Introduction to Research Data Management & Data Management Plans for Qualitative Data

Doctoral School of Social Sciences + FWB Data Ambassadors – April 8th, 2025 - ULiège

Judith BIERNAUX
ULiège
Research Data Officer
jbiernaux@uliege.be

Joëlle DESTERBECQ
UCLouvain
Research Data Officer
joelle.desterbecq@uclouvain.be

Jonathan DEDONDER
UCLouvain - IACCHOS
Research Logistician
jonathan.dedonder@uclouvain.be

Ferdinand TEUBER
UCLouvain - ISPOLE
Research Logistician
ferdinand.teuber@uclouvain.be

Christophe LEJEUNE
ULiège
Professeur Associé
christophe.lejeune@uliege.be

Introduction to Research Data Management & Data Management Plans for Qualitative Data

Doctoral School of Social Sciences + FWB Data Ambassadors – April 8th, 2025 - ULiège

Disclaimer

L'objectif de cette journée est une séance de formation au data management et aux DMP, à leur plus-value, et leur **application** aux données qualitatives.

Il n'est pas une formation aux méthodes d'analyse qualitative, qui sont nombreuses, diverses, complémentaires, et abondamment traitées dans d'autres sources.

Discussion

Who has re-used qualitative dataset? (Either from literature, archive, or own team)? What materials exactly?

Why won't you / can't you **reuse** qualitative dataset? What **difficulties** do you face?

Research Data

What are research data ?

Factual elements: figures, texts, images, sounds, measurements, results of recordings, computer programmes, etc.

Raw (i.e. not processed, manipulated or transformed in any way) **or derived from raw data** (i.e. obtained after transformation of raw data)

Quantitative (figures, measurements, statistics, survey answers) <u>or</u>

Qualitative (interview, speeches, recordings of speakers, videos)

On which the research is grounded

Necessary according to the scientific community to **<u>validate the results</u>** of the research

Can be stored on any **support** (paper, digital, etc.) and in any **format** (.png, .mpeg, .svg, .wma, .pdf, .txt, .xml, etc.)".

DORANum, Inist-CNRS, Définition des données de la recherche.
DOI : 10.13143/B4ZC-8E79. Nous traduisons

Research Data

A diversity of research data !

Research Data can take a diversity of **forms** and **formats**

- Figures and measurements
- Observational data
- Interviews, surveys
- Texts
- Drawings, maps or plans
- Audiovisual
- Photographs
- Experimental data
- Geospatial data
- Medical imagery
- Code, etc.

→ in all kind of file formats (.png, .mpeg, .svg, .wma, .pdf, .txt, .xml, etc.) .xml, etc.)

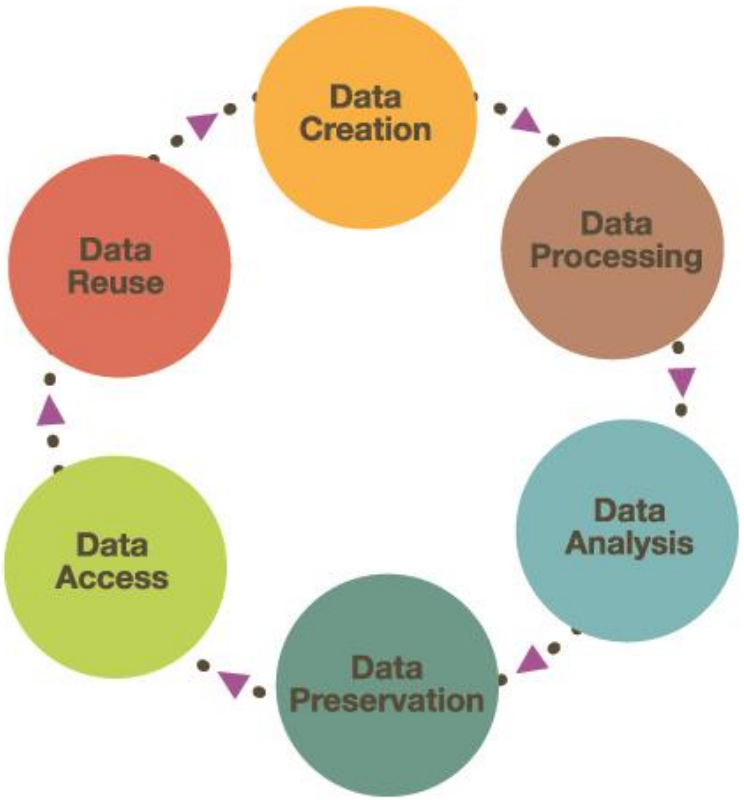
Data Management

Data Management

Data Management
The action of collecting, organising, storing, processing, analysing and sharing research data

Data management Plan

- ✓ A DMP is a **management tool**. Its purpose is to summarize the **description and evolution of the data sets** in your research project.
- ✓ It includes **every steps of research data lifecycle**
  - It prepares your data for **sharing, re-use and long-term preservation.**
- ✓ It helps **navigating** the specificities of your datasets: regulations, privacy, ethical concerns, storage needs, publication possibilities, costs, …
- ✓ The DMP is continually **updated !** A DMP is a living thing, it can evolve along the research project.



Data Management Plan

The DMP in practice :

- A DMP is a set of questions, usually web-based, that works as a **checklist of attention points** to guide the researcher through the data lifecycle.
- More and more **funders require** a DMP to be drawn up
- Some funders require to **fill in template DMPs** as deliverables, and those are therefore reviewed
- Some DMP templates are made **available online**, as examples, for researchers to use without any obligation or without any review
- These online tools usually **provide guidance** and examples of best practice

FNRS Template :

1. Data collection / description
2. Data documentation and data quality
3. Data storage and back-up
4. Ethical and legal requirements
5. Data sharing and preservation
6. Responsibilities and resources

Data Management Plan

You do not have to memorize everything: try writing a Data Management Plan!

- Log into **dmponline.be** with your university SSO
- Select a **template** (if none: generic or HE as generic)
- Read through the list of **questions** and start planning
- Ask your RDO for help :)

