The second is an interdisciplinary body of the International Science Council<sup>4</sup>, geared towards institutional research data services and was launched in 2008. Both RDA and WDS regularly promote events to disseminate open data and data sharing practices and increase networking.

In more detail, RDA is a community-driven effort of all actors involved in open research data management — practitioners, researchers, producers, users and funders — for all stages of the data lifecycle. By the beginning of 2023, it had over 13,000 members from 145 countries mainly organized along interest groups and working groups. These groups produce recommendations and reports that cover a wide range of subjects along e.g., “social hurdles on data sharing, education and training challenges, data management plans and certification of data repositories, disciplinary and interdisciplinary interoperability, as well as technological aspects”. Affiliation to the RDA is open to individuals and organizations that adhere to its principles of openness, consensus, community-driven, inclusivity, harmonization, non-profit and technology-neutral.

---

1 <https://www.nsf.gov>

2 <https://www.rd-alliance.org>

3 <https://www.worlddatasystem.org>

4 <https://council.science>

Some of the standards produced by its groups have been adopted by, for instance, the United Nations' Food and Agriculture Organization<sup>5</sup>. It is recognized by the European Union in its connection to the European Open Science Cloud. Besides its many groups, it also hosts the Funders Forum, which congregates funders from all over the world in discussions about open data policies and data sharing.

WDS, on the other hand, is geared towards enabling and fostering trusted data services for science, towards “*universal and equitable access to scientific data and information*”. Concerned with trusted data services, and quality-assured data, its members are organizations that are responsible for the management of large data repositories — e.g., the Australian Antarctic Data Centre, the Chinese Astronomical Data Center, the Digital Repository of Ireland or the Oak Ridge National Laboratory — to name a few. It promotes networking across its members, and certification of their repositories, covering a wide spectrum of domains, such as oceanography, geophysics, biodiversity, astronomy or the social sciences.

---

5 <https://www.fao.org/home/en>

## 4 Open Source Software

Open Science requires that the tools and instruments necessary for the scientific practice be highly available to scientists throughout the world so that experiments can be reproduced, and results verified by third parties. In particular, in this century, computer software has become the most ubiquitous form of tool used by scientists. Software is now a fundamental component of the toolset used by a wide variety of sciences including not only Computer Science but also Physics, Chemistry, Biology, Engineering, etc. It is also increasingly used in the Social Sciences and Humanities. Software is used for data cleaning, data processing, data visualization, as well as for creating models and carrying out predictions. Specialized algorithms are coded in the form of libraries, scripts, and accompanying metadata. Without sharing all these artifacts, it is impossible to reproduce scientific research and validate its correctness.

Thus, to fully achieve openness in science, it is essential that the software libraries, scripts, packages, and applications used and developed by scientists be readily available to any researcher interested in reproducing or extending a certain piece of research. When it comes to publicly funded research, it is highly desirable that related *Research Software*, that is, all the software produced by a research initiative, be Open Source to ensure that its benefits are made available to the widest possible community.

### 4.1 Free/Open Source Software

The Free Software Movement was initiated in 1983 when the GNU project was founded by Richard Stallman as a means to develop a robust body of software, including an operating system and associated tools and applications that would be freely available to anyone interested. In 1998, the term “Open Source Software” was adopted by a part of this movement as a strategy to refer to this new, collaborative way of developing and sharing software with commercial companies. Open source software is now highly successful, adopted by the major IT companies in the world, which both use and produce software distributed under open source licenses. Nowadays, thousands of high-quality software components are available as open source, covering all layers of the software stack, from low-level drivers and operating systems to high-level libraries, applications, and frameworks. Within this context, two concepts have emerged – free software, and open source licenses, which go hand in hand.

**Free Software** — The documentation published by the Free Software Foundation<sup>1</sup> defined Free Software as the one that gives its users four freedoms: (1) to run the software for any purpose, (2) to study and modify the software (access to the source code is a precondition for this), (3) to redistribute exact copies of the software, and (4) to distribute modified versions of the source code<sup>2</sup>.

---

1 <https://www.gnu.org/philosophy/free-sw.html>

2 <https://www.gnu.org/philosophy/philosophy.html>

These four freedoms leverage the principles of Open Science. For instance, *permission to run* the software for any purpose allows researchers to reuse existing software without having to buy or build it from scratch to perform their own studies. *Permission to study and modify the source code* supports reproducibility and replicability by disclosing research software and related artifacts. It also increases transparency (visible workflows), auditability, and reliability (results can be verified by third parties and anyone can detect and correct a bug or a malicious feature). *Permission to redistribute copies* enables sharing replication packages, which, in addition to the raw data, provide the code necessary for their analysis and interpretation in different settings. Finally, *permission to distribute modified versions* of the software enables researchers to develop their own work by reusing and expanding someone's workflow, codebase, or tool, and to share the new knowledge for the benefit of others.

**Open Source licenses** — This type of license was defined to comply with the aforementioned Free Software Definition and, as such, they also leverage Open Science by providing a safe legal framework for sharing. To be classified as Open Source Software, a piece of code must be distributed under a license<sup>1</sup> formally approved by the Open Source Initiative<sup>2</sup>, preferably one that is labeled as “popular, widely used or with strong communities”. For instance, the GNU General Public License<sup>3</sup> (GPL), written by Richard Stallman in 1989, is a popular, widely used open source license.

## 4.2 Software life cycles and practices

Successful Open Source projects adopt a development model that results in high-quality code that gets adapted quickly to different situations. Several established practices are used and recognized as important contributions to the maintainability of high-quality software. These include the use of public repositories, version control, collaboration among peers, code review, automated testing, adoption of standard formats and interfaces, and good documentation.

Open source is often described as developed under the so-called “Open Source software development workflow”, which is driven by means of open repositories, and hosted by version control systems. The latter allow multiple versions of the same software to coexist, and possibly be referenced by research experiments and articles. Open repositories (see section 3.5 on data repositories) can be used for sharing several kinds of digital research artifacts, such as algorithms, data, code, reports, and workflows, supporting reproducibility, reducing redundancy, and promoting open scientific collaboration. While the repository is often publicly available for reading and downloads, access for modifications and uploads is restricted to a limited group of developers selected by meritocracy.

---

1 <https://opensource.org/licenses>

2 <https://opensource.org/>

3 <https://www.gnu.org/licenses/old-licenses/gpl-1.0.html>

Collaboration among peers is continuous and feedback is frequent via shared access to the source code and multiple communication channels such as forums, mailing lists, and IRC (a real-time chat used by software developers since the late 1980s). Code review, a very common community practice (“given enough eyeballs, all bugs are shallow”), fosters enhanced software quality by means of sharing, collaboration, and peer review and can be applied to other research assets. Automated testing increases reliability, and the use of standards promotes easier integration with other software. Constant and continuous documentation is a recommended practice to keep user guides, manuals, and other relevant documents updated with respect to the latest version of the software. Thus, best practices used in open source software development can be thought of as good practices for Open Science in general, since they seamlessly support availability, peer collaboration, workflow transparency, reuse, and reliability.

### **4.3 Open Source and Research Software**

Software is a major element of science in the 21st century and Open Source Software is a fundamental means of achieving Open Science in this context. Research software is required to be readily findable, accessible, interoperable, and reusable; Open Source Software is the first step in achieving that.

