OPEN SCIENCE

RESEARCH DATA

PASSPORT FOR OPEN SCIENCE

INDEX

KEY

Underlined text is explained in the glossary.
▼ refers to tools which are given as examples.
↗ indicates an external link.

The digital version of this guide is available at www.ouvrirlascience.fr ↗

As part of the French Passport for Open Science collection, this guide covers the main concepts involved in managing and disseminating research data. It is aimed at researchers like you, whatever your discipline! As you read on, you will find explanations of what research data are, the issues involved in managing them wisely and the benefits of sharing them, as well as how you can best be supported in managing and sharing them.

Isabelle Blanc
National Chief Data and Software Officer
French Ministry of Higher education and Research

WHAT DO WE MEAN WHEN WE TALK ABOUT RESEARCH DATA?

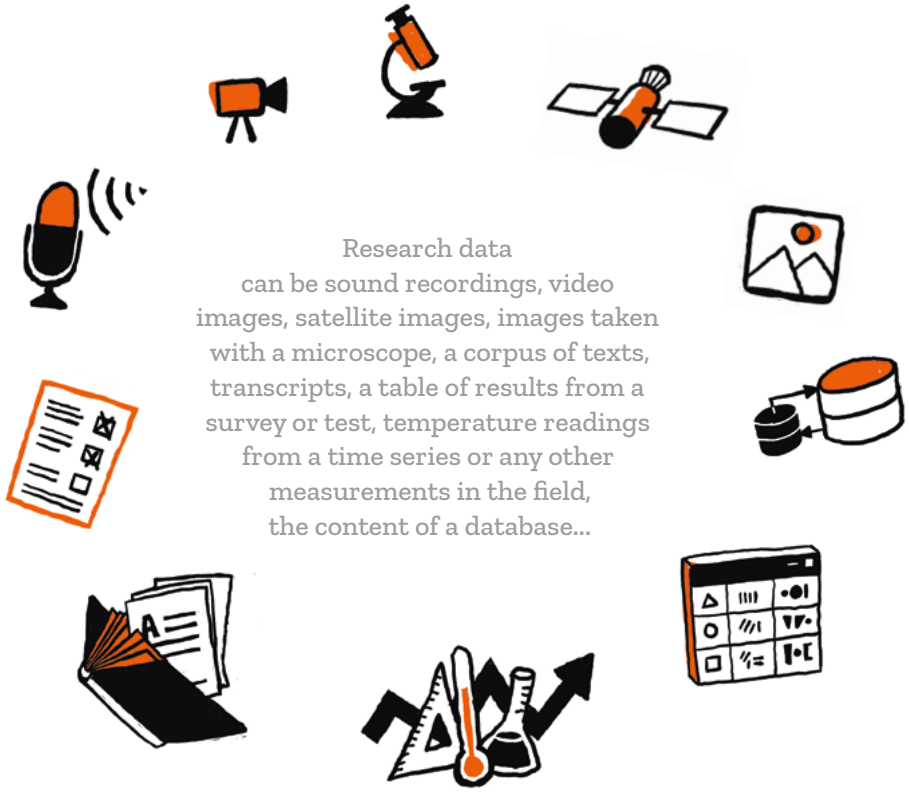
Everything you need to know about research data

There are many definitions, but the most commonly used is that of the Organisation for Economic Co-operation and Development (OECD), which defines research data as:

"**factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research,** and that are commonly accepted in the scientific community as necessary to validate research findings**".**

WORTH KNOWING

Source code and software must not be considered as data. They come with specific challenges, practices and recommendations when it comes to sharing and openness. Consult the booklet entitled *Source code and software*



Research data can be sound recordings, video images, satellite images, images taken with a microscope, a corpus of texts, transcripts, a table of results from a survey or test, temperature readings from a time series or any other measurements in the field, the content of a database...

Research data can be characterised based on the following:

- How they were obtained: data produced as part of experiments or analysis using instruments, observation data, data collected during a survey or field sampling, etc. You can also produce your own data or reuse data produced elsewhere.
- Type: textual, audiovisual, digital, imaging, observation, genomic sequencing, etc. produced using certain measurement, analysis or observation instruments
- Format: data in open or proprietary format.
- Production context: industrial partnership, laboratory in a restricted-access zone, etc.
- Legal regime: personal data (General Data Protection Regulation), covered by secrecy (e.g. professional, defence or industrial), subject to a confidentiality agreement, contractual obligations (contract governing access), etc.
- Critical nature: sensitive, confidential, etc.

All of these notions are used to describe data and make up what we call metadata.

Scientific metadata provides information about the data, in particular: protocol and context in which there were obtained, time references, settings of instruments used, analysis tools and software, etc. using the controlled vocabulary of the research field.

Documentary metadata provides more specific information about things like the institute and the individuals who produced the data, conditions of use and access, the dataset's persistent identifier, the identifier of the publications and software code linked to the data, etc.

A dataset is a coherent set of data in a single project, relating to a single topic or collected in a single location. All of the data in a dataset can therefore be described with mostly shared metadata.

Key steps for making research data open

Over the course of a research project, the data are collected, generated or reused and then stored so they can be processed and analysed. They will then be structured, cleaned and sorted so that only the relevant data for dissemination or publication are kept where possible in a data repository.

Furthermore, some data, especially observations made over time, are also archived for long-term storage.

These different stages punctuate the research project and make up what is called the **data lifecycle. Sound data management** should make it **Findable, Accessible, comprehensible for humans and machines, i.e. Interoperable, and Reusable**. This is what is known as the **FAIR principles**. These cover the different ways research data are constructed, stored, presented, shared and reused (see also "FAIR principes", p. 13-15).

The challenge underpinning the **FAIRification** of research data is ultimately to ensure they can be reused by the team who produced them as well as by others and directly by machines in order to feed into further research, meta-analyses and large-scale models (climate, biodiversity, pandemics, machine learning, etc.).

Legal status of research data

One of the aims of public research is to ensure free access to scientific research data. This is recognised in Article L112-1 of the French Research code. The applicable legal regime for open science is now governed by the general principle that research data must be as open as possible and as closed as necessary. When public, research data are also subject to a principle of openness by default (Open Data), introduced by the French law for a digital Republic (*Loi pour une République numérique*) and now enshrined in the French Code governing relations between the public and the administration (*Code des relations entre le public et l'administration, CRPA*).

Article L. 533-4 of the Research code (*Code de la recherche* in French) adapted from Article 30 of the French Law for a Digital Republic (*Loi pour une République numérique*), further provides for the free reuse of research data after publication where the data:

- is the product of research more than 50% publicly funded,
- is not protected by a specific law or regulation,
- has been made public by the researcher, institute or research body.

Public institutions act as the guarantors for the implementation of the opening of public data. If your research is in France, you can consult the guide to applying the French law for a Digital Republic to research data entitled ***guide d'application de la Loi pour une République numérique*** to find out more about how this works in practice.

Research data management must therefore be reasoned, with openness as the guiding principle and restriction the exception. The decision to maintain restrictions on data must be based on other legal justifications that form exceptions to the general principle of openness. Drawing up a Data management plan is the ideal time to consider these legal issues (see also "Legal issues, honouring exceptions", p. 20).