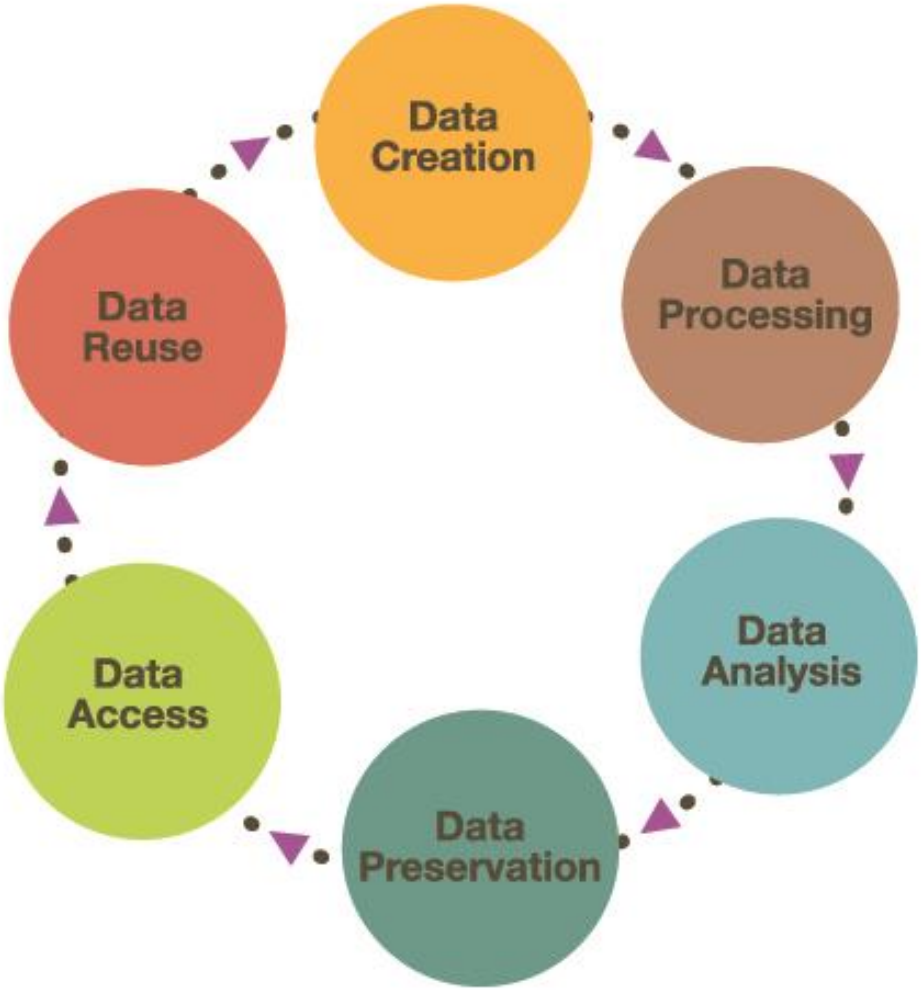
Data Management benefits

How does it helps me ?
Besides supporting your data planning:

Funders obligation
Editors demand

Reproducibility crisis
Digital transition
Legal and ethical considerations
Engagement: science should be public and transparent

Publication is only about 20% of RDM!



Open Data and RDM benefits

Most European funding agencies encourage sharing scientific results, methods and data.
They refer to the **« as open as possible, as closed as necessary »** principle.

Increases **visibility**, impact and improves reputation

Facilitates scientific **reproducibility**

Ensures data **sustainability,** even after staff rotation for example

Optimises return on investment from funders

Selection criterion for some funding programs (or awards)

Facilitates **collaboration**

Open data sharing accelerates COVID-19 research

Artist's impression of COVID-19 open access data sharing. Credit: Spencer Phillips

Summary

- Open access increases the visibility of research data and information, giving scientists the ability to build upon and react to existing research quickly
- EMBL-EBI launched the European COVID-19 Data Platform to enable rapid access to datasets and results pertaining to the SARS-CoV-2 outbreak
- Open access data sharing has greatly accelerated COVID-19 research and helps further our understanding of the biology, transmission, and spread of the SARS-CoV-2 virus

Victoria Hatch, EMBL-EBI News, Oct 19, 2020

Open and FAIR Data

Open Data :
"Open data is data that can be **freely used, re-used and redistributed** by anyone – subject only, at most, to the requirement to attribute and share-alike"

Open Knowledge foundation, Open Data Handbook.
https://opendatahandbook.org/guide/en/what-is-open-data/

≠ data available on the internet

https://book.fosteropenscience.eu/

FAIR Data principles

Findable

Data are **discoverable** and easy to find in a non-equivocal manner, by both humans and computers.

Accessible

Data are made available in a **sustainable** way, even after the project is over.
Users know **how to access** data.

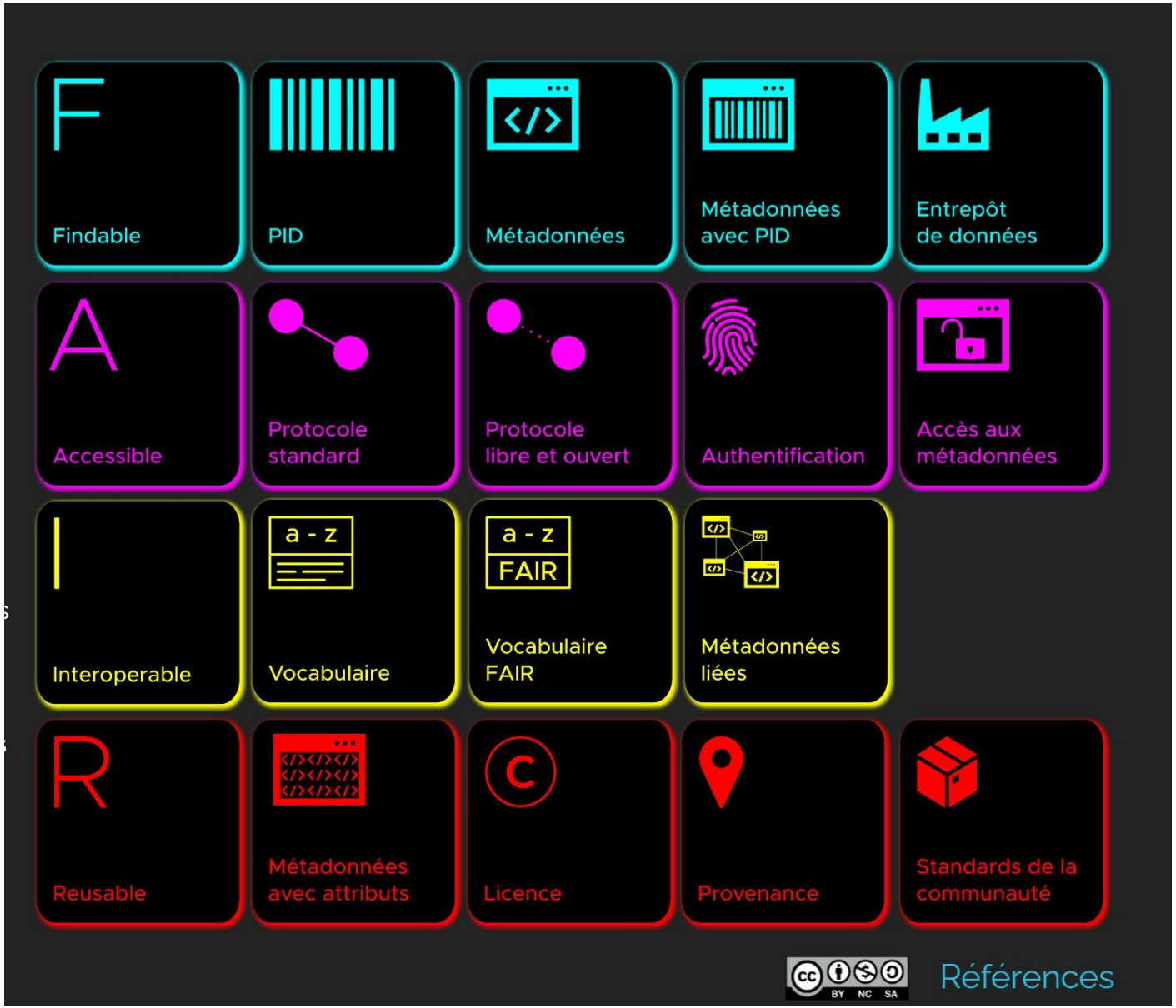
Interoperable

Data can be **operated** / exchanged / compared between a **variety** of institutions, workflows, software, applications, systems, ...

