DATA MANAGEMENT PRIMER FOR RESEARCHERS

RESEARCH DATA

LOVE THEM, CARE FOR THEM

Luana Farias Sales
Luis Fernando Sayão

Luana Farias Sales

Brazilian   National Archive
Brazilian Institute for Information in Science and Technology – IBICT

Luis Fernando Sayão

National Commission for Nuclear Energy – CNEN/CIN

Joana Sales Marques (illustrator)

Gilda Gama de Queiroz (translator)

Maria Luiza Machado Campos (revisor)

PREFACE

Hello! My name is **Joana**! I work in a lab where I am a leader of a team of researchers

Dr. Joana' team

Our research generates much valuable data for our own team as well as for other researchers.

I LOVE my data, but sometimes I feel really drowned when I don't know how to handle them. As long as good research supposes good data, however, I know they are supposed to be organized and meaningful, so to increase the worth of my group`s research!!!

That is why I am here with you fellow researchers, **to share some tips on how to process well your research data**.
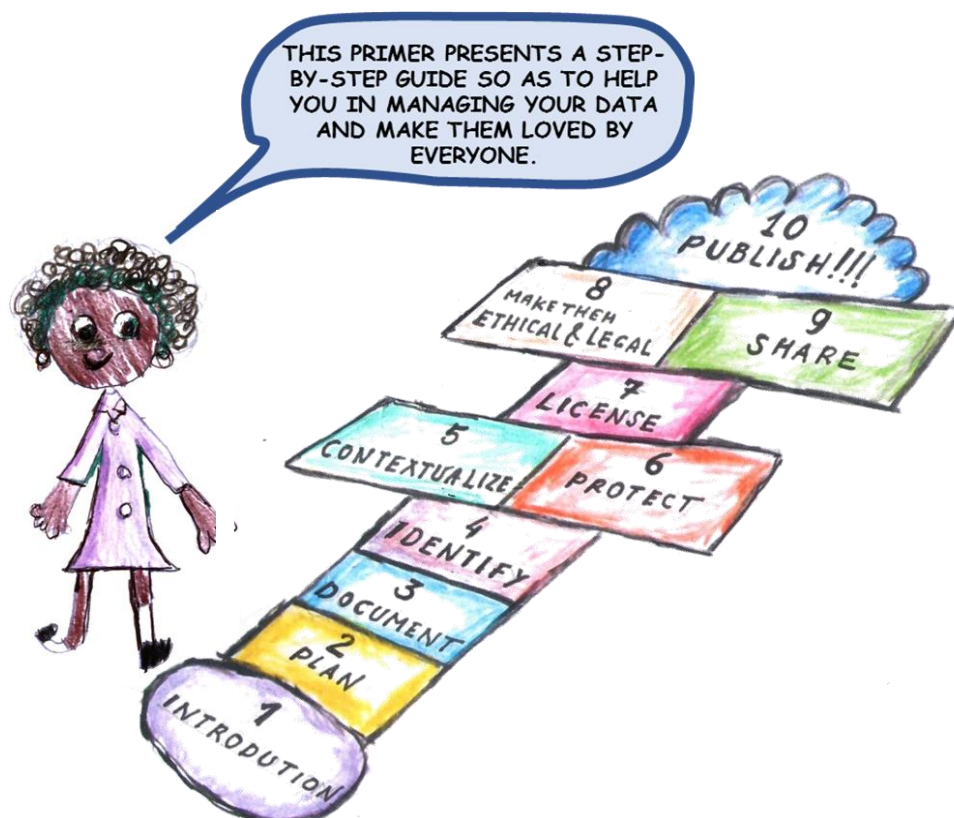
1 INTRODUCTION

My data are as important as **my articles**. Thus, I need to learn **how to manage them along with my research**. Data management will allow my data to be:

- ✔ **PLANNED** and **ORGANIZED**;

- ✔ **DOCUMENTED** so that they can be **MEANINGFUL** to myself as well as to other researchers. Besides, they should have their **PROVENANCE** proven;

- ✔ **STORED** in a **SAFE WAY** and **PRESERVED** for present and future uses;

- ✔ **SHARED**, **ACCESSED** AND **REUSED** for new researches;

- ✔ **PUBLISHED**, **CITED** and with my **AUTHORSHIP** properly recognized;

- ✔ **ALIGNED** with the **POLITICS** recommended by the **FUNDING AGENCIES**, and by my **INSTITUTION;**

- ✔ RESPECTFUL to **ETHICAL** and **LEGAL** principles that apply;

- ✔ **CONTEXTUALIZED** by links to other resources.

Good data management must take into account certain items such as technical, organizational, disciplinary, structural, legal, ethical and economically sustainable aspects. Nowadays it is an essential part in the **responsible conduct of research projects**. It is also becoming part of the **culture of the researcher profession**.

2 PLAN YOUR DATA

Research data management starts with **planning**. This essential phase should **start** when the research is **still being delineated**. It should consider how the data will be **cared for during the research development**, how they will be **archived** and **preserved** after the project closes and, in addition, how they will be **shared** while respecting the researcher conditions, the involved institutions as well as ethical and legal standards.

Many questions can already be formulated and answered at this point:

It is worthwhile to remind that this planning has another objective: to serve as a basis for the elaboration of a **very important document** which is demanded by most of the funding agencies, and sometimes even by your own institution: the **DATA MANAGEMENT PLAN (DMP)**.