KEY STEPS FOR MAKING RESEARCH DATA OPEN

BEGINNING OF A RESEARCH PROJECT

PLANNING
Implementation of a data management plan and data FAIRfication

Think about the following key questions: what type(s) of data will you produce? What volume of data? How will you describe them? Process them? Analyse them? Share them? Store them? Is there a specific legislative framework for the dissemination protocol?

COLLECTION CREATION STORAGE

Ask yourself whether it is necessary to produce new data and consider the possibility to reuse previous research.

Prioritise open source software and open formats for better compatibility between tools.

Read up on the procedures for storing the data, security protocols, access rights and data recovery in the event of an incident.

DOCUMENTATION

Do an inventory of the preexisting data used and the new data collected. **Describe** them using **scientific metadata**: creation date, provenance, collection method or protocol. **Use** the controlled vocabulary of your academic community, specify the type and format.

ANALYSIS PROCESSING COMPUTATION

Think about the need to process the data and the environmental and social impact of your research.

ARCHIVING

All documents are intended to be kept permanently or to be destroyed when they have no further administrative value.

Beyond the life of the project, some data are stored for a very long time. **Ask for** more information from the archive department at your institute.

PREPARING TO SUBMIT
Structure and description

Sort your data, selecting which ones need to be submitted and published.
Make sure your chosen repository is trustworthy.
Prepare the data files, indicate the documentary metadata: contributors, establishment, etc., associated files, and the README file that describes them.

DEPOSIT

Deposit files. **Define** reuse rights (licence) and data access rights: open or restricted access.
Update data set versions.

PUBLICATION TO OPEN UP OR SHARE
(restricted access)

Remember to link data to associated publications and vice versa thanks to their persistent identifiers.

REUSE

Find data in various catalogues and repositories. **Check** the conditions for reusing data, set out in the user licence. **Cite** the data using its persistent identifier.

DATA STILL BEING PROCESSED

VALIDATED DATA

CITABLE FAIR DATA

WHY DISSEMINATE DATA?

Sharing and opening up research data **facilitates their reuse both by you and others**, whether team members from your project or research team, or the scientific community as a whole.

Disseminating your data **helps increase the visibility of your work** and allows you to be more cited according to a study published in the journal PLOS ONE, **scientific articles with open data are cited 25% more**.

The dissemination of research data contributes to the transparency of the scientific approach and increases the level of trust in science among citizens.

It also contributes to the **reproducibility** of science, clarifies the way the data were produced, analysed and processed and thereby constitutes a strong marker of scientific and ethical integrity.

It is also worth disseminating data that did not lead to a publication or resolve an initial scientific hypothesis. Such data can be useful to other researchers in exploring new hypotheses, conducting new research, including in other fields, or highlighting negative results.

WORTH KNOWING

The digital sector is enjoying strong growth. It is a consumer of abiotic resources responsible for multiple forms of pollution, and through its impact on the environment and society it exacerbates the strain being put on our planet's limits. Digital research data are part of this growth, and so to avoid increasing your environmental footprint, it is essential to: 1) allow the reuse (FAIR principles) of existing data before trying to produce new data, and 2) document with as much detail and clarity as possible the use and impact of your data. To align open science with environmental aims, it is crucial to make your data findable and accessible, but also to destroy any data that will no longer be useful because it has not been described adequately. These best practices when it comes to data-sharing and destroying redundant data are a way to reduce the digital footprint of data.

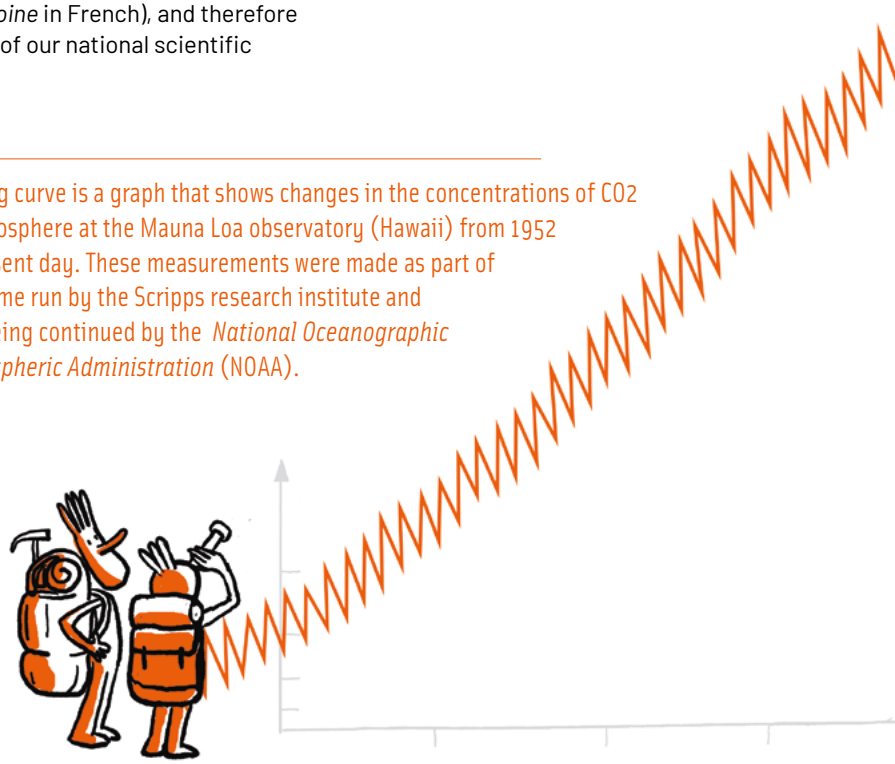
The collection and analysis of data are very costly phases. **Data that are neither shared nor disseminated are therefore a loss for the research team**. The report by the European Commission, Cost of not having FAIR research data, released in 2019, **estimates the cost of poor research data management at €3 billion for France**, due to time lost, non-optimised storage costs, licence costs and problems of duplicated research .

Some research data are unique. This is true of long-term data that monitor environmental parameters, for example. Such data fall under public archives according to the Heritage Law (*Code du Patrimoine* in French), and therefore form part of our national scientific heritage. Using precise descriptions and by sharing and ensuring the openness of on-the-ground observational data, it is possible to constitute time series and conduct analyses over several decades, for example to evaluate the impact of climate change.

The sharing and openness of data play an increasingly central role in public policies. The dissemination of data is included in the recommendations under the French National Plan for Open Science and institutional road maps. It is also a way to respect both legal obligations and the demands of funders as well as certain journals.

EXAMPLE

The Keeling curve is a graph that shows changes in the concentrations of CO2 in the atmosphere at the Mauna Loa observatory (Hawaii) from 1952 to the present day. These measurements were made as part of a programme run by the Scripps research institute and are now being continued by the *National Oceanographic and Atmospheric Administration* (NOAA).