Reusable

The data are **sufficiently described** and can be shared with as few restrictions as possible, as the ultimate goal is to optimise data reuse.
A **clear license** defines the conditions for reuse.

FAIR Data principles



Degrees of data sharing

- The aim is to practice as much **open data** as possible : **"as open as possible, as closed as necessary"**
- But open data is **not a panacea**, not even an obligation !
- The **obligation** is for data **to be FAIR**
- There are **different degrees** of data sharing

Degrees of data sharing

The aim is to practice as much **open data** as possible : "**as open as possible, as closed as necessary**"

Data that cannot be shared : closed

Restricted or controlled data

Open data

FAIR data

Personal and sensitive data **(GDPR)**
Confidential data or data subject to secrecy (e.g. trade secret)
Third-party data
Data protected by copyright and / or database right

Data with commercial / economic valorisation potential
Data with ethical risks for the subject(s)
Data with a shared responsibility (consortium, contractors, …)
Large datasets (i.e. text mining, web scraping, …)

Data with no restriction

Open Data and FAIR Data

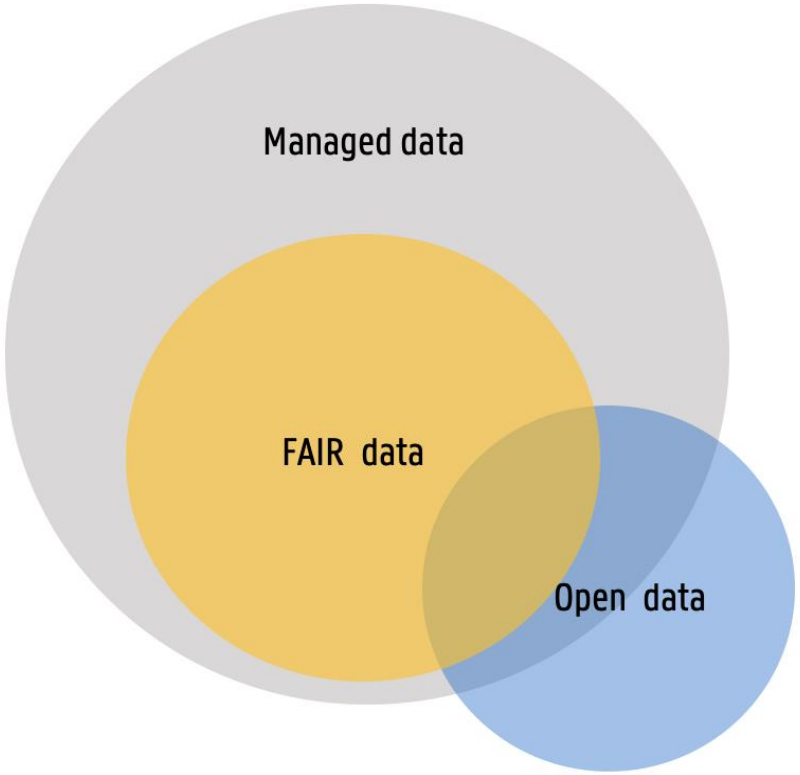


Image: UGent

- FAIR is not the equivalent of Open
- Data **can be FAIR <u>but</u> not Open** (fox example, personal data)
- But open data **needs to be FAIR to be useful**
  - Publicly available data may lack sufficient documentation to meet the FAIR principles, such as licensing for clear reuse
  - Making your data freely and openly available does not translate to it being reusable

Research Data Management, Dr. Sara El-Gebali. https://orcid.org/0000-0003-1378-5495
- https://www.openaire.eu/how-to-make-your-data-fair

Open Data and FAIR Data

- Data management is **not an administrative obligation**, it is a (new) practice in research communities.
- It follows the ever-evolving community standards of **digitization**, **complexification** and **sharing**, which are reflected in funders and editors demands.
-
- **The mostly used standard are the FAIR principles** (which stem from an IT perspective)
- It is not a pressure and certainly **not an injunction to openness**, there are degrees of data sharing
- It can translate in many forms, over which **you have agency:** FAIRness also applies to methods, protocols, results, software, even physical objects.

Check out this recent webinar about open qualitative research challenges!

Especially in human sciences…

Qualitative data
= data which do not come from a measurement

Different types of qualitative data


Audio files


Images and videos


Maps


Tables, excel/csv files


Texts and books, physical or pdf


Work of art, movie, painting, song…

Social and Humanities data = Among the most important to preserve, because they are unique, difficult to replicate, and are much valuable for teaching, research and historical research (long term)

Especially in human sciences…

Qualitative approaches: deductive or inductive or… abductive

INDUCTIVE: Based on data

1. Iterative procedure
2. Categories creation until saturation is reached
3. Build distance from pre-existing categories
4. Be mindful of conceptual contamination

DEDUCTIVE: based on theories

1. Design a theory, and relations between concepts
2. Pre-existing categories, that are completed, specified or modified

ABDUCTIVE: back and forth between data & theory

1. Starts with a set of existing theories
2. Identifies theoretically anomalous empirical cases
3. New tentative explanations
4. To be tested on existing and new data

Especially in human sciences…

Qualitative data are special

- More difficult to **anonymize**
- More complicated to **select**
- More difficult to **plan (less predictable)**
- **Raw and curated** data are very different
- Important data: what is said, but also what is **NOT said**
- **Intimacy** between the researcher and his/her data: implicated, immersive experience, empathy - **different culture** of data and data sharing
- **Plenty** of data collection methods and theories: so this training cannot be generalized to all qualitative approaches

Especially in human sciences…

Qualitative data are special

When dealing with human, qualitative data, there are some layers of complication:

- **Legal compliance** to GDPR and anonymisation

Any type of human data is subject to GDPR and anonymisation is a very tricky thing to fully achieve in the Big Data context

-> On today's menu ☺

- **Project inheritance/longitudinal studies**

Some projects and data in qualitative studies are run in the very long term, and might go from one senior researcher to her/his student

-> Risks of data loss (can be mitigated with good RDM diligence)

=> Today's take-home message = what you can do at each step of the data lifecycle to ensure good RDM with your qualitative data

Data collection

Data quality control

Data collection: some best practices to ensure quality

- Using standardized methods and protocols (ethical, institutional)
- Calibration of questionnaires (languages), focus group guides, interview/case protocols
- Taking several observations or samples
- Checking the truth of the record with an expert
- Keep a distance with your subject

Data collection

Data quality control

Data collection: Interviews and Focus groups

1. **Interview:** two people, more organized, understandable
2. **Focus groups**:
   a. Conversation between many people (1 ID per respondent)
   b. (Dis)agreeing with each other, non verbal: takes more time to transcribe and analyse
   c. To help with the analysis:
   Two people involved: the animator (guide the discussions) and observer (field notes): helps who said what, and contextualize data (non verbal)
   Write a focus group report, summarize interactions

Data collection

Data quality control

Transcription

When data have been collected, it's time to transcribe them in a text format.