Research data management plan

The idea of a DMP is to plan how to generate / collect high quality and sustainable data over time and space. The **document structure** varies according to the standard adopted by your institution and/or the funding agency which will finance your project. The figure shows some information that may be contained in a DMP:
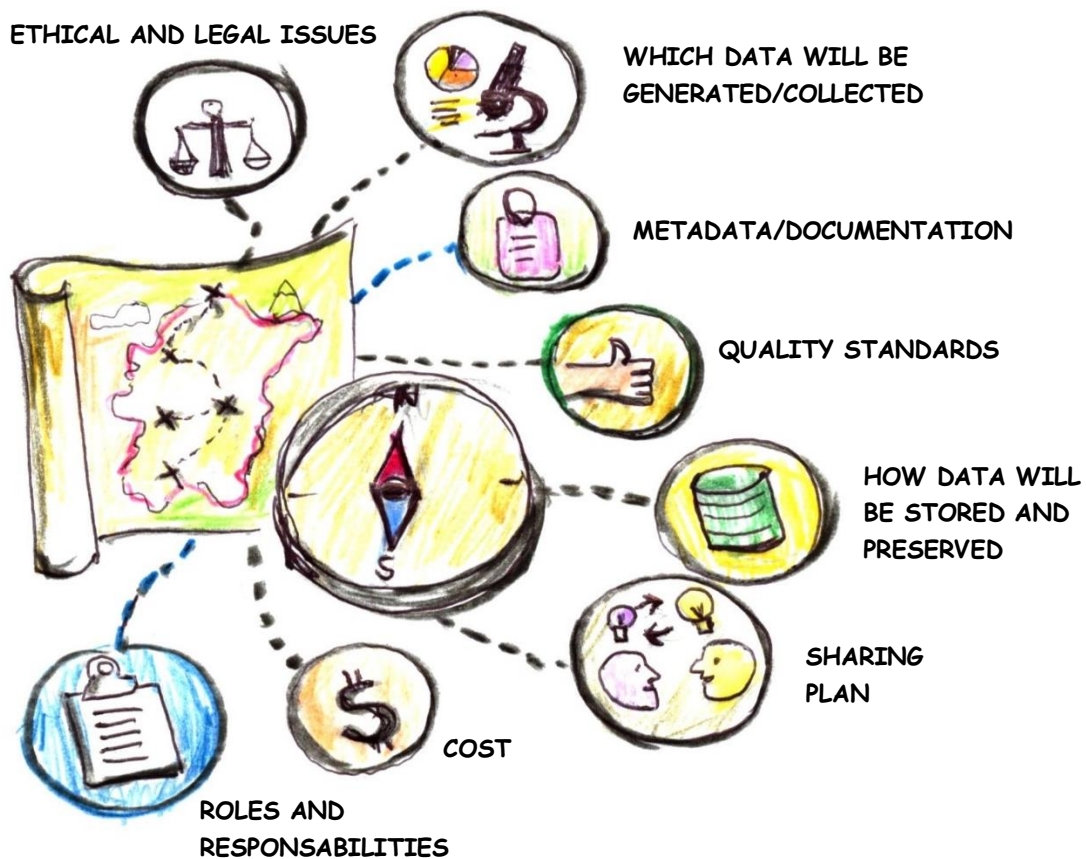
Consequently, DMP helps your project to adjust to **mandatory policies** issued by such institutions as well as to implement data management **according to the needs and purposes of the research**.

It is also a formal document that fully expresses how the data will be treated **during the research development and after it is finished**. Therefore, DMP is a "live" document, **that changes with the project and needs of its participants**.

The Data Management Plan is a key element for good data management (European Commission, 2016).

There are automatic tools and examples of templates on the Web, such as the Data Management Platform tool (**DMP tool**), to help you to develop your DMP.

3 DOCUMENT YOUR DATA

Do as Galileo, describe your data so that anyone can understand them, now and in the future.

Galileo´s notebook integrated his **DATA** (Drawings of Jupiter and its moons), key-**METADATA** (each observation chronometry, meteorological conditions, telescope properties) and **TEXT** (description of methods, analyses and conclusions). This allowed him to understand the Solar System and permited his readers to understand how he reached his discoveries.

A very important part of data management is assuring that yourself, your team and other researchers are able to **understand** and **interpret** data collections. **The reason is that data are not like articles or books, they do not speak by themselves, they need additional information to get contextualized, so that they become information and are able to transmit knowledge** over time and space.
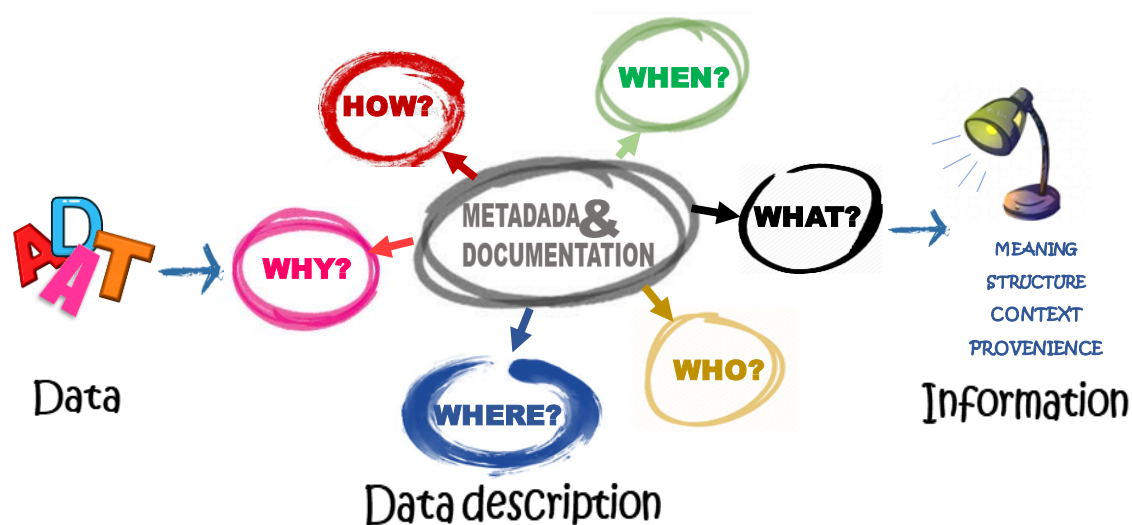
Which numbers are these?