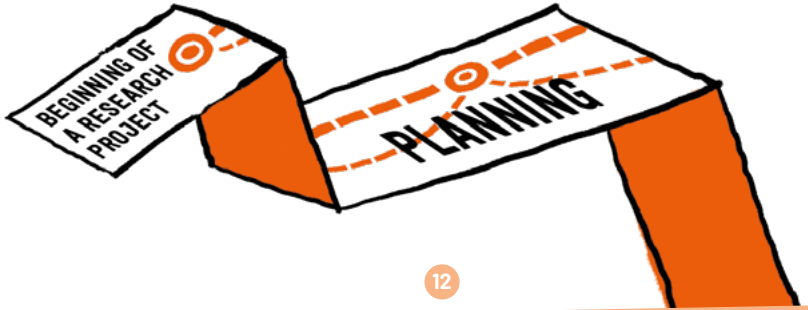
HOW TO DISSEMINATE DATA?

Preparing for data dissemination

Plan your data management

Data dissemination must be prepared at the beginning of the research project. To achieve this, a data management plan, or DMP, is a tool that will enable you to describe how data will be managed, stored, analysed and preserved and to anticipate how to open them up, subject to the legal and contractual frameworks etc. relevant to the project data. The DMP evolves over time and must be adapted to each phase in the research project.

The DMP comes in the form of a document divided into sections based on a model often recommended or even imposed by the supervising body or funding agency. You will find templates at ▼**DMP OPIDoR**. Its purpose is to provide an overview of the description and evolution of the datasets in the research project. It describes the data and how they are managed during the project and defines the procedures for their dissemination, reuse and preservation. It is all the more important to keep it up to date, since it is a data management steering document throughout the project and beyond.

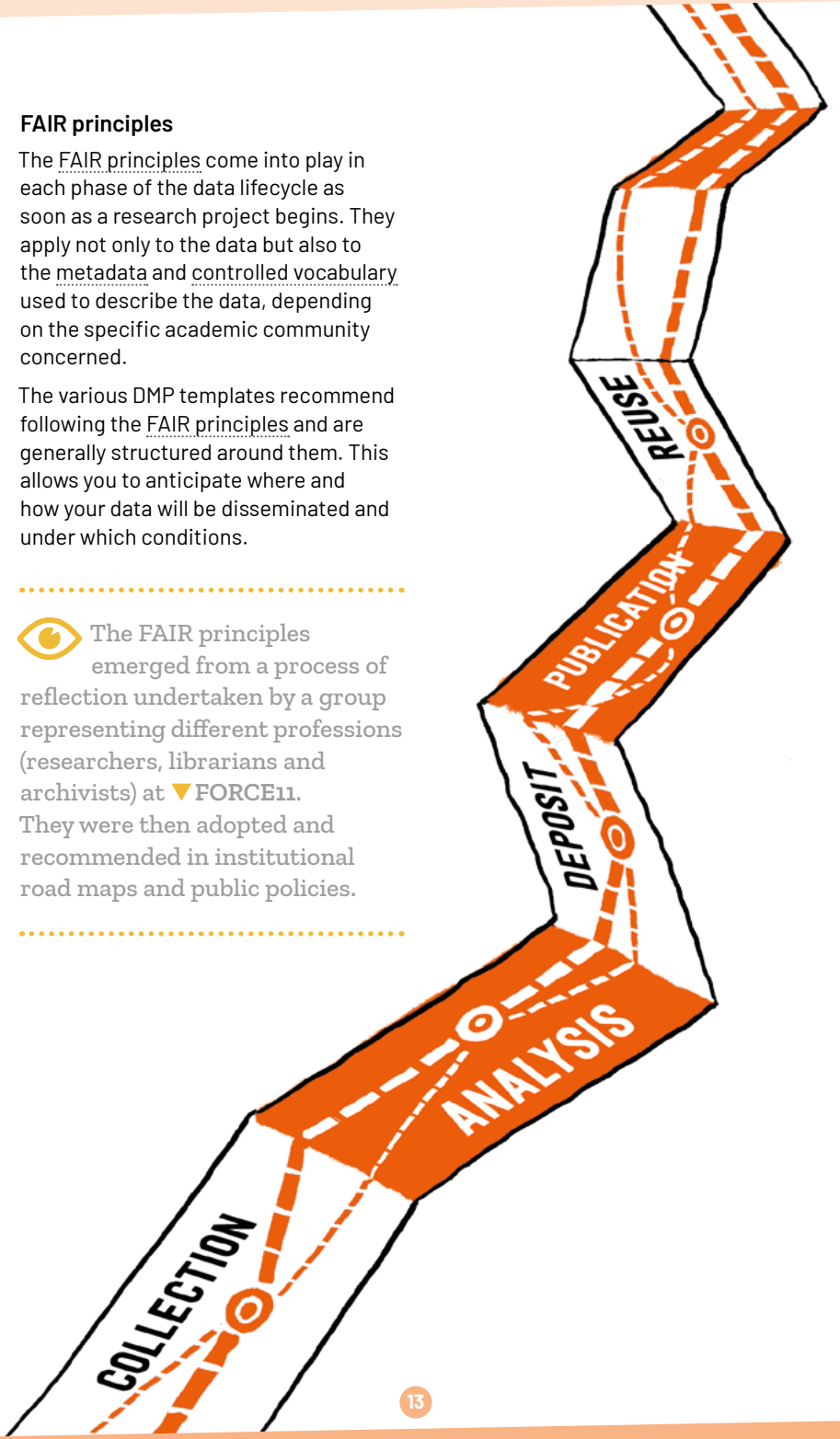


FAIR principles

The FAIR principles come into play in each phase of the data lifecycle as soon as a research project begins. They apply not only to the data but also to the metadata and controlled vocabulary used to describe the data, depending on the specific academic community concerned.

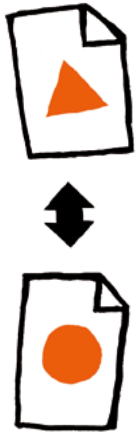
The various DMP templates recommend following the FAIR principles and are generally structured around them. This allows you to anticipate where and how your data will be disseminated and under which conditions.

The FAIR principles emerged from a process of reflection undertaken by a group representing different professions (researchers, librarians and archivists) at ▼**FORCE11**. They were then adopted and recommended in institutional road maps and public policies.



Ensure the data are ***findable*** for humans and machines via metadata indexing.

- Data should be attributed a Persistent IDentifier (PID) (e.g. Digital Object Identifier or DOI) to ensure stable access to the resource.
- Data should be described using scientific and document metadata.
- Data, or at least their metadata, should be indexed or recorded in a research tool, for example via submission to a repository or referencing in a data catalogue.



Make the data ***interoperable*** to ensure they can be used regardless of the computing environment used by humans or machines

- The data should be described at the beginning of their lifecycle using controlled vocabulary.
- The metadata should where possible refer to other data that can be linked together (e.g. naturalist data linked to climate and pedological data).
- The file formats used should be open and documented to ensure exploitable and persistent data using different tools.

Allow ***access*** to the data and metadata

- Online, using a standard, free and open protocol such as https.
- By authentication for data not in open access.
- The metadata should remain accessible even if the data are temporarily inaccessible or if there is restricted access to the data.