- If you can externalize transcription, it's fine; but try to transcribe at least one interview.
- Automatic transcription tools (Whisper, office, etc.) can help, but re-listening is essential to ensure accuracy and capture nuances.
- Take time to discuss with the transcriber, and tell exactly what you expect (laughs, hesitations, emotions)
- Number the line of text
- Pay attention to paralinguistic cues (laughs, pauses, sighs, tone shifts)—they often carry important meaning.

Jonathan : Le Research Data Management, c'est indispensable aujourd'hui. Mais on manque encore de directives claires...

Ferdinand : Oui, et surtout, la mise en... hum... application reste floue. Qui est responsable de la gestion des données à long terme ? *((lève les mains en l'air, interrogateur))*

Joëlle : ((*hoche la tête*)) Les chercheurs collectent, stockent... mais après ? Qui garantit la pérennité des données ?

Judith : Il faudrait une formation obligatoire. Trop de chercheurs ne savent pas organiser leurs datasets !

Christophe = Oui, et même quand ils savent, il y a toujours des différences d'un labo à l'autre...

Jonathan = C'est ça ! [On parle d'harmonisation, mais dans les faits...

Ferdinand : [C'est un chaos complet ! ((*incompréhensible en allemand*))

Joëlle : ((*soupire*)) Un chaos organisé, peut-être... mais un chaos quand même.

Christophe : Et sans cadre clair, on continuera à bricoler. ***Ad hoc***, toujours dans l'urgence...

Judith = Exactement ! [On nous demande de tout archiver, mais avec quels moyens ?

Jonathan : [Et puis après cinq ans, plus de serveur, plus rien... ***Gone with the wind***.

Ferdinand : ((*hoche la tête*)) Et personne pour assumer la responsabilité...

Data collection

Data quality control

Data checking

1. **Credibility** – Reflects the participant's reality (e.g. triangulation, member checking)
2. **Transferability** – Thick description to allow transfer to other contexts
3. **Dependability** – Transparent and documented research process
4. **Confirmability** – Findings shaped by participants experiences, not researcher bias
5. **Saturation** – Stop collecting data when all dimensions have been discovered; ensures depth, not just quantity
6. **Reflexivity** – Ongoing self-awareness of the researcher's role, influence, and assumptions in the research process

Qualitative Data Analysis Software

Multi-sources
Text (word, pdf, txt, etc.)
Audio and video
Images
Literature review

Time-saving
All data in one place
Large among of data
Images
Literature review

Theoretically neutral
Ethnography Text
Grounded theory method
IPA
Mixed methods

Useful software : CAQDAS

NVivo, Atlas.ti, MaxQDA, Dedoose...

Re-useability
Collaboration tools
Transparency
Not an analysis tool
Documentation

Metadata
Codebook
Individual characteristics
Fields notes
Personnal memos

Qualitative Data Analysis Software

Multi-sources

- Text (word, pdf, txt, etc.)
- Audio and video
- Images
- Literature review

Time-saving

- All data in one place
- Large among of data
- Images
- Literature revi

Theoretically neutral

- Ethnography Text
- d

Save time compared to manual coding (no data losses, quick on finding verbatims, makes analysis easier), which leaves more time to data interpretation

NVivo, Atlas.ti, MaxQDA, Dedoose...

Usability

- Collaboration tools
- Transparency
- Not an analysis tool
- Documentation

Metadata

- Codebook
- Individual characteristics
- Fields notes
- Personnal memos

Qualitative Data Analysis Software

Historique • The Ethnograph • NUD·IST • Kwalitan • Weft·QDA (FLOSS)	Classique • NVivo • Atlas·ti • HyperRESEARCH • Quirkos • Tams Analyser (FLOSS) • Taguette (FLOSS)
Mixte • MaxQDA • Provalis QDA Miner • RQDA (FLOSS)	Collaboratoire • Dedoose • Saturate • Cassandre (FLOSS)

https://hdl.handle.net/2268/206219

Data collection

Primary and secondary data

You can collect your data yourself… and/or check if there are any existing data that you can re-use!

Reuse of qualitative data (« secondary analysis ») can be an interesting option to the degree that it does not contradict your epistemological core tenets.

Introduce/ discuss	Write a research proposal and build your case on data from several datasets.
Efficency	Limit the data collection expense in using existing data (and test your hypothesis to them).
Originality	Secondary analysis benefits from the fact that primary data are richer than necessary to answer the original RQ. It can be a means to analyze social processes that have unfolded (unnoticed) over time and across space.
Compare/ discuss	Compare or discuss your research results with similar data, collected in other time/places, or with different methods

Data collection

A concrete example for the reuse of qualitative data in political science:

- ERC StG "Qualidem" (Grant No. 716208; 2017-2023; https://qualidem-erc.eu); PI V. Van Ingelgom & Co-PI C. Dupuy (ISPOLE, UCLouvain)
- **Goal:** Analysis the link between changes in public policy (e.g. Europeanisation & turn to neoliberalism) and evolutions of citizens' democratic linkages (e.g. political trust, political support, loyalty) from the 1990s to the 2010s.
- **Challenge:** How to retrospectively get longitudinal and comparative qualitative data on citizens' policy perceptions and experiences over the past thirty years?

Qualidem

Eroding Democracies - ERC Starting Grant

Data collection

- **Solution:** Collaborative re-analysis of existing qualitative data (31 individual & 47 collective interviews) by research team

Primary dataset	Primary data collection	Research topic	Type of interviews	Cross-national comparison*	Social composition	Number of groups/ interviews and participants in our dataset
Belot (2000)	1995-1996	Citizens' attitudes towards European integration	Semi-structured interviews	France and the UK	Young adults (3 categories of age) from varying socio-economic background (different level of education and coming from different regions of UK and France)	31 participants in individual interviews
CITAE (Duchesne et al., 2013)	2006	Citizens' reactions towards European integration	Focus groups	Belgium, France and the UK	Participants from varying socio-economic background (working class, white collars, managers)	24 FG and 133 participants
Mercenier (2019)	2014	Citizens' perceptions of the EU and their relationships to politics	Focus groups	Belgium	Young adults from different neighbourhoods with distinct socio-demographics	6 FG and 35 participants
RESTEP (Beaudonnet et al., 2022)	2019	Citizens' politicization of EU issues	Focus groups	Belgium, France, Italy and Portugal	Participants from varying socio-economic background (high and low education levels) – including students	14 FG and 69 participants
WelfSOC (Taylor-Gooby and Leruth, 2018)	2015-2016	Citizens' welfare state preferences	Democratic forums and focus groups	Denmark, Germany, Norway, Slovenia, and the UK	Participants from varying socio-economic background (self-employed, unemployed, ethnic minority)	3 FG and 34 participants

* Data from underlined countries are part of our secondary corpus.