04 27 56
01 16 44
02 01 17

Date?
Age?
Temperature?
Pressure?

Data need to be described and contextualized in order to be understood!

If you wish your data to be captured, identified, understood, reused and to have their provenance duly attested, they should be complemented by a set of information which will describe all their aspects.

The **description** that goes along with the data may be presented in two ways:

METADATA – set of standard and structured information which describes the data and thus, inform about PROVENANCE, ORIGIN, PURPOSE, AUTHORS, INSTITUTIONS, DATES, LICENSES, TECHNICAL DETAILS. The **disciplinary metadata** are intended for specific areas and describe the data with more precision, accordingly. Metadata should be read by people as well as by computers in Web search engines such as Google Scholar or other mechanisms, which brings visibility to your research.

DOCUMENTATION – documents that give support to the understanding of how the data were COLLECTED, GENERATED, PROCESSED and how they were STRUCTURED, ORGANIZED and NAMED. For example: associated research project, laboratory and field notebooks, interview scripts, code books, "read-me" archive and much more.

> The good documentation of research data is essential so they may be managed, identified, published, found, cited, reused and have their value credited to their authors.
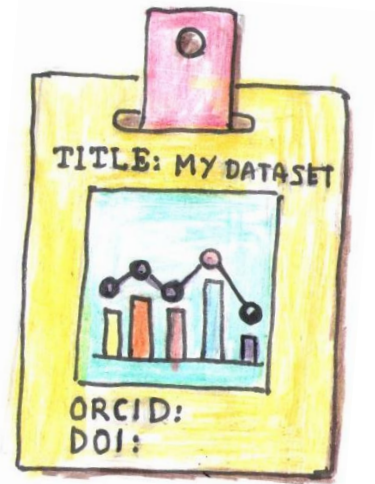
4 IDENTIFY YOUR DATA
so that they will be more cited

The research datasets, in the same way as journal articles, may be **identified** by persistent, standardized and globally unique identifier schemes.

The appropriate identification will permit your data to be **accessed, cited and preserved and their authorship credited to you and your team.**

The **personal researcher identification**, however, adds to more precise identification of your datasets.

So, **two types of identifiers** are very important:

DATA PERSISTENT IDENTIFIER

In order to be correctly cited, your datasets need a persistent digital identifier. The **DOI (Digital Object Identifier)**, frequently used for academic articles can also apply to data.

DATASET IDENTIFICATION
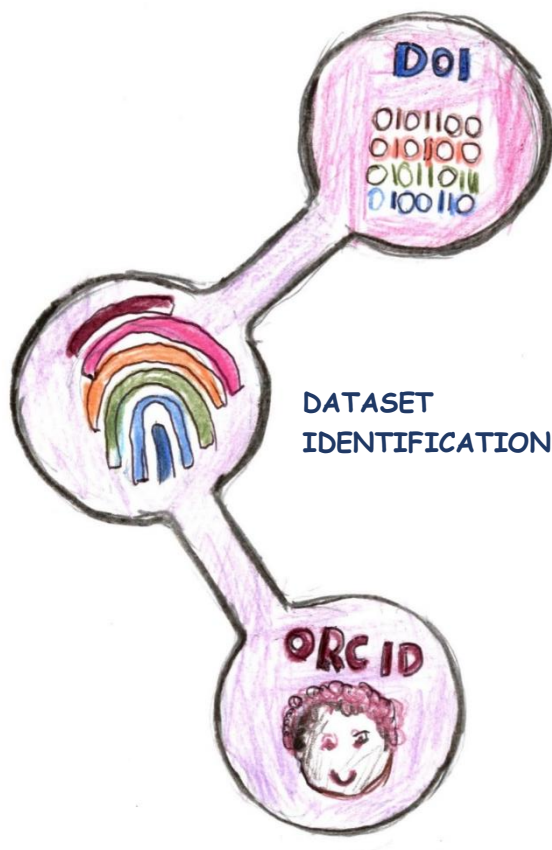
AUTHOR PERSISTENT IDENTIFIER

A way to link the data to you is to create a personal persistent identifier. The **ORCID** is usually used, and it distinguishes you from every other researcher and links you to all of you research activities.

Reminder: give also a **descriptive title** to your datasets!

5 CONTEXTUALIZE YOUR DATA
Create links to other resources

Besides metadata and documentation that help in giving the data some meaning, it is quite important that they are linked to other information resources aiming to create an "ecology" of data and information.