Allow the data to be ***reused*** for future research

- The metadata should have several useful attributes to facilitate the comprehension and reuse of the data.
- A licence for reuse should be associated with the data.
- The description of the data should indicate their provenance.
- The structure of the data should be in line with the standards of the scientific community to facilitate their analysis.

Implications of data management choices

Certain choices will have an impact on the quality of adherence to the FAIR criteria and on the environment, via the dissemination of the data: for example, the choice of data repository, choice of vocabulary or choice of format.

Certain platforms or repositories will suggest using standards for your data and metadata. This is the case on ▼**GBIF**, the platform for biodiversity data which suggests the ▼Darwin Core standard for data and ▼**EML** (Ecological Metadata Language) for metadata. This helps satisfy the criteria *I* and *R* of the FAIR principles. ▼**Progedo** suggests the ▼**DDI** (Data Documentation Initiative) standard for metadata when describing data from surveys and other observation methods in social science, behavioural science, economics and healthcare.

Your choice of repository and the way you document your metadata will also influence the potential for the automated reuse of your data. For example, tabular data will have to be digitally transformed so they can be read by machines and thereby included in linked data.

▼**FAIR-Aware** is an online tool that enables you to test your level of knowledge of the FAIR criteria.

Some tools can be used to receive suggestions on how to improve a dataset's level of FAIRness using its persistent identifier, or PID. Examples are ▼**FAIR-Checker** and ▼**F-UJI**. These are limited as they use an automatic approach and can therefore result in bias, so they should be used with precaution.

IN THE FIELD

JOSHUA G.

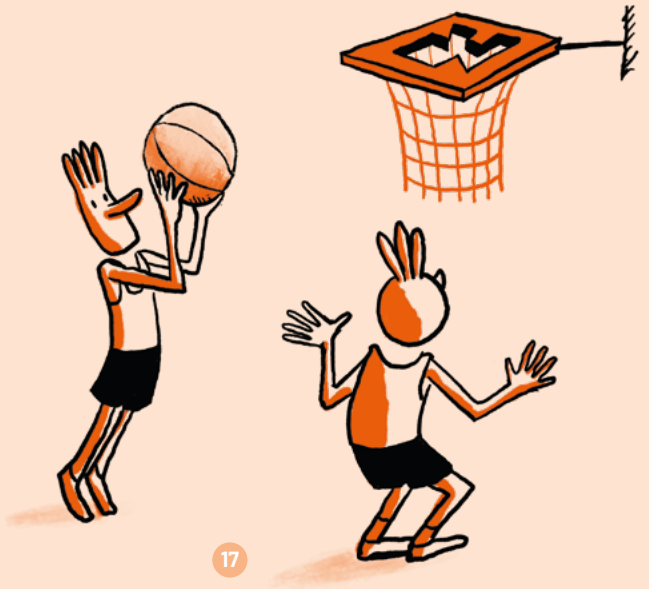
PhD student in biomedicine at the University of Lyon

Signal processing is a major component of research in many fields. While working on my thesis, I had the opportunity to acquire, manipulate and process biomedical imaging data obtained using several different methods. This helped me realise that beyond the logic of algorithms, the format of the data itself is of key importance and is often overlooked in publications.

An increasing number of researchers make their tools available, but the lack of homogeneity in data formats often makes it particularly laborious and time-consuming to use them, sometimes leading to discouragement. New solutions are being developed by the scientific community, and everyone would benefit from being able to use them with simplicity.

In my case, this is happening with the BIDS (Brain Imaging Data Structure) format, which focuses on helping the neuroscience community using a unique standard and an expanding ecosystem of applications that can benefit from it.

This is partly why I feel that harmonising research data is an important objective. And so I endeavour to invest in community-led initiatives driven by this need.



Useful contacts

At a local level, working closely with research teams, many experts are on hand to help you with each phase in your research project:

- The ▼**Recherche Data Gouv research data management clusters** (*Atelier de la donnée*) provide expertise, support and training in and around data. They bring together various data stakeholders (researchers, lecturers, librarians, IT engineers, archivists, jurists, etc.) from one or more institutes. If there is no data management cluster available where you are and no support system at your institute, you can consult the ▼**SOS-PGD** directory which may be able to help you find the right contact.
- "Open science" or "data" advisors may have been appointed in your lab, research structure or university library.
- Some fields also have specific designated contacts. In the humanities and social sciences, the ▼***Maisons des sciences de l'homme*** can offer support.

Those working in the social sciences can also reach out to the University Data Platforms (▼***Plateformes universitaires de données*** *-PUD*) which are overseen by ▼**Progedo**.

At a national level, the ▼**Recherche Data Gouv** ecosystem represents many different actors, including the ▼**data management clusters** and six ▼**thematic reference centers**:

- ▼**CDS** (*Centre de Données Astronomiques*) for astronomy and astrophysics.
- ▼**Data Terra** for earth systems and the environment.
- ▼**PNDB** (National Biodiversity Data Unit or *Pôle National de Données de Biodiversité* in French) for ecology & biodiversity.
- ▼**Huma-Num** and ▼**Progedo** for the humanities and social sciences.
- ▼**IFB** (French institute of bioinformatics) for biology and healthcare.

You will also find self-learning resources at ▼**DoRANum**, the ▼***réseau URFIST*** (French URFIST netwrok), ▼**Couperin**, ▼**CoopIST** or on the ▼**Recherche Data Gouv** ecosystem.

Look for info at your institute: Are there designated "data" advisors? Is there a data management cluster? Contact the person in charge of open science at your institute or university library. If you are a PhD student, you can also contact your doctoral school.

Disseminating research data

Legal issues, honouring exceptions

Research data are subject to a general principle of openness by default, but this comes with certain exceptions set down by law.

One must therefore ask whether there are rights that will ensure legitimate interests be taken into account so as to determine the applicable regime for the data collected and produced over the course of a research project. This might be about protecting intellectual property (authors' rights and *sui generis* database rights), biodiversity, personal data or secrecy (professional secret, national defence secret, etc.). Some sector-specific data (healthcare, environment, archaeology, etc.) must also adhere to particular legal regimes.

Have you produced data as part of your research but are unsure how to share and disseminate it?
Here are a few questions to guide you:

DATA OWNERSHIP

Was your research conducted as part of a partnership?

YES → What is the nature of the partnership?

- **Private** (research more than 50% funded with private funds) → **Shared or open according to the terms of the research contract**
- **Not for profit organisation** → **Partial sharing of data or with controlled access**
- **Individual contributions** (e.g. participatory research) → **Partial sharing of data or with controlled access**

NO

SPECIFIC REGULATION LIMITING DISSEMINATION

Does your work involve processing (collection, analysis, etc.) personal data?

YES → Contact the Data Protection Officer (DPO) → Can the data be anonymised?

- YES → **The data can be made open**
- NO → **Partial sharing of data or with controlled access**

NO

Data relating to security, defence or a protected scientific sector?

YES → If it relates to domestic security or falls under the Protection of the nation's scientific and technical potential (PPST), contact the defence security civil servant (*Fonctionnaire de sécurité de défense*, FSD) → **Restriction based on FSD's advice**

NO

Data subject to professional secrecy?