Open Source Software can be searched and retrieved from repositories based on identifiers and descriptors, using different criteria such as keywords, programming language, software version, among others. Such criteria are encoded into the metadata records that describe a given set of software modules. Accessibility is encouraged in Open Source Software by using open repositories and providing explicit, well-defined sharing licenses and sufficient documentation. The definition of programming interfaces and input/output formats and use of standards may promote interoperability and also reusability. Within the Open Science ecosystem, Open Source Software should be citable, sustainable, and recognized as a valuable research output, along with research articles, data, and metadata. However, essential work is still needed to allow credit for software to become more well defined and traceable in science. Indeed, while many see software as yet another kind of data (namely, executable data), software’s specific characteristics have fostered a large movement within the Open Science community towards, e.g., FAIR software development and properties [Katz et al 2021].

### **4.4 Examples of Successful Open Source Research Software**

Open Source Software has been a primary component of research in Computer Science over the past decades. There are hundreds of examples of systems, libraries, and tools that have promoted the rapid development of Computing research in the past. Examples include the Unix operating system with its open source distributions such as Free BSD and Linux, scientific writing typeset-

ting tools such as LaTeX, statistical libraries and languages, such as R, and so on. More recently, a large collection of Artificial Intelligence (AI) and Machine Learning (ML) libraries such as Scikit-learn, TensorFlow, and PyTorch contribute to the rapid development of the AI and ML fields.

In Brazil, one of the open source projects with highest impact is the Lua programming language<sup>1</sup>. Lua is a highly flexible and efficient programming language that is used both by academia for research in programming languages and by the industry as a scripting and systems integration language. The original Lua journal paper has over 800 citations. Any developer in the multi-billion gaming industry knows the Lua language, as it is widely used in hundreds of companies in dozens of countries.

Other sciences have also benefited from open source software. The Genomics community, for example, developed EMBOSS, an open source software analysis package specially developed for the needs of molecular biologists. It provides support for sequence alignment, rapid database searching with sequence patterns, protein motif identification, nucleotide sequence pattern analysis, as well as presentation tools for publication.

As yet another example. ROOT<sup>2</sup> is an open source data analysis framework used by high energy physics and others. It was born at CERN and received contributions from developers worldwide. Currently, over 1 exabyte of scientific data are stored in ROOT files. The Higgs boson was found with the ROOT framework.

---

1 <http://lua.org>

2 <https://root.cern/>

## 5 Some additional aspects

Another set of benefits from Open Science comes from their increased visibility outside the scientific community. This effect has been clearly seen during the COVID-19 pandemic: preprints and open access articles were accessible to the general public, to government, to practitioners (e.g., in medicine, veterinary, agriculture, environmental conservation), news media. Such documents fueled an intense social debate about discoveries, initiatives and their advantages and disadvantages. This communication beyond the research community brings an effective contribution to the visibility of science in society and to the increase in the benefits gained from research results.

By the same token, the pandemic gave rise to a large number of COVID-19 data repositories, some of which containing publications, but most exposing data, which were accessed and downloaded thousands of times. Just as an example, by September 2021, the Zenodo Coronavirus Research Community repository<sup>1</sup> contained over 1600 open COVID19-related published digital assets, comprising articles, code, software, and several types of datasets, such as clinical data, tweets, mobility data and others, covering from regional to country or worldwide geographical expanses. Additionally, software tools and platforms dedicated to COVID-19 data analyses were also made openly available, again contributing to collaboration through reuse.

Considering that a considerable amount of the science created in the world is funded by taxpayers, it seems a good thing that they should have access to the results, even when the interpretation of such results might be difficult and create long discussions. Also, the international experience shows that improving the openness of science requires not only conquering the hearts and minds of the research community, but also a prominent involvement of research institutions (institutes, hospitals, universities) as well as research funders and sometimes even businesses and industries. The role of institutions is essential for the sustainability of initiatives and for their continuity and permanence.

### 5.1 Citizen Science

The increasing access to information technology devices and software has offered lay people the possibility of contributing to and participating in scientific initiatives, in what is now known as “citizen science”. Though usually considered as participation via data collection, it has extended options for citizen participation, offering individuals opportunities to get involved in research projects at distinct stages, including the design and even the responsibility for launching an experiment. In citizen science, the largest number of cases concerns data collection — namely, when non-scientists actively engage in collecting data, for instance in biodiversity or environmental studies. However, there are increasingly examples of their involvement in, e.g., data curation, data labeling and classification, data analysis, development of open source scientific

---

1 <https://zenodo.org/communities/covid-19/>

software, as well as participation in the preparation of open training materials and even, in some cases, project evaluation. In some citizen science platforms, projects are proposed by and executed by non-scientists, but evaluated by a scientific committee.

Citizen science thus plays an important role in the Open Science movement. First, it raises awareness about the general value of science. Second, it attracts young people to STEAM disciplines (Science, Technology, Engineering, Arts and Mathematics). Third, it stimulates in society a “scientific way of thinking”, based on facts, carefully considering available evidence, open to change in face of new, better evidence. Additionally, one could point out that, through this awareness-raising, it can help populate the virtual world with sound information, to make an opposition to the flood of misinformation, ideology, and fake science.

There are several examples of systematic citizen engagement in data collection for digital processing purposes, a long time before the term “citizen science” was coined — in particular, to help monitor environmental conditions or record observations of nature. For instance, the Sidney Streamwatch program, launched in 1990 in Sidney, Australia, was created as part of an effort to educate school children in the importance of water as a natural resource, in which dozens of schools in Sidney were given equipment so students could monitor water quality, and report results. Since then, this program has run continuously, having celebrated its 30th anniversary in September 2020<sup>1</sup>. Indeed, some of the first sites for monitoring water quality at the national level appeared in Australia in the same decade, with data being contributed by citizens and by county administration.

Bird watching — and recording citizens’ observations of birds — is yet another early example of citizen engagement before “citizen science” became known as such. One of the largest collections of this type in the world is the eBird project<sup>2</sup> hosted by Cornell University. Launched in 2000, according to its site, by mid-2020 it contained more than 100 million bird sightings from all over the world, with an average 20% growth rate in participation every year. Here, citizens help not only through contributing with observations but also in validating and curating the data, thereby helping to provide reliable input to research in, e.g., biodiversity.

While most citizen science projects are connected to the natural sciences, yet another example is that of astronomy, in which the GalaxyZoo project [Pinkowski 2010] has involved millions of people in the identification and classification of galaxies since 2007. While novel machine learning algorithms are being developed for this identification, at that time humans were needed for this task. The first result was the classification, in 3 weeks, of 1 million galaxies by the Sloan Digital Sky Survey “astronomers estimated it would take three to five years to categorize all million galaxies. In the first year, 50 million classifications were made by 150,000 people. Galaxy Zoo became the world’s largest database of galaxy shapes.” [Pinkowski 2010]. Galaxy Zoo is now one of the many projects of Zooniverse<sup>3</sup>, a platform “powered by people”, in which over 1 million

---

1 Sidney Streamwatch <https://www.streamwatch.org.au/>

2 <https://ebird.org/home>

3 <https://www.zooniverse.org/>

volunteers come together to use computational tools to help and participate in scientific observations of the universe.