Your data should be, at a minimum, linked to the academic literature that have them as a basis, like your published journal articles. Conversely, the articles should reference the data in the same way that you reference other resources.

But the data may be linked to their own versions, as well. They may be linked to other data sets, projects, laboratories, researchers, reports etc.

This navigable contextualization will allow data to be more accurately understood and to be located and accessed by various means, for example, through a link in an article. And, therefore, this will allow better visibility for your research and more citations for your articles and data.
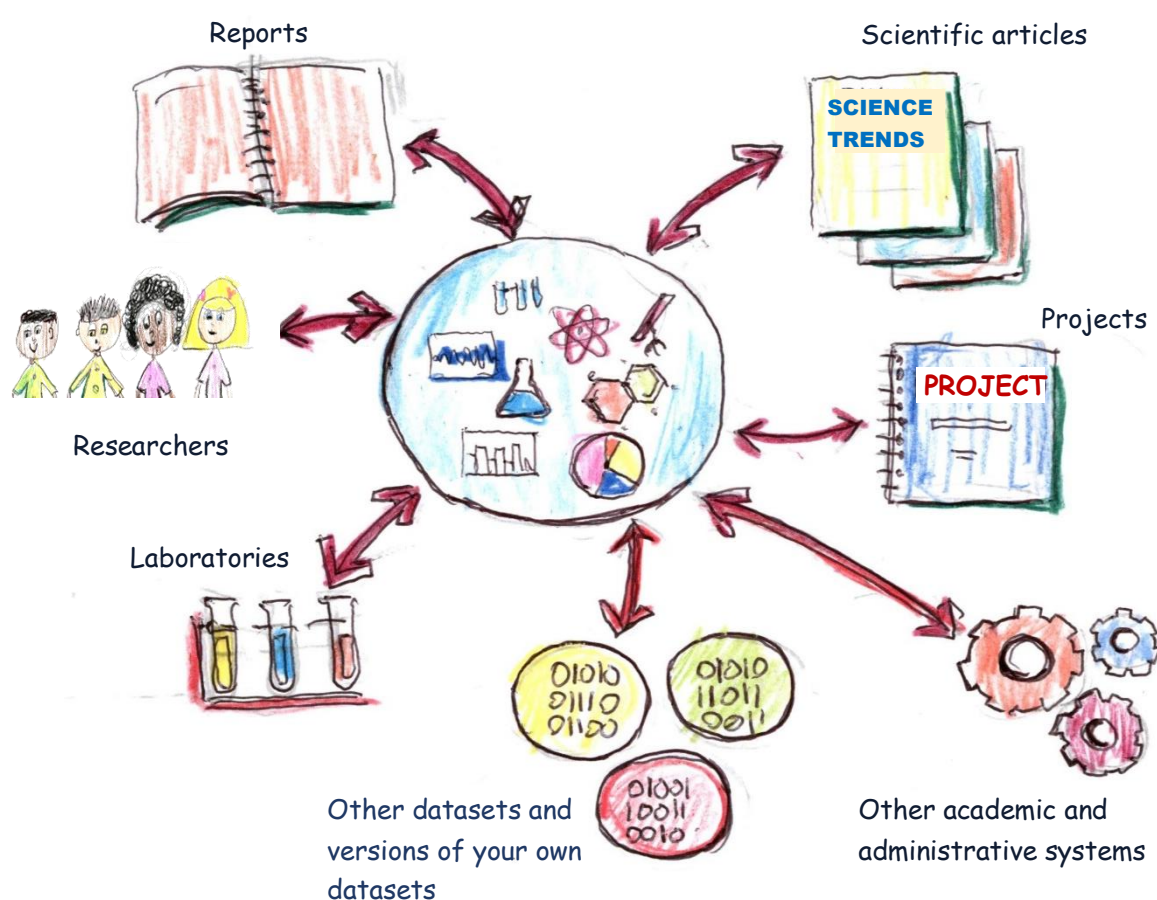
6 PROTECT YOUR DATA

Your data are under the menace of various risks all the time. Risks such as physical damage, losses, thefts, hackers invasion. Data in digital format are digital objects and therefore they are still susceptible to technological obsolescence!

To maintain the integrity and confidentiality of your data, it is necessary that they are adequately protected during the course of your project, until they are archived in a trusted environment, such as a repository or data center.

Physical Access – control the access by unauthorized persons to the local where the data, computers and media are kept.

Passwords – change the passwords frequently to ensure that only the present project members can access the data.

Antivirus - maintain updated protection versions in all of your computers

Firewall – if your computer is linked to Internet, use firewall.

Confidential data storage – don´t keep confidential data in computers that are linked to external networks.

Archive sending – personal or confidential data should be cryptographed before they are sent elsewhere.

7 LICENSE YOUR DATA

When you choose a repository to publish your data, a license is granted to the data. The license is a legal agreement between you, as the data creator or the depositor, and the repository that will take care of the data.

The license associated to your data will settle what kind of permission another researcher who intends to reuse your data will receive.

The postulates of Open Science advise the licenses to permit:

> ✔ The data to be at disposal to the broadest possible audience;

> ✔ That they apply to a wide variety of reuses.

Nevertheless, you can apply more restrictive licenses, in some cases where the open license will be inappropriate. For instance:

> ✔ Datasets including sensible personal data where a term of consent is not granted;

> ✔ Patentable projects data;

> ✔ Whenever there is no permission from the copyright owner.