YES → Medical secrecy, business secrecy (protocols, methods, algorithms), statistical secrecy (INSEE), research involving a human individual (RIPH) → **Partial sharing of data or with controlled access**

NO

Did you answer NO to all these questions?
Once complete and validated scientifically, the data should be made open.
Local support staff can assist you with the management and preparation of your data dissemination.

Data covered by the protection of the Nation's scientific and technical heritage, defence secrecy.

France's Heritage code and the provisions relating to access to public archives govern the length of time for which such data must remain confidential. After this, the data can be freely disseminated. This also applies to other types of exceptions.

Data covered by professional secrecy

Patents protect the inventions, not the underlying data that made them possible. Nonetheless, before obtaining a patent, one must be careful that the dissemination of the data does not in any way reveal details of the invention. Once the intellectual property title has been secured, the data associated with the invention can be opened up provided there is no other impediment to doing so.

Data produced by laboratories in restricted-access zones

Such data are not automatically excluded from the principle of openness by default: to determine which data must be kept confidential, one must ask the authorised parties to indicate which restrictions apply to dissemination, such as the civil servant responsible for security and defence at the institute. Then, as with any other project, it is up to the research teams to identify the finalised public data that can be made open.

Personal data

Management of personal data must be planned as carefully as possible. In order to process, collect, record, modify or even transmit personal data, you must contact the data protection officer (DPO) at the institute to which your unit director belongs for registration in the record of processing activities (in accordance with the General Data Protection Regulation – GDPR) or to request authorisation from the *Commission Nationale de l'informatique et des libertés* (CNIL).

Sensitive data are a particular category of personal data which include, for example, specific information about an individual's racial or ethnic origins, political opinions, religious convictions, health, lifestyle or sexual orientation. It can also relate to genetic or biometric data, generated in order to identify an individual in a unique way. The collection and processing of sensitive data are prohibited in principle, but the GDPR provides for exceptions in the context of research.

Protocols for the dissemination of personal data

Data containing personal information can be made public after they have been processed using encoding, anonymisation or pseudonymisation, depending on the level of confidentiality and the nature of the data processed. The level of confidentiality is to be determined in collaboration with the DPO at your organisation and the objectives of your research project.

- Anonymisation makes it definitively impossible to identify the person. ▼**Amnesia** is a tool that will allow you to anonymise your datasets.

- Pseudonymisation prevents others from identifying an individual without using third-party data. Unlike anonymisation, pseudonymisation is reversible. It involves substituting identifiers (surname, first name, etc.) with indirect identifiers (alias, number, etc.).

Fully anonymised data no longer contain personal information and can therefore be open, provided it is possible to establish that the reidentification, even indirect, of the individuals concerned is no longer possible. Pseudonymisation is a data protection measure, but the data remain subject to the regulation on personal data and cannot therefore be open.

Licences

When publishing data, it is highly recommended to associate them with a licence in order to define how they can be reused and modified.

In France, a decree lists the licences that administrations can use to disseminate public data. These are ▼**Etalab** licences, which provide the producers and reusers of the data in question with the necessary legal security, authorising their reproduction, redistribution, adaptation and commercial exploitation while making it mandatory to cite their provenance.

In addition to the Etalab license, we also recommend adding the ▼**Creative Commons** licences. These allow you to customise the degree of openness you want, and with the CC-BY license, to credit the producers of the datasets. A list of licences is usually proposed by the data repository which will be responsible for storing and disseminating the data submitted.

Practical issues

Contributors in the team

Throughout the data lifecycle, different people contribute to their openness: the researcher who considers which data to open up when drafting their data management plan, the professionals who support the research process and accompany the different data management phases, the data protection officer who advises the researcher on the conditions under which personal data can be opened up, the scientific supervisor of the project who submits the data to a repository so they can be reused, and finally the publishers of data papers.

To acknowledge these various contributors properly when disseminating your results, you can consult ▼**CRedit**, a taxonomy that identifies up to 14 different roles within a research project.

Choosing a data repository

Choosing your data repository is crucial because they are not all equally compatible with the FAIR principles. In order for data to be easily *Accessible*, they must be made available in a data repository. In order for data to be *Findable*, they must also be referenced in catalogues or on platforms using a persistent identifier.

When publishing a dataset, the data repository will assign it a single persistent identifier. The more data are described using rich and detailed metadata (title, producers, date, summary, format, persistent identifier, conditions of access and use, geographic and time metadata, etc.), the better they will be indexed and therefore easier to find.

To meet data quality objectives, some repositories moderate data before they are published and suggest to the depositor ways to improve the description of their datasets based on clearly defined criteria set out in a curation guide.

For greater visibility, sharing and reuse of the data produced or collected as part of scientific projects, there is a diverse offer of data repositories: thematic or disciplinary, focused on trust or certification, institutional or sovereign, generalist, etc.

The plateforms ▼**Cat OPIDoR**, ▼**re3data.org** and ▼**FAIRsharing.org** list several repositories. In order to identify the one best suited to you, it is useful to find out about each one's business model, functions and characteristics to make sure it will cater to your scientific, documentary and technical needs (disciplinary field, type of data accepted, limited volume). When submitting an article, the publisher may ask you to provide the associated data with a view to disseminating it. Read up on the best practices on the guide ▼**Sharing data linked to publications** (*Partager les données liées aux publications*).

WORTH KNOWING

Some repositories offer the possibility to publish data under an embargo so that initially only the metadata are made public. This enables the data to be flagged up and cited, but does not give access to the files themselves, meaning they are not available for public consultation or download. Nonetheless, it is possible for the depositor to grant certain identified individuals access to the data files under an embargo. The Recherche Data Gouv repository allows you to determine the access rights. For example, when submitting an article, if the journal requests access to the data for the reviewers, by submitting them to the repository, you can provide temporary and targeted access to the data files, which will only be made open access once you decide to do so after publication of the article.

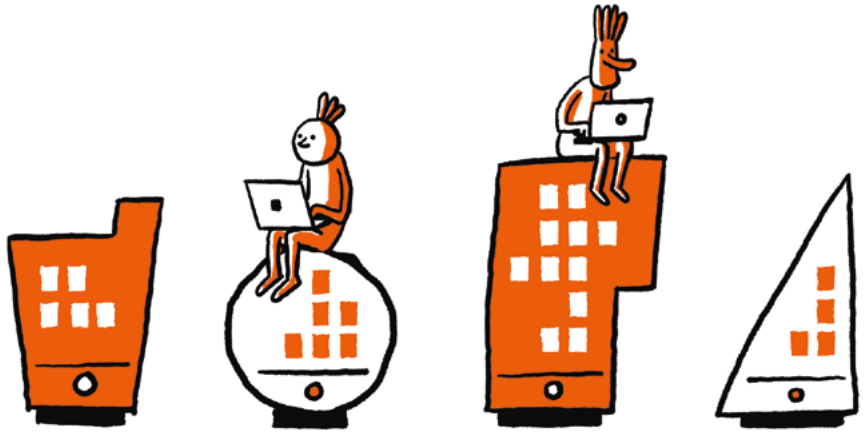
▼**Science Europe** via the guide ▼**Criteria for the Selection of Trustworthy Repositories**, tells us that a trustworthy repository should satisfy the following four criteria:

- systematically attribute a persistent identifier to data and/or datasets,
- suggest a standardised and open data description framework,
- set out the conditions of access and the framework for reuse through the granting of licences,
- guarantee a certain level of preservation and accessibility in the long term for both the data and metadata through the implementation of a dedicated policy and governance.