Today, Zooniverse has evolved into a full-fledged platform for people to develop and share their own citizen-science projects. Among those new applications, Planet Hunters<sup>1</sup> (where people contribute to the discovery of new planets based on transient eclipse data from space or ground-based observatories) has been particularly successful in going a step forward and enabling “citizen-led” science. In such cases, scientific results or discoveries are led or autonomously conducted by citizen-scientists based on the data and the software and methods available in the platform. In the case of Planet Hunters, it is now recurrent that citizen scientists are co-authors to the papers (published in international refereed journals) or even lead papers of their own (not in refereed, but moderated journals). In the particularly successful example of Planet Hunters, the scientific results achieved are sound, and citizen-led papers are being referenced in journals as the discoverers of some new extrasolar planets. According to the founders of the project, the key factors for successfully enabling citizen-led science are: to develop the citizen-science project in a way that the participants are trained in the basics of the science and methods behind the task, and to have the opportunity to interact with the citizen-scientists (through the online platform, for example) at the various stages of the process when questions or doubts arise.

More recently, citizen scientists have also become engaged in contributing to research in the social sciences and humanities — see for instance the discussion in [Tauginiene 2020]. In particular, the notion of citizen science has become an object of study in the social sciences and humanities.

Citizen science is also being increasingly recognized and practiced in Brazil, in particular to help data collection and curation, as well as a lever for education and creating awareness of science. Many examples appear in the context of biodiversity (e.g., within SiBBr , the Brazilian Biodiversity Information System).

## ***5.2 Open Science and Biodiversity research***

Biodiversity is an excellent example of a multidimensional field of study. The more data available, the better the modeling. The more data available, the better the processes for protecting, preserving, and using biodiversity. The more access to information on biodiversity, the better and deeper the social involvement with the processes necessary for environmental quality and, therefore, for quality of life.

---

1 <https://www.zooniverse.org/projects/nora-dot-eisner/planet-hunters-tess>

Thus, not only does technical and scientific information lack adequate circulation among peers, but the decoding of information into a form accessible by society must be available. Needless to say, this information needs to be reliable and, thus, systemic peer review routines are not ruled out. However, we must think about the temporal stability of the databases, which must be as broad as possible — the completeness of these databases must be central to their curatorial agenda. Therefore, as the cost of maintaining such an initiative is significant and as continuous management is necessary, it is best that these databases be connected with solid organizations and with diversified financing. It is moreover necessary that these initiatives find shelter in the policies of the state. Finally, there are two aspects that need to be considered and addressed in a robust way. The first involves trafficking plants and animals. It is possible that this aspect should involve continuous monitoring of the use and users of databases by international security organizations. The data on biodiversity also refers to rare species with high financial value. By making the locations and characteristics of these species widely public, it will be easier to have access to these species, further facilitating trafficking. The implications of animal and plant trafficking are numerous, including human health. The second point involves the strategic security issue of countries that have extensive biological heritage, such as Brazil and several Amazonian and African countries, for example. It is necessary to ensure security and sovereignty over these national assets, as they have relevance for environmental conservation and for processes related to the bioeconomy.

### ***5.3 Ethics, Privacy, and Security***

The demand for open data in the scientific community is not a new trend; however, it has significantly increased over the past few years as the recognition of the advantages of very-large (big-data) datasets for research has been coupled with the scientific and technological advances in data management and analysis. At the same time, international concerns were raised regarding the ethical issues involved, especially when information on human subjects is shared. As already emphasized, the science community has been working under the premise that data sharing should be “as open as possible, as closed as necessary.” While the notion of “Security” is all-encompassing and applies indistinctly to all kinds of digital objects, there are distinct ethical and legal procedures in many sciences, as well as issues involving Intellectual Property of data. The same reflections apply to software, but to a lesser degree — perhaps because open software is designed and developed by specific communities that share documentation and coding principles.

Open Science implies sharing and collaboration, and thus introduces the risk of misuse or unethical use, also called *dual use* of the knowledge shared through papers, data, software, methodologies, and all kinds of research outputs. Therefore, one of the biggest challenges currently facing the scientific community is to balance the incentive for sharing, while avoiding, contravening, or compensating for the risk of misuse. To help guide the different stakeholders involved, there are already many recommendations, legislation, and literature produced about the ethical aspects of open data and data sharing, some of which are referred to below.

### 5.3.1 *Data privacy protection legislation*

The protection of personal information has led to recent legislation in many countries, such as the European General Data Protection Regulation (GDPR), which is applied to all organizations processing personal data in the European Union. Many countries are following suit, such as Brazil (LGPD) or South Africa (POPIA — this latter in effect as of July 1st 2021).

All such laws clearly define the legislators' preoccupation with the consequences of the networked open digital world and their implications to research that involves the life cycle of personal information. While GDPR came into effect in 2016, bringing in its wake a large set of laws, professions, and computing implementations, other legislations are only now prompting creation of regulatory bodies, and review of practices and legislation of ethics committees. For Instance, POPIA (the South Africa Protection of Personal Information Act No. 4 of 2013) is giving rise to the creation of domain-specific Codes of Conduct that interpret it under distinct research domains, notably in health and social sciences.

Such Codes apply not only to the region (or country's) inhabitants, but also to the privacy of personal information as a whole — not only the use of identifiers, but also data that allows deriving identification (e.g., images or voice recordings). For instance, in South Africa, there is concern with personal information of children from other countries, in particular those from regions where such concerns do not arise. Indeed, researchers from many African countries are becoming increasingly worried about the lack of regulation concerning private data from indigenous populations, and so are researchers from countries such as Canada, the USA or Costa Rica (to name but just a few). The GIDA coalition (Global Indigenous Data Alliance<sup>1</sup>) is an example of multinational collaboration in the subject, but still lacking participation from South America.

In Brazil, the law on personal data protection – LGPD (law 13.709/2018) - is less constraining when it comes to research performed for 'academic purposes only' according to article 4 of the law. However, if the research involves partners in the private sector, the data collected are under the protection of LGPD. Furthermore, if research data is collected from protected documents, such as medical records, or interviews and recordings (e.g., in the Social Sciences), research data may also be under the protection of the law.

Thus, although the Brazilian LGPD does not rule over data collected exclusively for academic research purposes, Brazil has since 1996 ethical regulations on research on human subjects, as well as a very organized system of approval for this type of research, under the CEP/CONEP system (Comitês de Ética em Pesquisa/Comissão Nacional de Ética em Pesquisa)<sup>2</sup>. In this context, all research protocols performed with human subjects in Brazil have to be evaluated by a Research Ethics Committee (Comitê de Ética em Pesquisa, CEP). The main piece of legislation used by the CEPs in Brazil is Resolution 466 from the National Health Council (Conselho Nacional de Saúde — CNS), approved on December 12, 2012, which applies to all studies involving humans.

---

1 <https://www.gida-global.org/>

2 <http://conselho.saude.gov.br/comissoes-cns/conep/>

According to this resolution, research in human subjects can only be performed if the principles of autonomy, non-maleficence, beneficence, justice, and equity are applied. Furthermore, the prior free (non-coercive) and informed consent from the research participant is a central piece of the process, which also includes several guarantees to the research participant, such as the right to refuse participation and withdrawing consent at any moment, including exclusion of the data collected, or the “right to be forgotten” a concept that is present in LGPD. Thus, it is noteworthy mentioning that the above referred CNS-Resolution already contained the major principles of protection of personal data included in the current Brazilian and international recommendations.