Qualidem

Eroding Democracies - ERC Starting Grant

From Dupuy et al. 2022, p. 134.

Data collection

Primary and secondary data

You can collect your data yourself… and/or check if there are any existing data that you can re-use!

Consult institution data repositories or other relevant data repositories:

https://www.re3data.org/ - browsing by subject
https://commons.datacite.org/ -> keyword based search
https://doranum.fr/depot-entrepots/depot-et-entrepots-fiche-synthetique_10_13143_a3d4-7553/

Whyte, A., *Where to keep research data: DCC checklist for evaluating data repositories* (v.1), Edinburgh: Digital Curation Centre, 2015. https://www.dcc.ac.uk/guidance/how-guides/where-keep-research-data

Some examples

- EOSC: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
- SODHA : https://www.sodha.be/
- Bequali: https://bequali.fr/en/
- Qualitative Data Repository: https://qdr.syr.edu/
- TROLLING (Linguistics)
- Qualidatanet: https://www.qualidatanet.com/en/
- UK Data Service: https://ukdataservice.ac.uk/

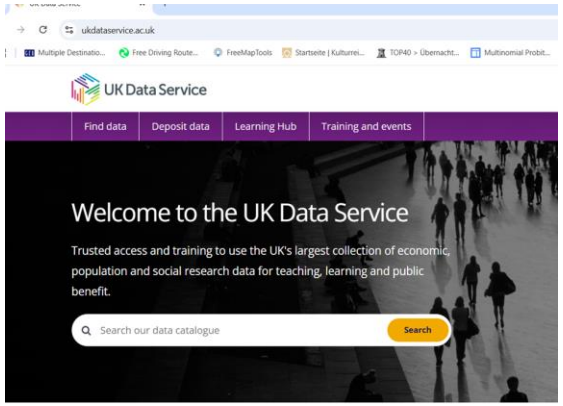
Or contact the authors

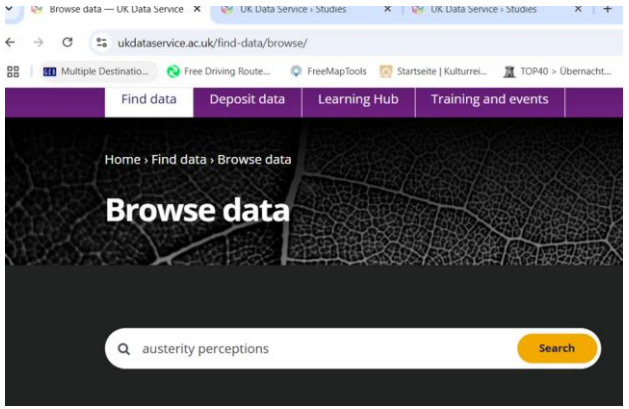
Always check the quality, read metadata and documentation. Check with an expert.

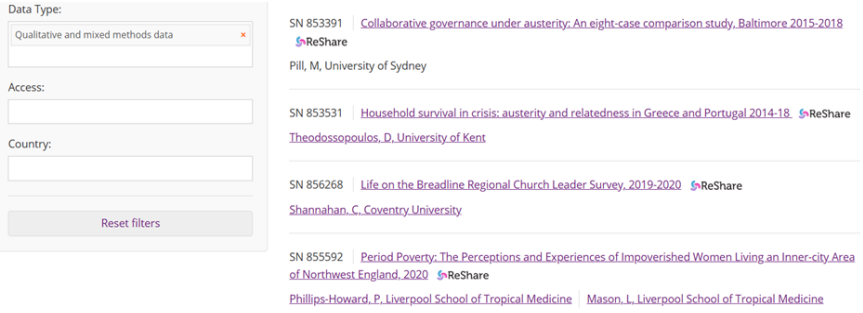
Data collection

A short illustration for finding secondary data

Imagine that, in the context of a project, you want to re-analyse some data on the perception of austerity policies in a liberal welfare regime in the midst of economic crisis. What can I do?







Public perceptions of threat in Britain: Security in an age of austerity

Details | Access data

Details

Title:	Public perceptions of threat in Britain: Security in an age of austerity
Study number (SN):	851004
Access:	These data are safeguarded
Persistent identifier:	10.5255/UKDA-SN-851004
Data creator(s):	Stevens, D, University of Exeter

Coverage and methodology

Dates of fieldwork:	31 March 2012 - 31 July 2013
Country:	United Kingdom
Observation units:	Group Individual
Kind of data:	Numeric Text
Method of data collection:	1. Method: Internet survey. Sampling procedure: British citizens over 18 from ICM internet panel. Observation units: individuals. Data files: 1. Cases: 2004, including booster sample of 251 British Muslims. Variables: 756. 2. Method: Mini-focus groups. Observation units: individuals in groups of 3. Data files: 20 transcripts.