There are other questions you should ask yourself when choosing a trustworthy repository:

- Is there a repository used by peers in your field of research?
- Does the repository meet the national open science policy objectives and satisfy the guidelines set out in the FAIR principles?
- Will it assign a persistent identifier to your data?
- How long will the data be stored?
- What type of moderation does it use?
- Does it offer the possibility of an embargo?
- Is it recommended by funding agencies?
- Is the repository certified?

For communities that lack a thematic repository, the ▼**Recherche Data Gouv** repository offers a sovereign, trustworthy and multidisciplinary service for publishing data. It will attribute an identifier and licence to your dataset that will enable you to be cited.



WORTH KNOWING

Certification is a way to award labels to repositories for a fixed period based on a list of criteria established by recognised bodies. Until 2023, only three French repositories had been granted ▼**CoreTrustSeal** (CTS) certification: ▼***Centre de Données Astronomiques de Strasbourg*** (CDS), ▼***IFREMER-SISMER*** and ▼***Institut d'astrophysique spatiale*** (IDOC). However, several repositories are deemed "trustworthy" when they provide a certain level of services and guarantees. CTS certification should not be a non-negotiable criterion when choosing where to store your data.

ATTENTION!

Submitting data to a repository does not mean it will be kept for a very long time. It is important to distinguish between storing, saving and archiving data.

Storage simply means that data are held digitally for the duration of the project, whereas the aim of saving data is to duplicate them on various digital devices.

Archiving is a process which at the end of the project allows you to conserve selected data for a very long time. All data repositories disseminate data, but only some of them offer data archiving services in partnership with organisations like ▼**Quetelet-Progedo**.

IN THE FIELD

JULIETTE G.

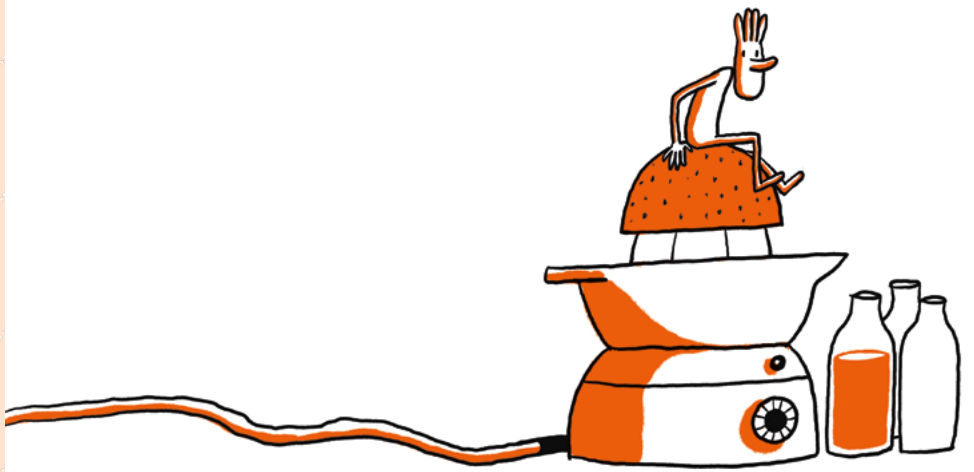
PhD student in hydrology
at Gustave-Eiffel University

I work on the risk of flooding in France, particularly around the Mediterranean basin. The aim of my thesis is to apply and evaluate a model that could predict the impact of a sudden rise in water levels.

As part of my thesis, I had to do a training course on open science. Before that, I didn't know much about it, but now its importance is self-evident. It's our job as researchers to make our results available to anyone who needs them, and to be totally transparent in providing all the data that enabled us to draw our conclusions.

I had the opportunity to publish my first article in an open access scientific journal, and at the same time I published the dataset ↗ on the ▼**Recherche Data Gouv** platform. My article is currently going through the peer review process, and the reviewers can consult my dataset to make sure that my work is reliable.

As much as possible, I've decided to proceed in this way: whenever I publish results, they will always be in open access, and I will provide the code and data I have used. I'm convinced that a real awareness of open science is emerging, and I'll do my best to keep up with the movement.



CHEDID S.

PhD student in geotechnical engineering at the University of Nantes and Gustave-Eiffel University

I'm part of the team working on the geotechnical centrifuge (Lab. GERS-CG), and there are two aspects to my research: experimental and digital. During my PhD project, after publishing the results in an international journal, several researchers contacted me to gain access to the experimental data.

Mindful of the importance of the centrifuge trial results for research, and after consultation with my supervisors, I decided to publish ↗ the data on a platform that provides open access to the public: ▼**Recherche Data Gouv**. Disseminating these data will enable other researchers to use them, thus avoiding having to collect new data from scratch and allowing more in-depth analyses.

In addition, this approach to open science will help increase the transparency and dissemination of my work, and increase public confidence in the results obtained.

From a personal point of view, data dissemination has enabled me to make my work more visible and gain greater recognition within the scientific community. This will help enhance my reputation and that of the Gustave-Eiffel University (Lab. GERS-CG), as well as my impact as a young researcher.

WHAT'S NEXT?
PREPARING FOR THE FUTURE

Promoting your data

As well as submitting your data to a repository, you can opt to promote them in a data paper. This is an article that describes an original dataset, with a view to it being reused. It contains a detailed description of the dataset (production context, producers, associated rights, etc.) as well as access to it, often in the form of a persistent link towards the data repository.

Data papers follow the same editorial and review process as regular scientific articles. There are a number of different journals that publish data papers. They may be multidisciplinary, disciplinary or thematic. They may specialise in data papers or be traditional journals that offer a data paper section. ▼**CoopIST** provides key points in understanding how to structure the content of a data paper, how to choose a journal in which to publish it and how to evaluate it.

You will find different criteria and examples of journals for each discipline.

Linking your data to your other studies

Using a persistent identifier, citations are made easier and more stable since this type of identifier provides a single path to the dataset. In a publication, the associated data, authors and contributors will be unequivocally linked in the long term and with stability thanks to the persistent identifier, regardless of the form of the information used to describe them in the different institutions.

▼**DataCite** is a non-profit organisation that assigns dataset identifiers at an international level. The French agency responsible for assigning DOIs to data, DataCite France, is run by Inist-CNRS. The provision of certain persistent identifiers is part of complementary services such as the automated formatting of citations, made possible by Digital Object Identifiers. The DOI is automatically assigned by the repository in which the data are held.

Identifying the different versions of a dataset

Disseminating a dataset is the first step. It can be updated later if it changes over time. Most repositories record the different versions of datasets to ensure transparency in their evolution.