The Creative Commons licenses offer a wide range of alternatives applied to data, datasets and data bases.

Nevertheless, it is important to remember that the licenses you may attribute to your data may be also determined or restricted by the repository you will choose to publish them.

8 MAKE YOUR DATA ETHICAL AND LEGAL

Collecting, using and sharing data in the course of a scientific research demands that ethical standards and legal obligations be respected. The most relevant within the research scope are those concerning **data protection** and **intellectual property**. Depending on the extent of the partnerships among researchers,, **international laws** may also apply.

THE ETHICS OF THE RESEARCH DATA

When the research involves people, it is expected that the researcher keeps a behavior adjusted to a **strict ethical code** which should permeate the whole research cycle, specially the sharing of data phase. In this context, **two types of data are very important**:

PERSONAL DATA

These include data related to living individuals who can be identified from the data or their combination with other information.

SENSIBLE PERSONAL DATA

These are data on race, ethnic origin, political opinion, religion or philosophical beliefs, sexual orientation, health, biomedical and genetic records, etc.
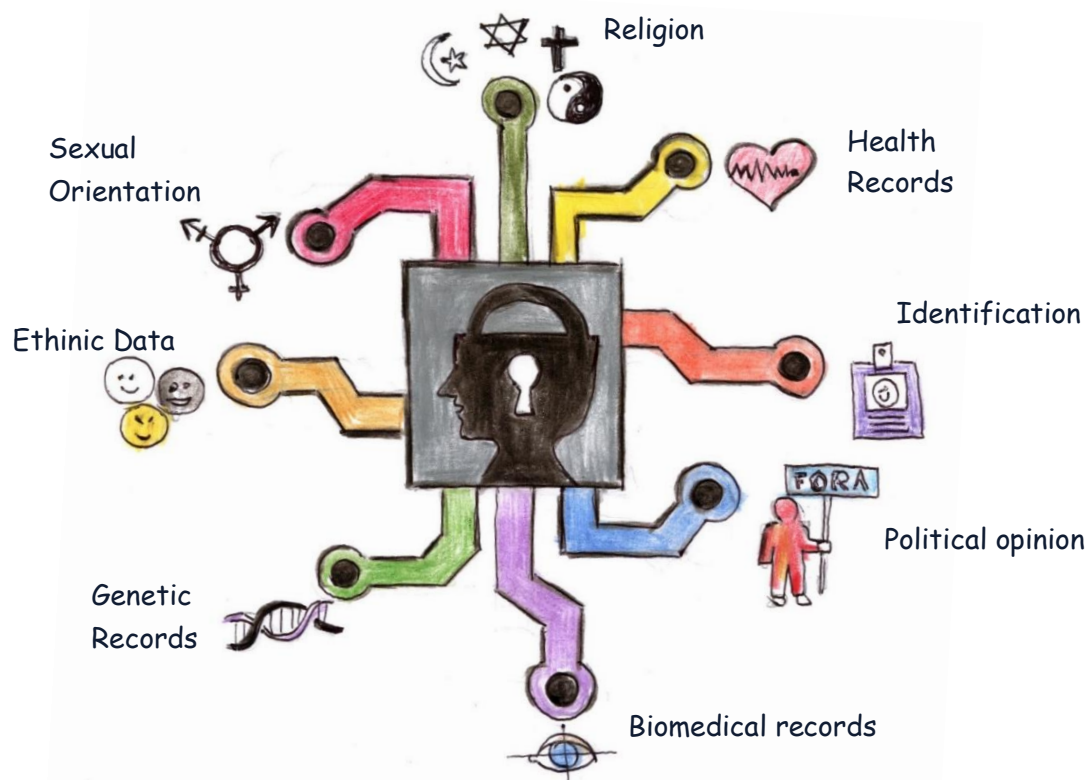
INFORMED CONSENT TERM

Informed consent refers to a communication process which allows a researcher to deliver information to an individual to permit him to take an informed choice on his voluntary participation in a research.

DATA ANONIMIZATION

Includes a series of procedures to submit the data that contain personal information in order to avoid the identification of individuals.

COPYRIGHT

Copyright is an internationally recognized intellectual property right automatically attributed to the author of an original work, such as a scientific research.

However, **copyrights do not apply to facts, ideas and concepts**, but to the specific disposition they are expressed. For instance, in the case of research data, to **spreadsheets, data bases, computer software, models, reports** etc.

Consequently, **copyright also applies to research data** and it is an important item to take into consideration in the creation, sharing and data reshaping or reuse.