Furthermore, several international entities dedicated to the ethical and responsible sharing of data have issued recommendations, such as the “Framework for Responsible Sharing of Genomic and Health-Related Data”<sup>1</sup>, by the Global Alliance for Genomics and Health<sup>2</sup>. The document has been translated to different languages, including Portuguese, and highlights the benefits of data-sharing for the advancement of science and medicine, especially in the context of Precision and Personalized Medicine. At the same time, the document issues recommendations for the protection of the privacy of research volunteers and the personal information shared, including genomic information. The Framework is constructed based on the current knowledge about data-sharing, but most importantly under the “recommendations of Article 27 of the 1948 Universal Declaration of Human Rights”. In addition, the Framework also considers an implementation since it aims to “facilitate compliance with the obligations and norms set by international and national law and policies.”

In yet another domain — Social Sciences and Humanities — ethical concerns pose new challenges when it comes to openness, again for possible violation of privacy of individuals, but in this case associated with authorship (including collective authorship). In the case of indigenous populations, this can extend to ethnic communities, in which notions of privacy and of authorship are collective and extended to the entire community. Therefore, the issue here connects the problems of anonymity and cultural appropriation. This has given rise to research groups within the Open Science movement dedicated to establishing policies and mechanisms to share (while protecting) indigenous data, as well as to the integration of Indigenous researchers in the production and analysis of Indigenous data. This originated the definition of the so-called CARE principles for indigenous data governance and sovereignty (Collective benefit, Authority to control, Responsibility and Ethics) [Carroll et al 2020], created under the joint umbrellas of GIDA and RDA.

Within the biological sciences realm, there are at least three ethical aspects that need to be considered in the present context. The interaction of man with plants and animals, in the first place, tells a story of learning (ethno-knowledge) that can constitute a map of discoveries of great importance and, thus, the complete opening of the information needs to be accompanied by protective care of the information. Likewise, secondly, it is necessary to consider sensitive species of plants and animals from the economic-social, ornamental, and medical points of view,

---

1 <https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/framework-for-responsible-sharing-of-genomic-and-health-related-data/>

2 <https://www.ga4gh.org/>

among others, which since time immemorial have aroused interests in academia, but also outside academia, mainly, with regard to plants and animals highly priced in the market. Disclosing their locations can put these species at risk. Finally, sentinel animal and plant species that indicate geological, environmental and social security processes need, at times, ethical protection not only for themselves, but also for the environments and communities that eventually live and interact with these places.

### 5.3.2 *Data and algorithm ethics*

In computer science, the notion of ethics in software development dates back to the seventies. With the so-called data deluge, and the increasing presence of artificial intelligence (AI) algorithms to extract knowledge from massive data volumes, the concept of ethics in computing has been extended to encompass data and algorithms. For instance, one of the first codes of ethics in computing was proposed in 1992 by ACM (Association of Computing Machinery), the USA society for computing professionals and researchers. It has since been updated and followed by other scientific associations. The Brazilian Computer Society (SBC) published its Code of Ethics in 2013, at the same time creating its own Ethics Committee. ACM's present code, from 2018, includes clauses for the responsible practice of data collection and use, and the design and development of algorithms and software [ACM 2018].

The terms “data ethics” and “algorithmic ethics” are treated under different perspectives. The first involves the processes along the data life cycle, including, e.g., concerns on the so-called “data bias”, in which inappropriate data collection may induce algorithms to produce biased results. This, in turn, may lead to inadequate decisions, such as those that result in prejudice in gender, age, social class, religious beliefs and others. Algorithmic ethics concern issues in algorithm design and software implementation. This implementation, in turn, may result in the same kinds of prejudice, e.g., violating individual privacy or creating and propagating fake news. Breaches in data and algorithmic ethics may cause moral and physical harms to people and society. As a consequence, many graduate and undergraduate programs all over the world have created disciplines, or even degrees in such ethics. As well, these questions have motivated the creation of dedicated research centers that require collaboration between, among others, philosophers, humanists, computer scientists, and researchers from many domains. The Big Data Institute in Oxford, in which ethics play a major role, is an example of such a center<sup>1</sup>.

Though AI is a branch within computer science, it is often perceived as a separate research domain, given its pervasiveness when it comes to data analysis and decision support. Ethics in AI has given rise to a movement called “responsible AI”, which covers both data and algorithmic ethics. In AI, many computational models emerge from the analysis of the data — and thus biased data will result in biased algorithms. As well, biased algorithms may produce

---

1 <https://www.bdi.ox.ac.uk/study/cdt/ethics>

biased data. To meet this challenge, Unesco approved in November 2021 its recommendations on ethics in artificial intelligence [UNESCO 2021a]. It is worthwhile mentioning that these recommendations were approved in the same General Assembly that approved the Open Science recommendations — thus, in a sense, indirectly linking Open Science and AI ethics within a set of concerns with how science should be conducted.

## **5.4 Overall challenges**

Open Science is based on the principle of global scientific collaboration and sharing throughout all stages of scientific research and practice, having openness as the main principle. As such, the challenges posed are the same faced by any kind of scientific endeavor; moreover, while collaboration in research may be a rewarding (and in some cases required) experience, achieving and maintaining collaboration is hard. Under this perspective, the challenges of Open Science are the same as those of conducting any collaborative research, with the added requirement of openness. While the digital world helps to foster collaboration, sharing and advancement of science, it may also increase the digital divide. [Medeiros 2021] emphasizes the role of the human component in the implementation of Open Science, and how the many actors involved are both assets and barriers, depending on culture, education and domains involved.

### **5.4.1 Change in culture and attitude**

As portrayed in Figure 1, Open Science involves many actors, each of whom may be immersed in distinct scientific, economic, social and cultural environments. Open Science by Design [NAS 2018], as mentioned in section 3.2, requires that all actors who participate in a research initiative work with openness in mind from the time the research is conceived. This implies changes in attitude and culture, some of which require large efforts, e.g., to collect data or design a protocol thinking that data and protocol will be shared in the future with unknown people, and that they may be reused and repurposed. This requires adjusting mindsets, learning new skills (e.g., in documentation) and rethinking costs. Among the many changes required, the notion of “data sovereignty” must be rethought, entailing rethinking long-standing research traditions.

The term “data sovereignty” was coined in the context of data ownership, and access rights, together with privileges and responsibilities such ownership entails. The underlying idea is that the institution and/or researcher responsible for generating and depositing the data want to monitor the information of who accesses the data, and for what purposes — e.g., to avoid misuse or plagiarism, and eventually withdraw access rights. Though seemingly inconsistent with the notion of openness, in which the responsibility for ethical behavior lies on who uses the data, there is more to the concept. For instance, some countries in Africa have joined efforts to establish the African Open Science Cloud Platform (AOSP)[AOSP 2018], sponsored among others by the African Academy of Sciences. The AOSP is being designed

as a set of repositories and repository services to manage and share research outputs from the countries that participate in this coalition.