Open and FAIR Data

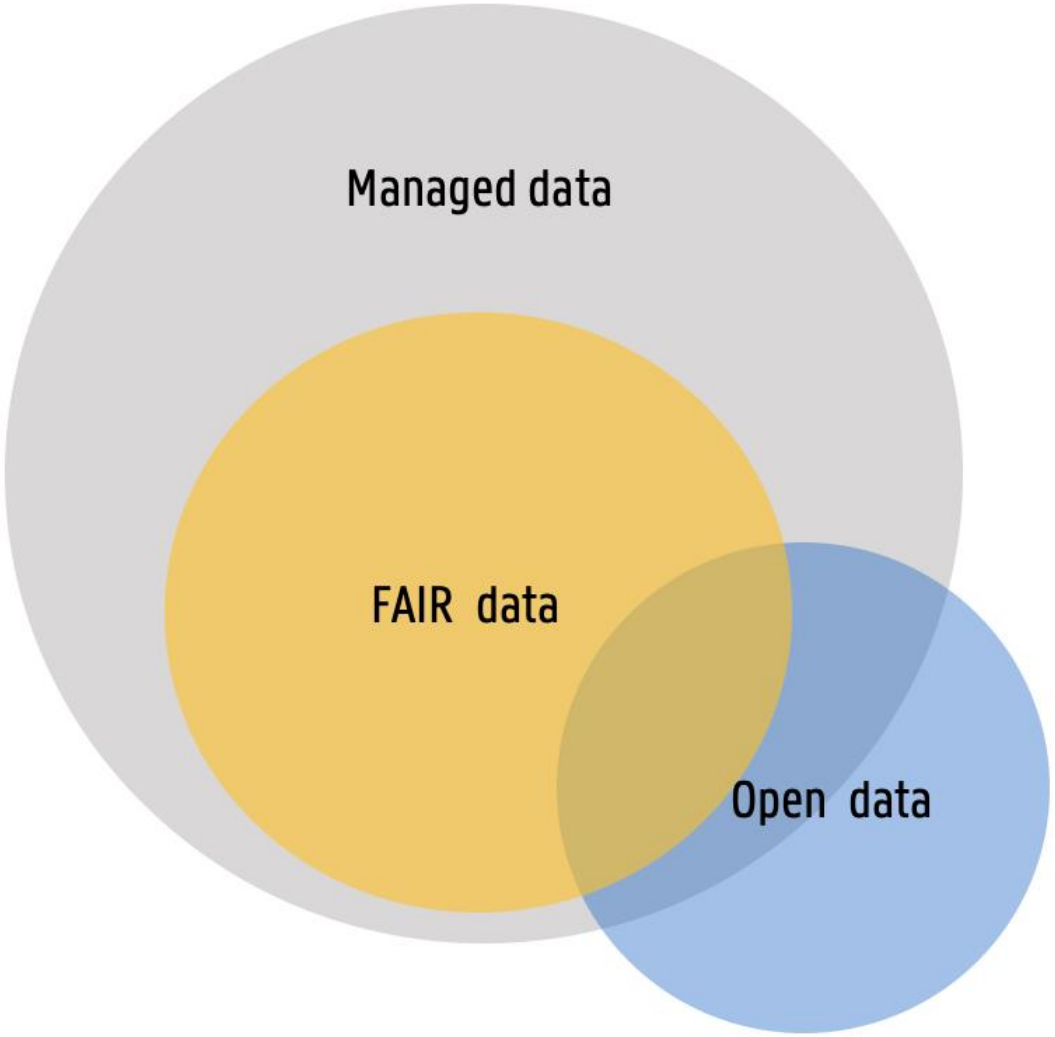
Open Data :
"Open data is data that can be **freely used, re-used and redistributed** by anyone – subject only, at most, to the requirement to attribute and share-alike"

Open Knowledge foundation, Open Data Handbook.
https://opendatahandbook.org/guide/en/what-is-open-data/

≠ data available on the internet

https://book.fosteropenscience.eu/

Open Data and FAIR Data



Message 1: sharing does not necessarily mean opening up everything

FAIR is the standard, data can be FAIR and not open

Message 2 : Open data does not necessarily mean that data can be re-used

Open data needs to be FAIR to be useful

Image: UGent

How to open up data?

Deciding which parts to publish	Selecting a Data Repository	Writing a data paper
Documenting data	Selecting a license	Choosing an embargo period
Completing metadata	Defining specific terms (DSA)	Versioning an existing dataset

Select a Data Repository

"A data repository is an online platform that is used to deposit completed datasets with the purpose to publish, share and/or preserve them. A data repository is database infrastructure that compiles, manages and gives access to data and associated metadata and documentation".

https://www.ugent.be/en/research/openscience/datamanagement/after-research/data-repositories.htm

How to select a good data repository?

A good repository:

- Is recognized by your peers
- Provides a persistent identifier such as a DOI or handle
- Comes with a few possibilities for **licenses** (CC, …)
- Has high documentation metadata standards with controlled vocabularies (therefore discipline-specific is usually better)
- Lets you keep all your rights

The whole point is **data FAIRness** – the repository structure nudges you towards filling in **metadata info** (author, date, keyword, references…) but also adding **documentation** (**read me files, …**) and sometimes even explicating **acronyms, measurement units, conventions,** …

https://doranum.fr/metadonnees-standards-formats/cours-introductif-sur-les-metadonnees_10_13143_vwce-g965/

Select a Data Repository

How to select a good data repository?

The Open Science Committee “has defined **a list of exclusion criteria** for **selecting trusted thematic repositories :**

- No moderation of deposits
- No permanent identifier
- No guarantee of infrastructure continuity
- Property Rights transfer
- Excessive pricing policy
- Localisation of data outside the European Union (=> GDPR)
- Repository restricted by institutional affiliation

If these criteria <u>are present</u> (or one of them), it is better <u>not to choose</u> this repository

Comité pour la Science ouverte – Collège Données de la Recherche, Arnould P.-Y. et al., “Sélectionner un entrepôt thématique de confiance pour la diffusion des données de la recherche : note méthodologique”, Ouvrir la Science !, France, Novembre 2023, Licence Creative Commons.

Decide what data to keep

Not everything needs to be shared at all costs, it is not all 'black or white'
Digital Curation Center **Data appraisal – Five steps to decide what data to keep:**

- 1.Consider **potential reuse** purposes - what aims ***could*** the data meet? (verification – validation, further analysis, further publication, learning and teaching, …)
- 2.Check for indications that it ***must*** be kept considering **legal or policy compliance** risks
- 3.Identify which data ***should*** be kept as it may have **long-term value**
  - Could the data have broad appeal e.g. as it relates to a landmark discovery, a significant new research process, or international policy and social concerns?-
- 4. Weigh up the **costs**
  - Which data management costs have already been incurred and therefore contribute to its value, and how much more is planned and affordable? Where will the funds to pay these costs come from?
- 5. Complete your **data appraisal**
  - This will list what data must, should or could be kept to fulfil which potential reuse purposes.

Whyte A., DCC, *Five steps to decide what data to keep: a checklist for appraising research data* (v.1), Edinburgh: Digital Curation Centre, 2014. Direct link : https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep#3

Decide what data to keep

In summary :

- Select data you **want to** publish, and **delete** those you have to (consortium agreement, legal obligations, GDPR requirements).
- For other data, consider their **uniqueness**, long-term **value** and **potential** of reuse.
- Keep certain data to **validate** your publication's results, for future teaching or research – traceability and reproducibility.
- Take also into account the **costs** (time, software, etc.) and efforts required to preserve these data (preparation, documentation, and storage steps).
- Depending on **legal and practical aspects**, you may state a **period of preservation**: some data will be obsolete in 2, 5, 10 or 50 years.

Where do I publish FAIR / open data ?

Two main publication behaviours are observed for data sharing practices :

Data as (annex to) a paper

Data as its own publication

Where do I publish FAIR / open data ?

Data as (annex to) a paper

Advantage of information proximity: the data and the results live together

Where do I publish FAIR / open data ?

Data as (annex to) a paper

Advantage of information proximity: the data and the results live together

Disadvantage of mismatch with the FAIR principles, confusion between paper and dataset metadata

Where do I publish FAIR en data ?

Cornell University

arXiv > physics.data-an

Data Analysis, Statistics and Probability

- Cross-lists
- Replacements

See recent articles

Showing new listings for Wednesday, 20 November 2024

Total of 6 entries
Showing up to 2000 entries per page: fewer | more | all

Cross submissions (showing 1 of 1 entries)

[1] arXiv:2411.11991 (cross-list from cond-mat.stat-mech) [pdf, html, other]

Spectral Coarse-Graining and Rescaling for Preserving Structural and Dynamical Properties in Graphs

M. Schmidt, F. Caccioli, T. Aste

Comments: 7 pages, 5 figures

Subjects: **Statistical Mechanics (cond-mat.stat-mech)**; Disordered Systems and Neural Networks (cond-mat.dis-nn); Biological Physics (physics.bio-ph); Data Analysis, Statistics and Probability (physics.data-an)

We introduce a graph renormalization procedure based on the coarse-grained Laplacian, which generates reduced-complexity representations for characteristic scales identified through th
the analysis of large graphs by decreasing the number of vertices. Applied to graphs derived from EEG recordings of human brain activity, our approach reveals macroscopic properties em
brain activity across scales, with more generalized patterns during rest and more specialized and scale-invariant activity in the occipital lobe during attention-focused tasks.