Models

Photos
Videos
Diagrams

Spreadsheets
Data bases

Simulation
Software
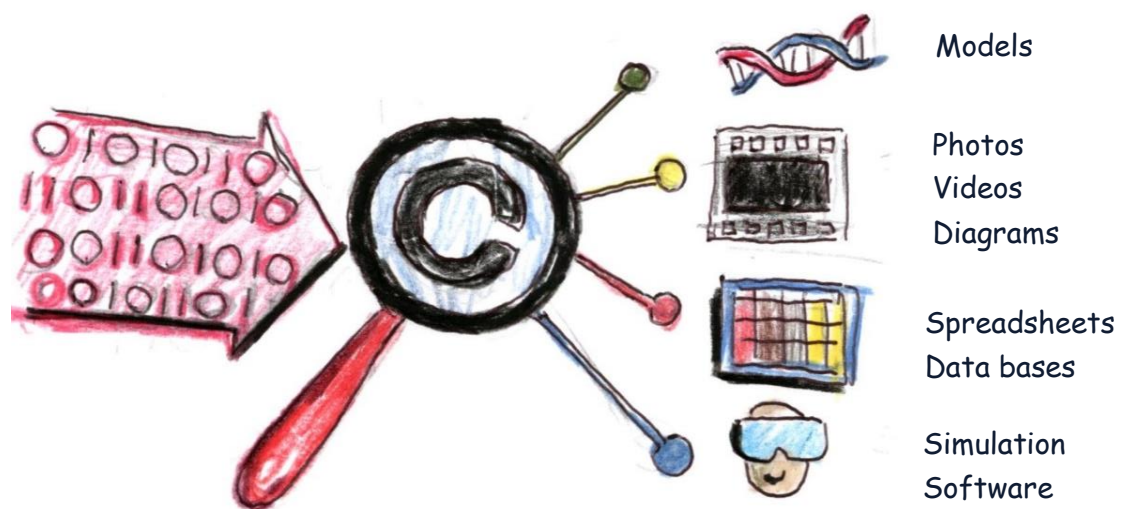
Thus, conflicts between the protection and sharing of personal data can be overcome by combining informed consent, anonymization, access control and copyright. In this way, these data can be ethically and legally used, reused and shared.

9 SHARE YOUR DATA

Love your data, but not to the extent of trying to hide them. Let other researchers love them, as well. **Data sharing is part of the scientific process!!!**

Your research data may be reused now and in the future by other researchers in your area and in other disciplines, since they have a value that exceeds their original value. There are some advantages for you and your team, such as:
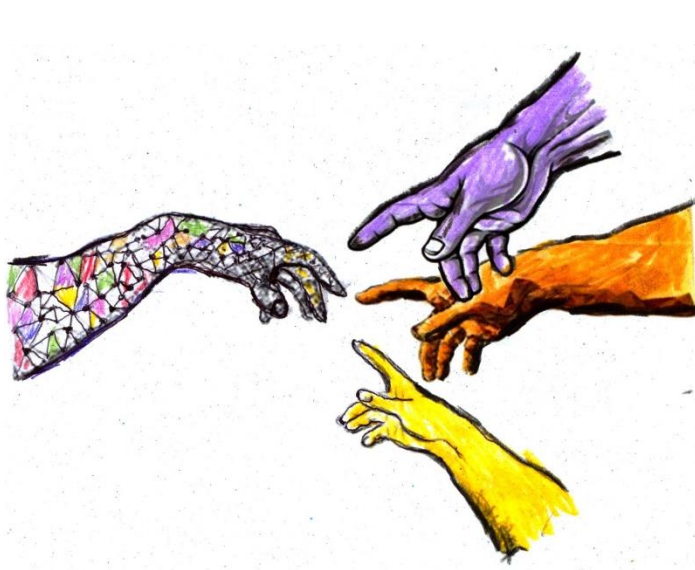
To avoid **effort duplication**

To stimulate **scientific debate**

To promote **new uses for the data**

To enable **new forms of collaboration**
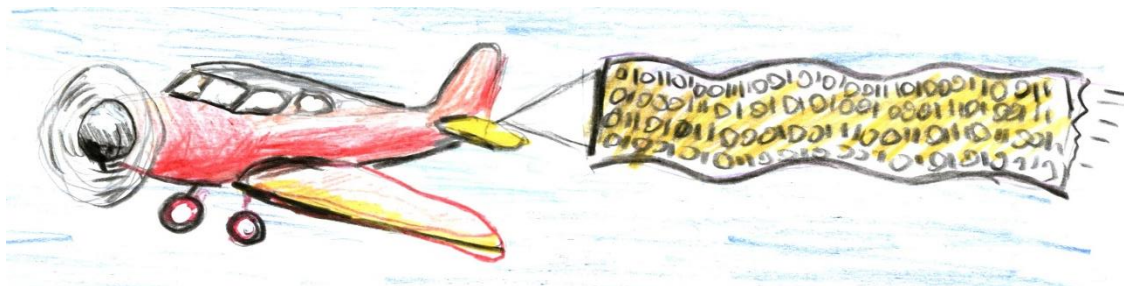
To enable **research validation**

Sharing your data will bring: more **visibility** to your research, more **citations** to your articles, credits to your **authorship** and more **credibility** to your research!