The underlying idea is that research data produced in a given region should be managed and shared from that region, and not elsewhere. There are several issues involved here. One of them is the monetization of data — namely, the possibility that when data are stored in outside repositories, over which the original data producers have no control, the entities that govern these third-party repositories can take financial advantage of these data, since sometimes the provenance is hard to prove. Another question is the way open repositories are managed — depending on their governing bodies and/or political evolution, these repositories may be suddenly closed, withdrawing access even from the original depositors or producers. For these reasons, the OECD published a report on repository policies, which include recommendations for repository governance, and actions for eventual repository closing or extinction [OECD 2017]. Sovereignty can extend to other outputs of Open Science (e.g., software).

## ***5.4.2 Training and education – educate for openness, train in open digital practices and good science***

When one mentions training and education for Open Science, the immediate reaction is to discuss curricula or specific subjects. First and foremost, however, Open Science is about science — and thus lifelong education on the value of science. And, for those who want to become researchers, on the practice of science, ethics and responsible behavior. In other words, for Open Science to prosper, science itself has to be valued — the goal of STEM and STEAM programs. Also, because Open Science relies on the digital, training should expose everyone to a minimum of digital literacy. As pointed out in [OECD 2015] “... citizens need to acquire the skills to take advantage of, use and reuse data sets shared by the research community.”

When it comes to specific disciplines, curricula should build competence and capabilities to work in a scientific environment with openness and sharing in mind — i.e., ultimately “open by design” [NAS 2018]. For instance, when a biologist goes on a field trip to collect specimens s/he should also consider by which means the digital rendering of these specimens can be documented so as to be reusable. Moreover, as sketched in Figure 1, actors in the Open Science ecosystem include researchers, facilitators at all levels and ultimately society. Hence, training — the level, the intensity and the subjects — depends on each person’s role in supporting, conducting or benefitting from research. As an example, Nature Masterclasses Online<sup>1</sup> has prepared multiple session tutorials on planning and preparing open data for analysis and managing research data. These courses are geared towards researchers in the life sciences and would need to be adapted to other fields.

---

1 <https://masterclasses.nature.com/>

Also, there is a tendency to presume that any curriculum should include the basics of data management, since repositories are at the center of any Open Science infrastructure and services. However, this does not take into consideration the fact that the ecosystem involves not only people directly associated with the research environment, but also lawyers or legislators, among others.

A useful starting point that takes this into consideration is the Open Science Training Handbook [Handbook 2018]. This is an online set of resources, in constant evolution, whose goal is to “create an open, living handbook on Open Science training”. Created in 2018, additional online materials have been progressively added, to support both instructors and trainees. This handbook is directed towards all kinds of training profiles — e.g., civil society, policymakers, librarians, or researchers, among others. For instance, materials for policymakers and funders include research and data ethics, responsible research and innovation, and FAIR data. Each person can build his/her curriculum in this “live handbook” (as it describes itself).

The rest of this section is centered on the training of those that are directly involved in a research effort, namely, the scientists and support staff. To start with, scientists increasingly recognize the need for some kind of basic background in data management and software development, if only to be able to better interact with computing experts and researchers. There is also a growing demand for learning the principles of data science; any such training must also include awareness of the dangers of data and algorithm biases. At the same time, researchers should become acquainted with the basics of how to design open experiments, collect and document data, and develop software, under the open-by-design principles. Training should take as much as possible advantage of available open data and software as educational resources, thereby illustrating to scientists the benefits of open resources, while at the same time showing how to document them for reuse and repurpose.

On the support staff side, Open Science relies on trained ICT experts and librarians. Both should be trained to support scientists and research activities, e.g., help prepare open data/software for appropriate sharing and preservation and create and maintain trustable repositories [Dawei et al 2020]. While ICT staff are primarily concerned with aspects associated with setting up e-infrastructures and programming support, data librarians and stewards are professional librarians trained in basic data management and the documentation of digital resources via metadata records, according to domain standards. This is complicated by the fact that, as mentioned in section 3, there are countless such standards, and very few consensual ones. Thus, data librarians often specialize in some specific domain, to better assist researchers in preparing their data for open publication. Preservation and archival, among others, require continuous collaboration between IT staff and librarians, and Open Science research environments must count on this cooperation to function appropriately.

Regardless of the various emphases and materials, there are two main axes to be considered in any Open Science training program. The first is that “Data is the 21st century’s new raw material” (see Francis Maude in the Foreword to the UK Government’s 2012 Open Data White Paper) [Open Data White Paper 2012]. Being so, it is not possible to think of the educational process at all levels without considering “data” as an invaluable resource. Once this “raw material”

becomes free, interoperable, and reusable, it can be the starting point for a good model of more transparent practices, which is crucial to reduce the friction between different sectors of society, thus contributing to a more collaborative work among different sectors of society.

The second is that software is a first-class citizen. Whereas scientific communities recognize the value of Open Access and, to a lesser degree, Open Data, they need to become more aware of the impact of their software contributions and make strategic investments in the maintenance and development of key research software. As such, they should support the creation and maintenance of software directories to highlight the impact of software produced in their disciplines and encourage reuse and credit for those involved in the production of software. By the same token, publishers should consider the efforts around FAIRification of open source research software to devise the metadata requirements of software journal publications. Training initiatives should address all levels of software assets and practices used in open source software ecosystems from coding, and testing to tooling, maintenance, and reuse, including open repositories, and communication channels.

### **5.4.3 Sustainability and Costs**

Open science does not come for free. Besides involving all the usual costs associated with creating and maintaining environments for enabling research, one must consider the additional costs for enabling collaboration and sharing, as well as for maintenance of the open e-infrastructure. The latter involves not only computing aspects, such as software, hardware and networks, but also the continuous training of staff, from ICT personnel to librarians to research staff. Training and education costs also include those aimed at researchers and involve all stakeholders that participate in the production of knowledge.

In Open Science, communication of results is not limited to papers. Thus, when mentioning costs, while Open Access may require paying to publishers for opening publications, the costs of creating and maintaining the openness of data and software cover a wide range of permanent expenses that include infrastructure (for repositories) and personnel. Indeed, here, the investment in training researchers and staff (and paying staff personnel) is long-term and may prove to be more expensive than paying for open publications.

Figure 11, already mentioned, extracted from NAS reports on the costs of creating and maintaining an open data infrastructure for biomedical research [NAS 2020], can be extended to include any kind of research environment. It illustrates the three main cost components of the needs of an Open Science infrastructure: the costs considered by the researcher in his/her project, the costs of managing research outputs in the so-called “active repositories” and the costs of managing long-term preservation infrastructure needs.

For instance, an example from the academic world of Astronomy is the Hubble Database, which is the best and simplest one (for an expert) to use among all astronomy observatories. The result is that over 60% of the papers published using Hubble data described results obtained from

analysis of archival data, not by the project Principal Investigators (including Nobel-prize winning results). This also led to new and unexpected uses of the data that greatly contributed to the public perception of science and of this billion-dollar telescope, whose lifetime was extended several times by NASA to three times the initially planned mission lifetime. The cost of this extension was far from negligible but compensated by actual scientific and social achievements.

On the other hand, a 2018 study<sup>1</sup> has shown (and this is valid for the case of Hubble and space missions in general) that low-level archiving and general-purpose tools for data analysis and handling usually account for no more than 30% of the full cost of a space mission (including lifetime operation costs). Though not negligible, in the case of Hubble, it more than doubled science productivity.