Long-term archiving

This means storing, guaranteeing access to and protecting the intelligibility of selected data over a very long period (more than 30 years). The selection process is governed by specific legislation and regulations that address questions of accessibility, legibility and scientific heritage. Not all data are destined to be archived for long periods due to reasons of cost and protection of the environment in particular. Data management support staff, archivists in particular, will be able to advise you on these issues.

In France, the ▼ **CINES** (*Centre Informatique National de l'Enseignement Supérieur*) is responsible for the long-term storage of research data and offers various archiving solutions. Any plans to archive data should be analysed in advance in collaboration with the archive department at your institute before you contact the CINES: which parts of the data have recognised scientific value to justify their long-term storage? How much space will they take up? What format are they in? How long would you like to store them? What budget is needed and for how long? What will be the environmental impact of storing your data?

IN THE FIELD

NAOMI T.

Assistant Professor in German Sociolinguistics

Nearly all my data, articles and think pieces are currently available in open access, but it has not always been the case. I feel it is important to say that practicing open science is a process, and you don't necessarily have to show everything right away!

For me, it began with my annotated corpuses. The annotated data came from parliamentary debates in France, Germany and the UK. The project involved bringing attention to the transcripts from these debates in XML-TEI format using a CC-BY 4.0 licence to facilitate their dissemination and reuse as quickly and as widely as possible.

I published that data with open access on ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la LANGue) as soon as I began my thesis and before even publishing the results.

The reuse of parliamentary data is a significant issue for democracy: while the transcripts from parliamentary sessions are all available on the respective websites of each parliament, the exploitation of the data for research purposes remains challenging.

This undertaking allowed me to promote the results of my research on a larger scale. Two data papers described the process to make it transparent and reproducible. And I'm very honoured to have received the "open science research data" award in the "data reuse" category thanks to this work ↗.

If I had just one piece of advice for you, it would be: go for it!

GOING FURTHER

GENERAL RESOURCES

Passport for Open Science collection:
https://www.ouvrirlascience.fr/passeport-pour-la-science-ouverte/?menu=3

▼**Ouvrir la science**: resources. Website of the French National Committee for Open science: https://www.ouvrirlascience.fr/category/ressources/

PLATFORMS OF INTEREST

▼**DoRANum** provides resources to assist the scientific community in the management and sharing of their data. Here you will find self-learning content divided into different themes (objectives, submission, management plan, metadata, etc.) and disciplines: https://doranum.fr

The Open Science Data Working Group ▼**Couperin**:
https://gtso.couperin.org/groupe-donnees/

▼ **Recherche Data Gouv** provides guides, virtual lessons and tutorials:
https://recherche.data.gouv.fr/fr/aide-en-ligne

▼ **CoopIST** (part of CIRAD): https://coop-ist.cirad.fr/gerer-des-donnees

▼**The URFIST network** offers training courses on-site or remotely:
https://sygefor.reseau-urfist.fr/#/

▼**EcoInfo** - Service group that is part of the CNRS to reduce the negative impact of digital technologies on the environment and society: https://ecoinfo.cnrs.fr/

▼**Labos 1point5** - a group of members from academia, spanning all disciplines and regions, who share a common goal: " to better understand and reduce the environmental impact of research, especially on the Earth's climate":
https://labos1point5.org/

PRACTICAL GUIDES

Guide for researchers on sharing data linked to scientific publications (2022). French Ministry of Higher Education and Research, Committee for Open Science. https://www.ouvrirlascience.fr/partager-les-donnees-liees-aux-publications-scientifiques-guide-pour-les-chercheurs/

Guide on the application of the law for a digital Republic to research data (2022). French Ministry of Higher Education and Research, open science committee. https://www.ouvrirlascience.fr/guide-dapplication-de-la-loi-pour-une-republique-numerique-pour-les-donnees-de-la-recherche/

Guide to research data management best practices. Data workshop, MITI – CNRS inter-network working group (2023).
https://mi-gt-donnees.pages.math.unistra.fr/guide/00-introduction.html

Integrating open science into your ANR project: a practical guide (2023). The Open Science Data Working Group Couperin. https://doi.org/10.5281/zenodo.7657818

LEARNING GAMES

Dép'Osez: a game about data repositories. Inist-CNRS, DoRANum (2023). DOI: 10.13143/AABD-HS57

GopenDoRe: a game about managing research data. Inist-CNRS, DoRANum (2022). DOI:10.13143/91td-qe92

THEMATIC RESOURCES

Law

• Robin, A. (2022). *Droit des données de la recherche-Science ouverte, innovation, données publiques*. Larcier.

• GDPR documentation from the CNIL, in particular on sensitive data:
https://www.cnil.fr/fr/definition/donnee-sensible

• List of licences that can be used for data and research code and software.
https://www.data.gouv.fr/fr/pages/legal/licences/

Digital sobriety

• GreenDate guide for a controlled impact of open data. OpendataFrance, beta version. https://opendatafrance.gitbook.io/greendata-pour-un-impact-maitrise-des-donnees/greendata/1.1-contexte

• Publications of EcoInfo: https://cnrs.hal.science/ECOINFO/browse/last

SOURCES

Bouchet-Moneret, F., *Les données personnelles de recherche et le RGPD* (2021). https://hal.univ-lorraine.fr/hal-03636697

Colavizza, G., Hrynaskiewicz, I., Staden, I.a, *et al.*, *The citation advantage of linking publications to research data* (2020). https://doi.org/10.1371/journal.pone.0230416

Code des relations entre le public et l'administration (CRPA) : applicable à compter du 1er janvier 2016. Légifrance (legifrance.gouv.fr)

Article L112-1 du Code de la recherche. Légifrance (legifrance.gouv.fr)

Article L. 533-4 du Code de la recherche. Légifrance (legifrance.gouv.fr)

Cost of not having FAIR research data - Cost-Benefit analysis for FAIR research data (2018). http://publications.europa.eu/resource/cellar/d375368c-1a0a-11e9-8d04-01aa75ed71a1.0001.01/DOC_1

Confederation of Open Access Repositories. *COAR Community Framework for Best Practices in Repositories* (2020). https://doi.org/10.5281/zenodo.4110829

Cotte, E., Sébire, F., *Modèle de logigramme de l'Institut Pasteur relatif aux questions juridiques liées à la diffusion des données de la recherche* (2022). https://pasteur.hal.science/pasteur-03587216

Dedieu, L., *Revues publiant des data papers* (2022). https://collaboratif.cirad.fr/alfresco/s/d/workspace/SpacesStore/4fdec919-30a2-46f1-8acb-2f0fa28fe5f8?attach=true

Données de la recherche – contexte juridique : qui a les droits, quelles obligations ? Doi : 10.13143/8dh5-d615 https://doranum.fr/aspects-juridiques-ethiques/qui-a-les-droits-quelles-obligations_10_13143_8dh5-d615/

Gaillard, R., *De l'open data à l'open research data : quelle(s) politique(s) pour les données de la recherche* (2014). https://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf

Hadrossek, C., Janik, J., Libes, M., Louvet, V., Quidoz, M-C., *et al.*, *Guide de bonnes pratiques sur la gestion des données de la Recherche* (2023). https://hal.science/hal-03152732