But you need to **publish your data**!!! How? We will see later.
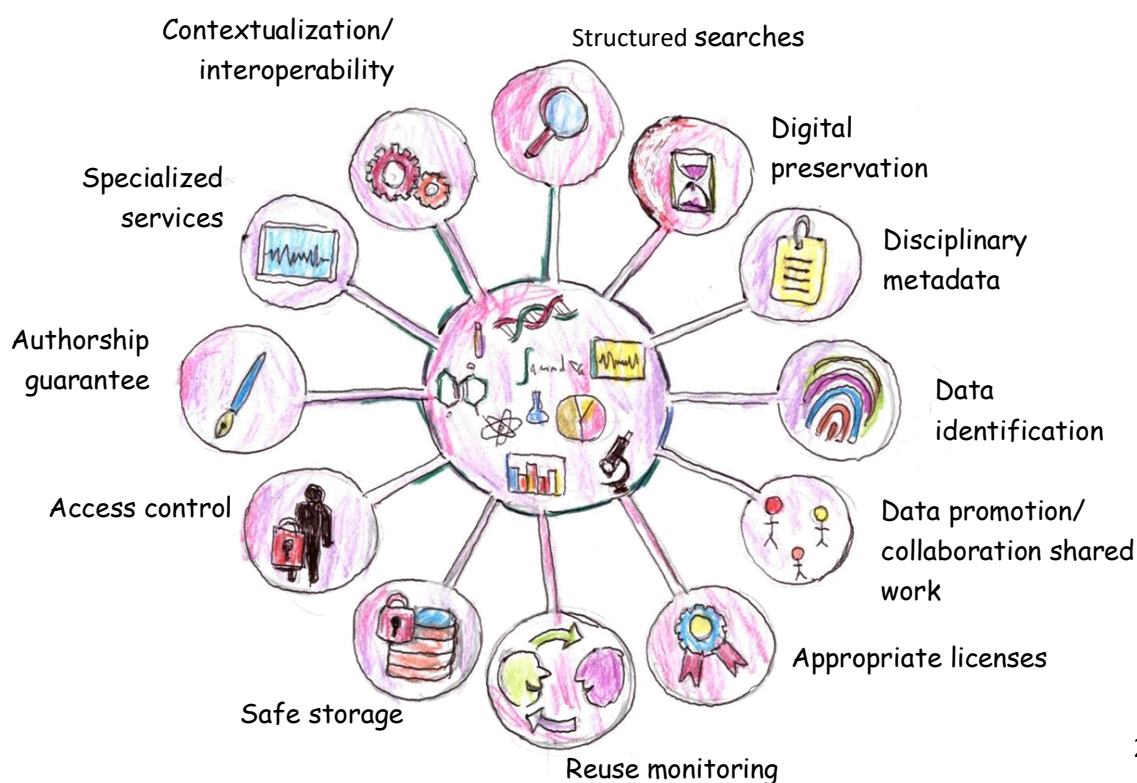
10 PUBLISH YOUR DATA

Show everybody how much you love your data. Publish them (even if you don´t want to release them yet)

You may share the data of your research project Web page or on your laboratory page. But these pages have a short life and they guarantee neither the integrity nor the persistence of your data.

If you wish your datasets to be identified, documented, contextualized, preserved, shared and licensed, with maintained integrity and safety, , so that to give your research more visibility, you should consider to **deposit them in a research data repository.**

The ideal is to publish your data in a data repository in your research area, that is, in a **disciplinary, domain specific repository,** that will care for your data on the long term, by maintaining them stable and authentic. This is true even when you do not want to release them to the public yet. The repositories take care of your data for the necessary lapse of time, without necessarily showing them to the public. See some advantages:

How to find a repository where to publish my data?

There are directories like the **R3data** that index and classify the data repositories. You may find the perfect repository to publish your data in there, or you may find collections that you may reuse as well.

But if **there is no discipline oriented repository** that fits your data, you may deposit in one of the various multidisciplinary repositories such as **Dryad, Dataverse, Figshare**, or still in an **institutional repository** maintained by the organization where you belong.

Remember: many scientific journals demand the publication of your data in a **data repository** !!!

DATA JOURNALS!!!

You may also publish a Data Paper, e.g., an article where you describe your data and publish them in a Data Journal. Your article creates a link to the data that are deposited in a repository. This means more publicity and visibility for your research!
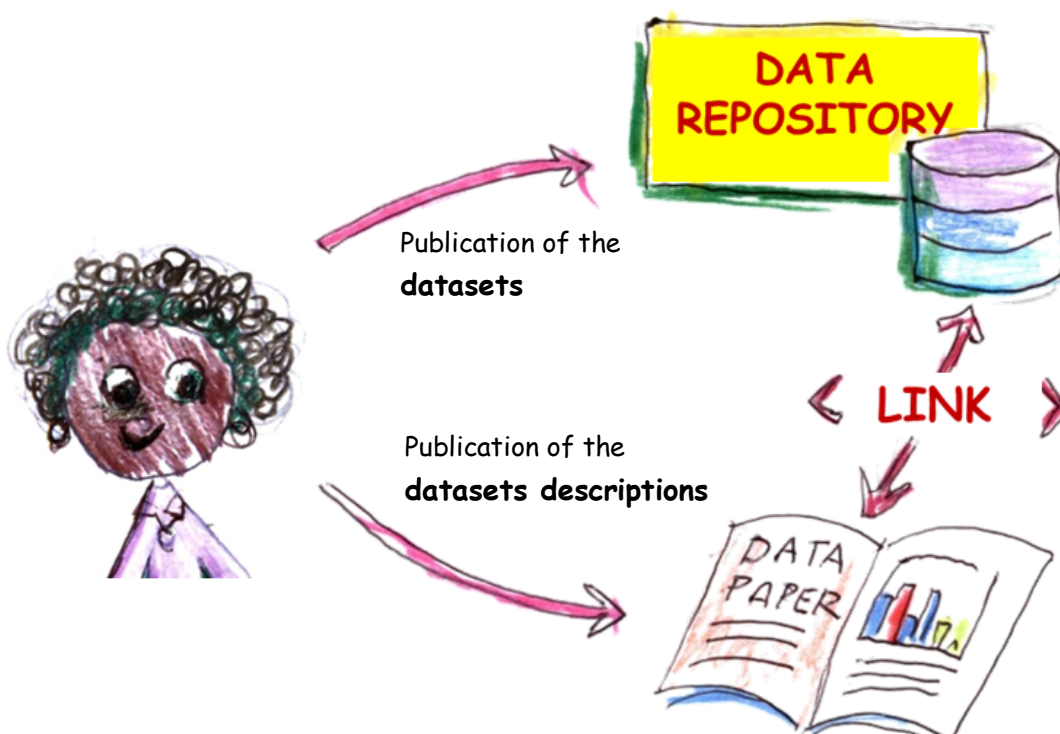
And finally, HAPPY END!!!