The Hubble example illustrates a basic principle, which is creating and maintaining digital research assets for Findability, Accessibility, Interoperability and Reuse (FAIR) [Wilkinson 2016]. It is costly and depends on physical resources and trained actors. Nevertheless, the cost of implementing openness is largely compensated by the opportunities it provides to the advancement of knowledge. Moreover, other savings should also enter this equation, e.g., openness facilitates reproducibility and auditability (and thus decreases fraud) and avoids duplication of data collection or software development.

Last but not least, any cost model has to take into consideration the cost of preservation. This, in turn, raises the question of which digital assets merit storing and sharing. According to the IDC (International Data Corporation) report of 2019<sup>2</sup>, by 2025 the “global datasphere” will have 175 zettabytes (computed based on annual growth). How much of these data can be reused? Which can be discarded, and at what cost in selection, or for future unknown developments? For many kinds of data, one can keep summaries, or descriptors, thus eliminating “superfluous” data. In other situations, such as simulations, one can preserve the software that generates the data, but this means that one must keep enough information to allow for reproducibility, e.g., the computational environment to re-execute that software. Moreover, some simulations are very costly in terms of computing resources — and, ultimately, environmental damage through energy consumption and heat generation. How do we know what will be useful in the future (and how do we define usefulness and future)?

---

1 Report on the United Nations/Italy Workshop on the Open Universe initiative  
[https://www.unoosa.org/oosa/oosadoc/data/documents/2018/aac.105/aac.1051175\\_0.html](https://www.unoosa.org/oosa/oosadoc/data/documents/2018/aac.105/aac.1051175_0.html)

2 IDC FutureScape: Worldwide Datacenter 2019 Predictions - <https://www.idc.com/getdoc.jsp?containerId=US42582518>

## 6 Recommendations on Open Science in Brazil

Brazil must increase its involvement in the Open Science movement to exact benefits from the new opportunities for the development of Science and Technology in the country. Such involvement might also be relevant to maintain the country's role as a first-class partner in international science, and also to continue to effectively contribute to Open Science initiatives. Brazilian Open Science policies should be openly discussed, promoted, and enacted at all levels — from the individual researcher to policymakers and the institutional level. There follow a few recommendations to foster actions towards this goal.

The Brazilian scientific community should:

- 1) Promote the discussion and recognition of Open Science initiatives as a means of supporting collaboration, knowledge production, national integration and decreasing inequity in research.
- 2) Promote the discussion of definitions, values and principles of Open Science with representatives of the Brazilian scientific community, institutional and governmental leaders, and other actors.
- 3) Promote interaction between government entities responsible for the Brazilian Open Science initiatives, scientific societies, research institutes, universities, and research funders, so that the official initiatives have the necessary degree of legitimacy to gain adherence and succeed.
- 4) Assist and inform Open Science initiatives from government, funding agencies, and academic institutions.
- 5) Promote Open Science practices, compatible with spending capabilities, within publicly funded research projects.
- 6) Disseminate, and promote the dissemination of information about best practices related to OS, at all levels, recognizing the need for sustainability in financial and political support to Open Science as a national long-term project, instead of short term, intermittent, initiatives.
- 7) Promote the discussion about the roles of the different stakeholders on sustainable Open Science actions (Government, funders, research institutions, researchers, and society as a whole).
- 8) Foster efforts by all actors across different disciplines to change and improve the research culture including rewarding researchers for sharing, collaborating, and engaging with society.
- 9) Foster the creation of Open Science actions, e.g., to propose strategies to educate and support members of the Brazilian research communities, institutions, publishers and funders to shift their practices and requirements towards Open Science.

- 10) Recognize and promote the need to incorporate diversity into the different solutions to generate scientific knowledge at all levels and instances, where diversity can be regional, financial, economic, cultural, and encompass methodologies, research and education environments, among others.
- 11) Support a sustainable and progressive move to full open access of the scientific literature and promote the discussion about best practices in Transformative Agreements for access to subscription and open journals.
- 12) Recognize and support Research Data and Research Software as first-class citizens of the research ecosystem, at par with Publications [UNESCO 2021] and promote good practices in Open Data management and Open Source Software.
- 13) Promote awareness of the state of Open Science practices and policies among academic institutions in Brazil, recognizing that the costs of Open Science extend beyond open publications, but also encompass all that is needed to support open data, open software, and shared open infrastructures.



## 7 References

ACM Code of Ethics, ACM. 2018. <<https://ethics.acm.org/>>, accessed in December 2021.

A. H. F. LAENDER et al. 2020. Abertura e Gestão de Dados: Desafios para a Ciência Brasileira. *Academia Brasileira de Ciências*. <<http://www.abc.org.br/wp-content/uploads/2020/09/ABC-Abertura-e-Gest%C3%A3o-de-Dados-desafios-para-a-ci%C3%Aancia-brasileira.pdf>>, accessed in February 2022.

A. LANDI et al. 2020. The “A” of FAIR – As Open as Possible, as Closed as Necessary. *Data Intelligence* 2(1-2), 47-55.

BEZJAK, S et al. 2018. Open Science Training Handbook. *Foster Portal* <<https://book.fosteropenscience.eu/>>, accessed in November 2021.

BOULTON, G. et al. 2012. Science as an Open Enterprise. *Royal Society*, <<https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/>>, accessed in November 2021.

BOURNE P.E., POLKA J.K., VALE R.D., KILEY R. 2017. Ten simple rules to consider regarding preprint submission. *PLoS Comput Biol* 13(5): e1005473. DOI: <<https://doi.org/10.1371/journal.pcbi.1005473>>.

BUDAPEST OPEN ACCESS INITIATIVE. 2002. In Budapest Open Access Initiative, <<https://www.budapestopenaccessinitiative.org/>>, accessed in July 2022.

CARRO, L. 2021. Hardware Aberto, uma análise de Possibilidades. *Computação Brasil*, 46, pp 29-31, **SBC**. <[https://www.sbc.org.br/images/flippingbook/computacaobrasil/computa\\_46/pdf/CompBrasil\\_46.pdf](https://www.sbc.org.br/images/flippingbook/computacaobrasil/computa_46/pdf/CompBrasil_46.pdf)>, accessed in November 2021.

CARROLL, S. R. et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1), p.43. DOI: <<https://doi.org/10.5334/dsj-2020-043>>

CORETRUSTSEAL REQUIREMENTS. 2020. v02.00-2020-2022. *Zenodo*, <<https://zenodo.org/record/3638211#.YuKV6XbMKUl>>, accessed in November, 2021.

DAVID, P. 2014. The Republic of Open Science - The institution's historical origins and prospects for continued vitality. *Stanford Institute for Economic Policy Research*, <<https://siepr.stanford.edu/publications/working-paper/republic-open-science-institutions-historical-origins-and-prospects>>, accessed in November 2021.

DILLO, I. et al. 2019. CoreTrustSeal Trustworthy Data Repositories Requirements: Extended Guidance 2020-2022. *Zenodo*, <<https://zenodo.org/record/3632533#.X57MYohKg2w>>, accessed in October 2021.