Inist-CNRS, DoRANum. *Métadonnées, standards, formats : fiche synthétique* (2017). https://doranum.fr/metadonnees-standards-formats/metadonnees-standards-formats-fiche-synthetique_10_13143_vbjs-6288/

Inist-CNRS, DoRANum. *Les principes FAIR* (2019). https://doranum.fr/enjeux-benefices/principes-fair_10_13143_z7s6-ed26/

Ministère de l'Enseignement supérieur et de la Recherche. *Feuille de route 2021-2024 sur la politique des données, des algorithmes et des codes sources*. https://www.enseignementsup-recherche.gouv.fr/fr/la-feuille-de-route-2021-2024-du-mesri-sur-la-politique-des-donnees-des-algorithmes-et-des-codes-50534

Ministère de l'Enseignement supérieur et de la Recherche. *Guide pratique pour une harmonisation internationale de la gestion des données de recherche* (2019). https://www.ouvrirlascience.fr/science-europe-guide-pratique-pour-une-harmonisation-internationale-de-la-gestion-des-donnees-de-recherche/

Organisation de coopération et de développement économique (OCDE), *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*, Paris, 2007, http://www.oecd.org/fr/science/sci-tech/38500823.pdf

Philippe, O., Rennes, S., Szabo, D., Martel, A-S., *Ouverture des données : ... Aussi ouvert que possible ... aussi fermé que nécessaire* (2022). https://hal.inrae.fr/hal-03659484

Science Europe. « *Criteria for the Selection of Trustworthy Repositories* ». [consulté le 10/07/203]. https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.scienceeurope.org%2Fmedia%2Ffffkb51ei%2Fse-rdm-template-2-criteria-for-the-selection-of-trustworthy-repositories.docx&wdOrigin=BROWSELINK

Scripps CO2 Program: Carbon Dioxide Measurements. [consulté le 10/07/2023] https://scrippsco2.ucsd.edu/

Soulimane, G., Bouchiha, D., Benslimane, M.S., *La ré-ingénierie des ontologies : État de l'art*. Conférence internationale sur l'informatique et ses applications, CIIA'2006, May 2006, Saida, Algérie. https://hal.science/hal-01585047

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.*, *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

GLOSSARY

Catalogue of (meta)data: inventory of data or metadata to make them findable.

Controlled vocabulary: reasoned and standardised lexicon designed to facilitate document searches and comparative analysis of data (list of keywords, glossary, thesaurus, taxonomy, ontology).

Curation (when submitting a dataset)**:** scientific curation involves cleaning, editing, and transforming the data with the aim of obtaining "clean", legible datasets that are easier to process. There is also documentary and technical curation, which involve verifying the metadata of data files to be stored in a repository with the aim of suggesting changes and improving the quality of the dataset descriptions.

Data repository: online service for the submission, description, search and dissemination of datasets. They can be multi-disciplinary or disciplinary. When they satisfy a series of criteria set out in the guide ▼**Criteria for the Selection of Trustworthy Repositories**, they are awarded a certification label designed to promote reliable and lasting data repositories.

Data paper: a publication that describes a scientific dataset, especially using structured information known as metadata.

Data Protection Officer (DPO): person responsible for protecting personal data within an organisation.

Dataset: aggregation, in a legible format, of raw or derivative data that presents a certain unity, combined to form a coherent whole.

Embargo: period during which the articles and research data stored on a platform are not freely available.

FAIR principles: these aim to make data findable, accessible, interoperable and reusable.

General Data Protection Regulation (GDPR): legal framework laid down by the European Union for the management of personal data. See: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&qid=1696951287977

Indexing: assigning distinct terms (keywords for example) to a document providing information about its content and making it easier to find.

Interoperability: capacity of different computer systems to dialogue with one another, exchange data, communicate without ambiguity and thereby interpret information correctly.

French Law for a Digital Republic *(Loi pour une République numérique)***:** French law from 2016 providing a legal framework for depositing certain versions of journal articles in open access repositories, if at least half of the funding came from the public sector. This legislation also treats research data as public data if the work is more than 50% publicly funded and addresses the specific case of partnership-based research.

Licence: legal text setting out the conditions for the dissemination and reuse of a particular output (for example free software licences, Creative Commons).

Linked data: initiative by the World Wide Web Consortium (W3C) to encourage the publication of structured data on the web, not in the form of data silos isolated from one another, but linked up to create a global information network.

Metadata: a set of structured information that describes, specifies and localises a resource with the aim of facilitating its findability, usage and management.

Persistent identifier (PID): a unique and stable reference for a digital object or topic (dataset, article, author, etc.). Examples: digital object identifier (DOI) and Open Researcher and Contributor ID (ORCID).

Personal data: data relating to a individual who is identified or identifiable, for example in correlation with other datasets.

Provenance: Provenance is information about entities, activities, and people involved in producing a piece of data or object, which can be used to form assessments about its quality, reliability or trustworthiness. (source: https://www.w3.org/TR/prov-overview/).

Reuse: use by any individual or entity of published data for purposes other than those for which it was produced or received.

Credits

Direction of publication
Ministry of Higher Education and Research

Editorial coordination
University of Lille

Scientific council
The Skills and Training College
and the Research Data College from
the Committee for Open Science

Project leader
Mónica Michel Rodríguez

Writers
Florence Bouchet Moneret, Romane Coutanson, Amélie Fiocca, Céline Hernandez, Alicia León y Barella, Émilie Lerigoleur, Mónica Michel Rodríguez, Pierrette Paillassard, Agnès Robin, Sara Tandar, Laura Tomasso

To create this guide, the working group drew from the educational content on the DoRANum, CoopIST, URFIST and COUPERIN platforms.

Translator
Myles O'Byrne
Proofread by Jennifer Morival

Graphic design
Studio 4 minutes 34
Studio Lendroit.com

Printing
L'Artésienne, Liévin

1st edition: February 2024
Printing finished on February 2024
5000 copies distributed

Thanks

The early-career researchers who shared their experiences of open science

Chedid Saade, Juliette Godet, Naomi Truan and Joshua Gobe

The PhD students who took part in discussions on the first edition of the guide

Marion Duthoit, Mathis Bachelot, Céline Barzun, Maxime Bedez, Paul Bellevile, Joan Bienaimé, Mélanie Bossu, Fabien Clouse, Violaine Courrier, Violette Delforge, Benjamin Demassieux, Meriam Meziani, Alexandre Van Outryve, Perrine Seguin

Experts consulted

Cécile Arènes, Alexis Arnaud, Flora Badin, Anne Baillot, Nathalie Barré-Lemaire, Laetitia Bracco, Marie Cros, Alina Danciu, Romain David, Delphine Du Pasquier, Emmanuelle Frenoux, Gaëlle Gauvrit, Candice Hector, Frédéric de Lamotte, Sylvie Le Bras, Gaëlle Leroux, Li Ling, Didier Mallarino, Gilles Mathieu, Lionel Maurel, Christelle Pierkot, Céline Rousselot.

This guide is part of the Passport for Open Science collection.
The digital version of this guide is available at www.ouvrirlascience.fr.

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté
Égalité
Fraternité

Ouvrir la science !

Université de Lille