Replacement submissions (showing 5 of 5 entries)

[2] arXiv:2310.05571 (replaced) [pdf, html, other]

Noise-Robust One-Bit Diffraction Tomography and Optimal Dose Fractionation

Pengwen Chen, Albert Fannjiang

Subjects: **Information Theory (cs.IT)**; Data Analysis, Statistics and Probability (physics.data-an)

This study presents a noise-robust framework for 1-bit diffraction tomography, a novel imaging approach that relies on intensity-only binary measurements obtained through coded aperture
iteration, to effectively recover 3D object structures under high-noise conditions. Theoretical analysis highlights the de-noising capabilities of the 1-bit scheme, with numerical experiments
A key contribution is the investigation of dose fractionation, revealing optimal performance at a signal-to-noise ratio near 1, independent of the total dose. This finding addresses the dose-c
reconstruction process, providing insights into algorithmic convergence and the interplay between eigenvector correlations and spectral gaps.

ive together

between paper and

Where do I publish FAIR open data ?

Where do I publish FAIR / open data ?

Data as (annex to) a paper

Advantage of information proximity: the data and the results live together

Disadvantage of mismatch with the FAIR principles, confusion between paper and dataset metadata

Disadvantage of fuelling the editorial business models (APCs, open access policies, collective and individual costs)

Where do I publish FAIR / open data ?

What about a data paper?

= a scientific article that describe the data you've produced during your research projects, and the management you've done.

- Published on a specific Data Journal, or in disciplinary journals
- Check whether this journal is peer reviewed.
- Useful to refer to a specific/innovative research design, data collection process or management procedures.

Where do I publish FAIR / open data ?

What about a data paper?

= a scientific article that describe the data you've produced during your research projects, and the management you've done.

- Published on a specific Data Journal, or in disciplinary journals
- Check whether this journal is peer reviewed.
- Useful to refer to a specific/innovative research design, data collection process or management procedures.

- A data paper is still a paper, still has paper metadata, and still fuels an editorial business model….
- … but it provides excellent documentation to a dataset that is otherwise published, with an extra citation as bonus !*

*unless it counts as salami slicing?

Where do I publish FAIR / open data ?

Data as its own publication

There are places that are made for sharing research datasets with appropriate metadata structure and standards = research data repositories

Where do I publish FAIR / open data ?

Data as its own publication

There are places that are made for sharing research datasets with appropriate metadata structure and standards = research data repositories

They usually look like a web page where one can upload files and fill in metadata fields, as well as select a sharing licence

Some of them are discipline-specific and some are open to all types of datasets, or even publications

Where do I publish FAIR / open data ?

Data as its own publication

There are places that are made for sharing research datasets with appropriate metadata structure and standards = research data repositories

Where to find specific repositories?

Ask your peers or check out the following links:
https://www.re3data.org/ - browsing by subject
https://commons.datacite.org/ -> keyword based search
https://doranum.fr/depot-entrepots/depot-et-entrepots-fiche-synthetique_10_13143_a3d4-7553/

Whyte, A., *Where to keep research data: DCC checklist for evaluating data repositories* (v.1), Edinburgh: Digital Curation Centre, 2015. https://www.dcc.ac.uk/guidance/how-guides/where-keep-research-data

Some examples

- EOSC: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
- SODHA : https://www.sodha.be/
- Bequali: https://bequali.fr/en/
- Qualitative Data Repository: https://qdr.syr.edu/
- TROLLING (Linguistics)

Where do I publish FAIR / open data ?

Data as its own publication

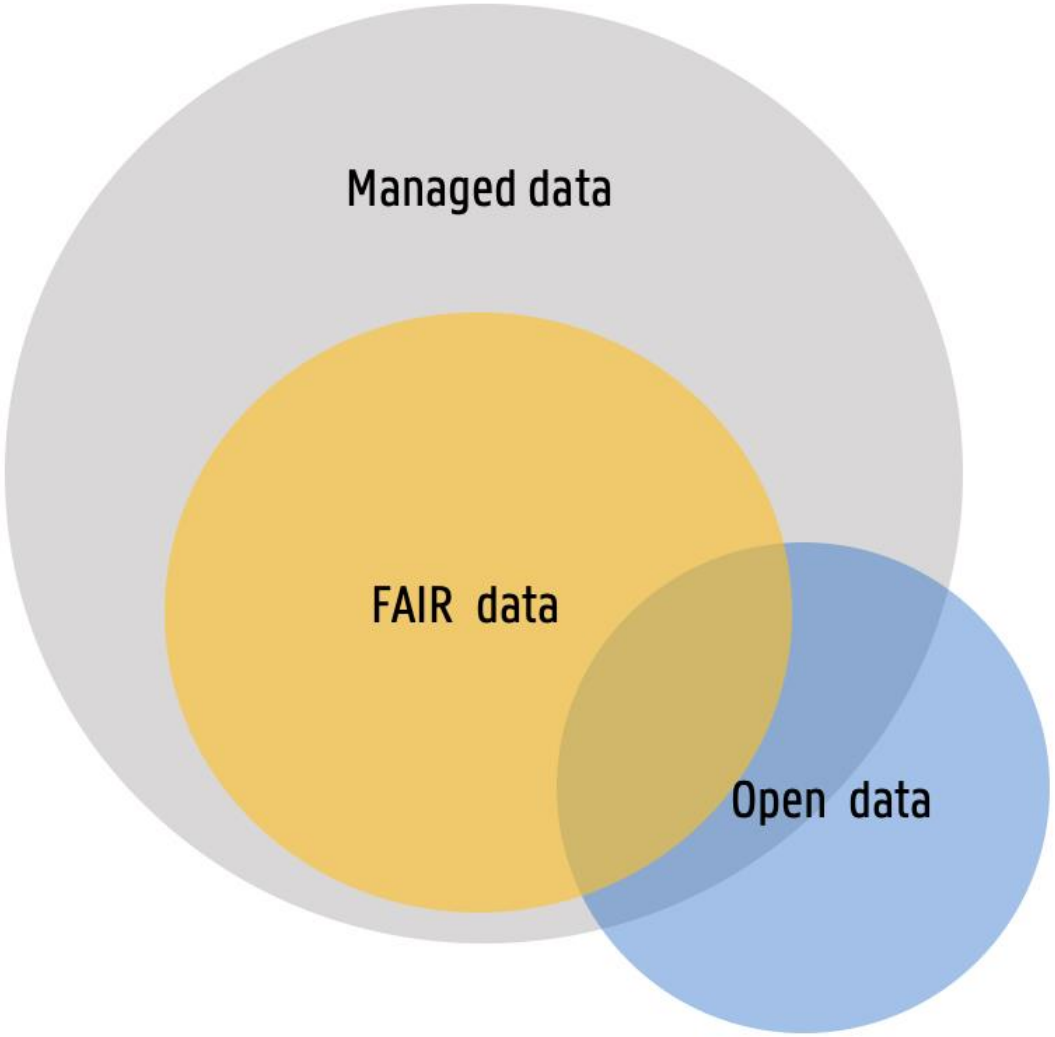
There are places that are made for sharing research datasets with appropriate metadata structure and standards = research data repositories

No specific repository? No problem!
Use a discipline-agnostic data repository.
Your university likely has one!