D.S. KATZ, M. GRUENPETER, T. HONEYMAN. 2021. Taking a fresh look at FAIR for research software. *Patterns*, 2 (2021), p. 100222, DOI: <<https://doi.org/10.1016/j.patter.2021.100222>>.

GIBBS, W. 1995. Lost science in the third world, *Scientific American*, <<https://www.scientific-american.com/article/lost-science-in-the-third-world/>>, accessed in January 2021.

GIBNEY, E., VAN NOORDEN, R. 2013. Scientists losing data at a rapid rate. *Nature* (2013). DOI: <<https://doi.org/10.1038/nature.2013.14416>>.

HODSON, S et al. 2018. The Future of Science and Science of the Future: Vision and Strategy for the African Open Science Platform (v02). *African Open Science Platform*. Report. DOI: <<https://doi.org/10.5281/zenodo.2222418>>.

LAAKSO M, BJÖRK BC. 2012. Anatomy of open access publishing: a study of longitudinal development and internal structure. *BMC Med.* 2012;10:124. 2012. DOI: <<https://doi.org/10.1186/1741-7015-10-124>>

LIN, D. et al. 2020. The TRUST Principles for Digital Repositories. *Scientific Data*, 7(144), <<https://www.nature.com/articles/s41597-020-0486-7>>, accessed in November 2021

MEDEIROS, C. B. et al. 2020. IAP Input into the Unesco Open Science Recommendation. *Interacademy Partnership*, <[https://www.interacademies.org/sites/default/files/2020-07/Open\\_Science\\_0.pdf](https://www.interacademies.org/sites/default/files/2020-07/Open_Science_0.pdf)>, accessed in November 2021.

MEDEIROS, C. B. 2021. The hidden dimension of Open Science: “Peopleware”. *Patterns*, Volume 2, Issue 11, 2021, 100385, ISSN 2666-3899.

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE. 2020. Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs. *The National Academies Press*. DOI: <<https://doi.org/10.17226/25639>>, <<https://www.nap.edu/catalog/25639/life-cycle-decisions-for-biomedical-data-the-challenge-of-forecasting>>, accessed in January 2022.

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE. 2018. Open Science by Design. Washington, DC: *The National Academies Press* <<https://www.nap.edu/catalog/25116/open-science-by-design-realizing-a-vision-for-21st-century>>, accessed in December 2021.

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE. 2019. Reproducibility and Replicability in Science Washington, DC: *The National Academies Press* <<https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science>>, accessed in February 2022.

OECD. 2017. Business models for sustainable research data repositories", *OECD Science, Technology and Industry Policy Papers*, No. 47, *OECD Publishing*, Paris, DOI: <<https://doi.org/10.1787/302b12bb-en>>.

OECD. 2020. Enhanced Access to Publicly Funded Data for Science, Technology and Innovation, *OECD Publishing*, Paris, DOI: <<https://doi.org/10.1787/947717bc-en>>.

OECD. 2015. Making Open Science a Reality", *OECD Science, Technology and Industry Policy Papers*, No. 25, OECD Publishing, DOI: <<https://doi.org/10.1787/5jrs2f963zs1-en>>.

OECD. 2021. Recommendation of the Council concerning Access to Research Data from Public Funding, *OECD/LEGAL/0347*, <<https://www.oecd.org/sti/recommendation-access-to-research-data-from-public-funding.htm>>, accessed in December 2021.

OPR. 1999. Editorial, Pros and cons of open peer review. *Nat Neurosci* 2, 197–198 (1999). DOI: <<https://doi.org/10.1038/629>>.

PACKER, A.L. AND MENEHINI, R. 2007. Learning to communicate science in developing countries. *Interciencia* [Caracas], 32(9), 643-647 <<https://wp.scielo.org/wp-content/uploads/PACKER-A.L.-and-MENEHINI-R.-Learning.pdf>>, accessed in January 2022.

PAMPEL et al. 2013. Making Research Data Repositories Visible: The re3data.org Registry. *PLoS ONE* 8(11): e78080. DOI: <<https://doi.org/10.1371/journal.pone.0078080>>.

PAVAN, C., BARBOSA, M.C. 2018. Article processing charge (APC) for publishing open access articles: the Brazilian scenario. *Scientometrics* 117, 805–823. DOI: <<https://doi.org/10.1007/s11192-018-2896-2>>.

PINKOWSKI, J. 2010. How to classify 1 million galaxies in 3 weeks. *Time*, 28 March 2010, <<http://content.time.com/time/health/article/0,8599,1975296,00.html>>, accessed in November 2021.

PLOMP, E. 2020. Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments. *Data Science Journal*, 19(1), p.46. DOI: <<http://doi.org/10.5334/dsj-2020-046>>.

ROSS-HELLAUER, T. 2017. What is open peer review? A systematic review [version 2; peer review: 4 approved]. *F1000Research* 2017, 6:588. DOI: <<https://doi.org/10.12688/f1000research.11369.2>>.

SCHILTZ, M. 2018. Science Without Publication Paywalls: cOAlition S for the Realisation of Full and Immediate Open Access. *PLoS Med* 15(9): e1002663. DOI: <<https://doi.org/10.1371/journal.pmed.1002663>>.

SCHIMMER, R., GESCHUHN, K., VOGLER, A. 2015. Disrupting the subscription journals' business model for the necessary large-scale transformation to open access. A Max Planck Digital Library Open Access Policy White Paper. *Max Planck Digital Library* <[https://oa2020.org/wp-content/uploads/pdfs/MPDL\\_OA-Transition\\_White\\_Paper.pdf](https://oa2020.org/wp-content/uploads/pdfs/MPDL_OA-Transition_White_Paper.pdf)>, accessed in December 2021.

SOTUDEH, H. 2020. Does Open Access Citation Advantage Depend on Paper Topics? *Journal of Information Science* (2020). 46(5): 696-709. DOI: <<http://doi.org/10.1177/0165551519865489>>.

SUFI, S et al. 2019. Report on the Workshop on Sustainable Software Sustainability (Version 1.0.0). *Zenodo*. DOI: <<http://doi.org/10.5281/zenodo.3922155>>.

SWAN, A. 2010. "The Open Access citation advantage: Studies and results to date". *University of Southampton Report*. <<https://eprints.soton.ac.uk/268516/>>, accessed in February 2022.

TAUGINIENE, L., BUTKEVIEIENE, E., VOHLAND, K. et al. 2020. Citizen science in the social sciences and humanities: The power of interdisciplinarity. *Palgrave Commun* 6, 89. DOI: <<https://doi.org/10.1057/s41599-020-0471-y>>, <<https://www.nature.com/articles/s41599-020-0471-y>>.

TENNANT, J.P., WALDNER, F., JACQUES, D.C. et al. 2016. "The academic, economic and societal impacts of Open Access: an evidence-based review [version 3; peer review: 4 approved, 1 approved with reservations]". *F1000Research* 2016, 5:632. DOI: <<https://doi.org/10.12688/f1000research.8460.3>>.

VAN NOORDEN, R. 2013. "The True Cost of Science Publishing", *Nature* p. 426, 495.

UK GOVERNMENT. 2012. Open Data White Paper - Unleashing the Potential. *GOV UK*, <<https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential>>, ISBN 9780101835329, Cm. 8353, accessed in January 2022.