Following these guidelines, your data will be more easily identified, found, and accessed;

They will be intelligently available, understandable and reused by other researchers, now and in the future;

They will be linked to other research materials such as articles, thesis and other datasets, creating a contextualization around your data;

They will be safe and digital preserved;

And they will respect the ethical and legal principles, and the rights associated with those, such as copyright.

AND THEREFORE, THEY WILL HAVE MORE **VALUE** FOR YOU AND FOR OTHER RESEARCHERS.

THEY WILL BE CLOSER to follow THE **FAIR PRINCIPLES**, the internationally adopted acronym that stands for **Findable, Accessible, Interoperable, Reusable**.

Accessible

Interoperable

Findable

Reusable
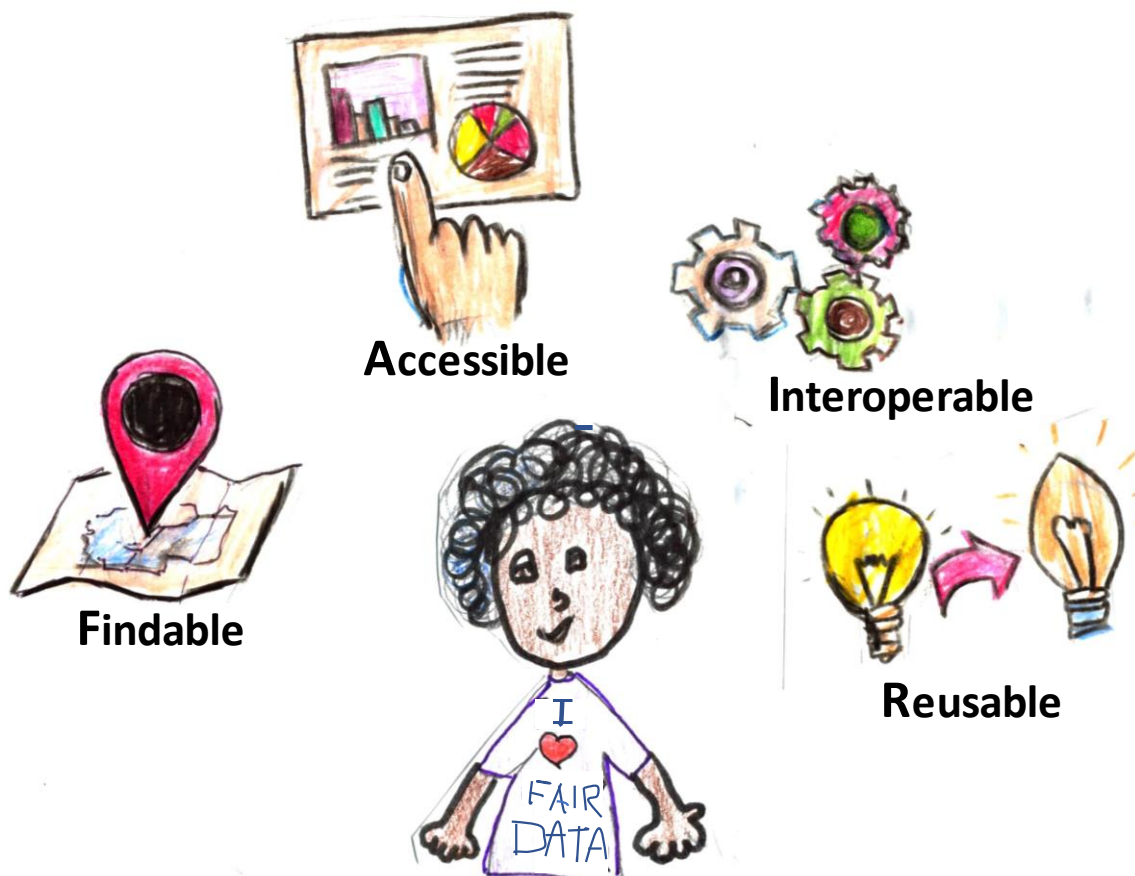
References and Consulted Works

CONSORTIUM OF EUROPEAN SOCIAL SCIENCE DATA ARCHIVES**.** Training Working Group. **Data Management Expert Guide**. Bergen, Norway: CESSDA ERIC, 2017-2018. Available from: <https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide>. Access in: 09 May 2019.

CORTI, Louise et al. **Managing and data sharing**: best practice for researchers. Colchester : Uk Data Archive, 2011. Available from: <https://ukdataservice.ac.uk/media/622416/trainingresourcespack.pdf>. Access in: 09 May 2019.

EUROPEAN COMMISSION. **Guidelines on FAIR Data Management in Horizon 2020**. July 2016. Available from: <https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf>. Access in: 09 May 2019.

SAYÃO, Luis Fernando; SALES, Luana**. Guia de Gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro : CNEN, 2015. Available from: <http://www.cnen.gov.br/images/CIN/PDFs/GUIA_DE_DADOS_DE_PESQUISA.pdf>. Access in: 20 Jul. 2020