UNESCO. 2021. UNESCO Recommendation on Open Science. Document SC-PCB-SPP/2021/OS/UROS. <<https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en>>. Accessed in January, 2022.

UNESCO. 2021. Recommendation on the ethics of Artificial Intelligence. <<https://unesdoc.unesco.org/ark:/48223/pf0000378931?posInSet=25&queryId=f4082765-2f1f-4710-a706-047db14472d1-draft-data-297>>. Accessed in January 2022.

WILKINSON, M. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018. DOI: <<https://doi.org/10.1038/sdata.2016.18>>.

## 8 Acronyms

<b>AAM</b>	Author Accepted Manuscript
<b>APC</b>	Article Processing Charge
<b>ASM</b>	Author Submitted Manuscript
<b>CARE</b>	Collective Benefit, Authority to Control, Responsibility, Ethics
<b>CC</b>	Creative Commons
<b>DMP</b>	Data Management Plan
<b>DOAJ</b>	Directory of Open Access Journals
<b>DOI</b>	Digital Object Identifier
<b>FAIR</b>	Findable, Accessible, Interoperable, Reusable
<b>FSF</b>	Free Software Foundation
<b>GIDA</b>	Global Indigenous Data Alliance
<b>GDPR</b>	General Data Protection Regulation - Europe
<b>IT</b>	Information Technology
<b>ICT</b>	Information and Communication Technologies
<b>LGPD</b>	Lei Geral de Proteção a Dados
<b>OA</b>	Open Access
<b>OAIS</b>	Open Archival Information System
<b>OASPA</b>	Open Access Scholarly Publishing Association
<b>OECD</b>	Organization for Economic Co-operation and Development
<b>OpenGLAM</b>	Open movement in Galleries, Libraries, Archives and Museums
<b>OPR</b>	Open Peer Review
<b>OS</b>	Open Science

<b>OSI</b>	Open Source Initiative
<b>RDA</b>	Research Data Alliance
<b>SDG</b>	Sustainable Development Goal
<b>STEAM</b>	Science, Technology, Engineering, Arts and Mathematics
<b>TRUST</b>	Transparency, Responsibility, User focus, Sustainability and Technology
<b>VoR</b>	Version of Record
<b>WDS</b>	World Data System



## 9 Glossary

- CARE** The **CARE** Principles for Indigenous Data Governance, published in 2019, are people and purpose-oriented, reflecting the crucial role of data in advancing Indigenous innovation and self-determination. These principles complement the existing FAIR Principles, encouraging open and other data movements to consider both people and purpose in their advocacy and pursuits. (<https://www.gida-global.org/care>)
- CC** **Creative Commons** is a nonprofit organization that helps overcome legal obstacles to the sharing of knowledge and creativity, mostly by providing licenses and public domain tools that give every person and organization in the world a free, simple, and standardized way to grant copyright permissions for creative and academic works; ensure proper attribution; and allow others to copy, distribute, and make use of those works. (<https://creativecommons.org/>)
- DOAJ** The **Directory of Open Access Journals** is a community-curated online directory that indexes and provides access to high quality, open access, peer-reviewed journals. DOAJ services are free of charge and all data is freely available. DOAJ also operates an education and outreach program across the globe, focusing on improving the quality of applications submitted. (<https://doaj.org/>)
- DOI** A **digital object identifier (DOI)** is a (persistent) unique identifier used to uniquely identify digital objects. Widely used in the identification of publications, it is now also adopted to identify data sets, which become thereby uniquely recognizable, helping to check reuse and citations. DataCite (<https://datacite.org>) is a global non-profit organization that provides DOIs for research data and other research outputs.
- FAIR** The **FAIR** Guiding Principles for Scientific Data Management and Stewardship, published in 2016, provides guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets. The principles emphasize machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.
- FSF** The **Free Software Foundation** is a nonprofit organization, founded in 1985, with a worldwide mission to promote computer user freedom. (<http://www.fsf.org>)
- GDPR** The **General Data Protection Regulation** is an European privacy and security law that lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data. (<https://gdpr-info.eu/>)

<b>LGPD</b>	The General Data Protection Law (GDPL), also referred to as LGPD in Portuguese ( <b>Lei Geral de Proteção de Dados Pessoais</b> ) is a Federal Law in Brazil that comprehensively regulates data protection. ( <a href="https://www.lgpdbrasil.com.br/">https://www.lgpdbrasil.com.br/</a> )
<b>OAIS</b>	The <b>Open Archival Information System</b> reference model provides recommendations on setting up archives delivering long-term preservation of and access to information (in particular, digital information) and creating preservation packages.
<b>OASPA</b>	The <b>Open Access Scholarly Publishing Association</b> is a non-profit association that supports and represents the interests of open access scholarly publishers and related organizations, and advocates for Open Access journals in general. ( <a href="https://oaspa.org/">https://oaspa.org/</a> )
<b>Open GLAM</b>	<b>Open GLAM</b> is an initiative to help coordinate efforts to aggregate, advertise, connect, and support open access to the “GLAM” sector (Galleries, Libraries, Archives and Museums), its cultural heritage initiatives and projects. ( <a href="https://openglam.org/">https://openglam.org/</a> )
<b>OSI</b>	The <b>Open Source Initiative</b> is a California public benefit corporation, founded in 1998, and actively involved in Open Source community-building, education, and public advocacy to promote awareness and the importance of non-proprietary software. ( <a href="http://opensource.org">http://opensource.org</a> )
<b>Repository</b>	Repository is defined as the (hardware and software) infrastructure and corresponding service that allows for the persistent, efficient and sustainable storage of digital objects (such as documents, data and code).
<b>Reproducible Research</b>	The definition of reproducibility covers a spectrum of aspects, depending on the context in which it is used. Generally speaking, reproducible research is any research whose associated documentation (articles, data, software) makes it possible to independently obtain similar results using the same methods but under different conditions (i.e., pertains to results). Some break the definition into levels of reproducibility, including <i>computationally reproducible</i> (also called “reproducible”): where code and data can be analyzed in a similar manner as in the original research to achieve the same results, and <i>empirically reproducible</i> (also called “replicable”): where an independent researcher can repeat a study using the same methods but creating new data.
<b>Research Software</b>	Research software is defined as “software that is used to generate, process or analyze results that you intend to appear in a publication (either in a journal, conference paper, monograph, book or thesis)”.

**STEAM**

**STEAM** is an approach to learning and development that integrates the areas of science, technology, engineering, arts, and mathematics. Originally limited to STEM, it has been extended to include the arts in many contexts.

**TRUST**

**TRUST** Principles enforce the need to develop and maintain the infrastructure to foster continuing stewardship of data and enable future use of data holdings. The TRUST Principles is a means to facilitate communication with all stakeholders, providing repositories with guidance to demonstrate transparency, responsibility, user focus, sustainability, and technology. (<https://www.nature.com/articles/s41597-020-0486-7>)

**Version Control**

Version control is the management of changes to documents, computer programs, large web sites, and other collections of information in a logical and persistent manner, allowing for both track changes and the ability to revert a piece of information to a previous revision.





WWW.ABC.ORG.BR



INSTITUTIONAL MEMBERS OF ABC



SUPPORT



ISBN: 978-65-981763-2-7