- ULiège: dataverse.uliege.be
- UCLouvain: Dataverse.uclouvain.be
- UMons: Dataverse under construction

Other ideas: Zenodo, OpenAire, partner institutions repository if applicable (most VL universities have one…)

Open Data and FAIR Data



Message 1: sharing does not necessarily mean opening up everything

FAIR is the standard, data can be FAIR and not open

Message 2: Open data does not necessarily mean that data can be re-used

Open data needs to be FAIR to be useful

Message 3: Achieving FAIRness is mostly about where to share datasets and metadata

A data repository is usually a good option

Image: UGent

Select an embargo period

Still a bit scared ? Select an embargo period

= Delay between the publication of the metadata and of the data files

May help ease into publication to balance between openness and value creation (usually commercial)

Select an embargo period

Still a bit scared ? Select an embargo period

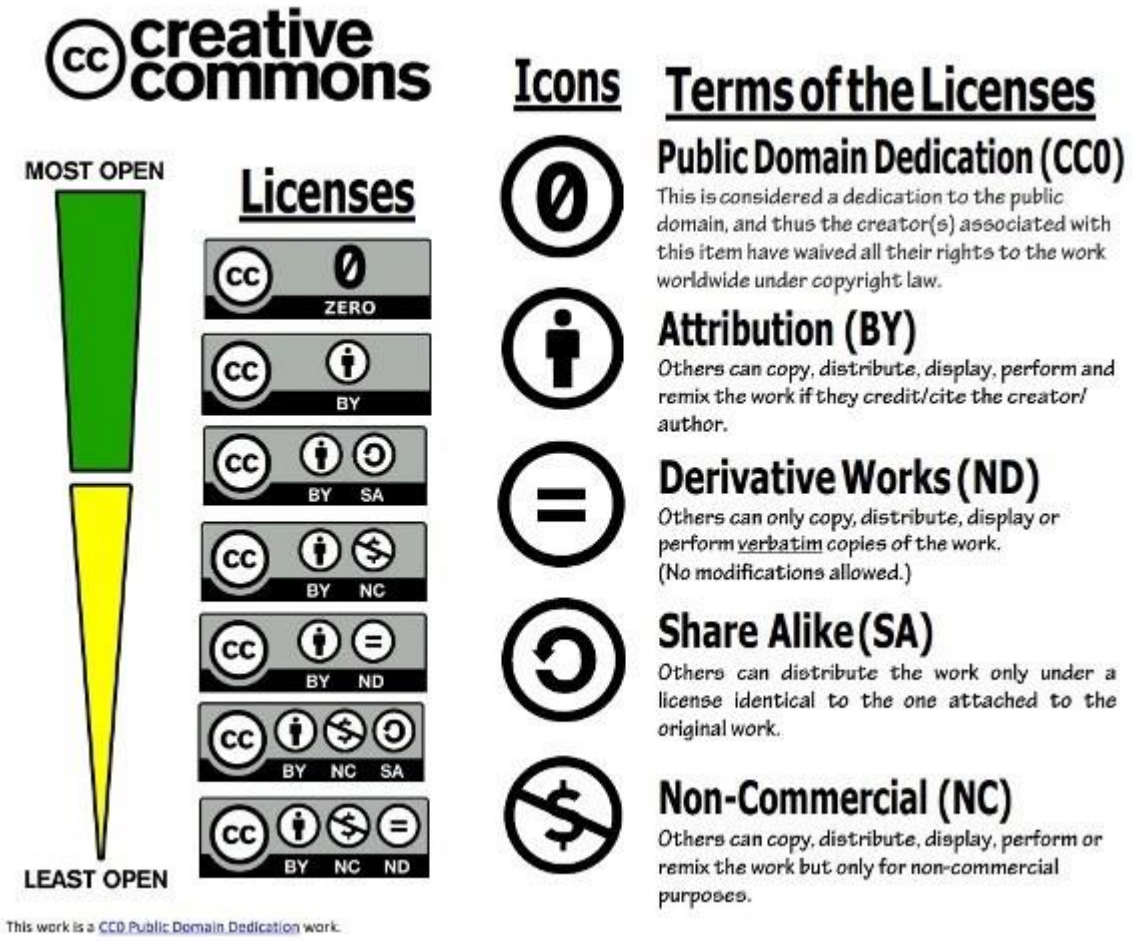
= Delay between the publication of the metadata and of the data files

- Maybe stated in your consortium agreement, funders' contract, patent, etc.
- Must be justified (but several reasons possible; maturity, value creation, strategic, …)
- Must be limited, must have a clear release date
- **Communicate** this embargo period with the repository you chose, and **program it** if possible (usually no early release possible!)

Choose a License or make a DSA

Still a bit scared ? Choose an appropriate license

- A license defines the terms of use of your dataset.
- Usually, open data = permissive CC BY 4.0 license, but there are other possibilities with more restrictions if needed.
- Example: the Creative Commons family



Choose a License or make a DSA

Choose an appropriate license

- A license defines the terms of use of your dataset.
- Usually, open data = permissive CC BY 4.0 license, but there are other possibilities with more restrictions if needed.
- Example: custom terms

British National Corpus User Licence

BNC User Licence

Please read and make sure you have understood the terms of this User Licence. Your use of the BNC is conditional on your acceptance of the terms and conditions specified here. When you have read the terms below, you will be asked to confirm that you accept them.

This Licence Agreement is made between the British National Corpus Consortium (hereinafter termed the BNC Consortium) of the One Part, and you, the reader of this document, of the Other Part,

WHEREAS the BNC Consortium has been created by a consortium of Oxford University Press, Longman Group UK Limited, W & R Chambers Limited, The University of Lancaster, The University of Oxford and The British Library, and

WHEREAS the BNC Consortium has obtained permission from a number of text providers (hereinafter termed the Text Providers) to include a sample of their texts in the British National Corpus, such texts or the categories thereof being listed in Appendix 1 to this Agreement, and

WHEREAS the Licensee is the end user in the manner detailed herein of the texts and/or the categories thereof listed in the said Appendix 1, which end user may be made up of academic researchers or researchers in commercial institutions, and

WHEREAS the Text Providers have empowered the BNC Consortium under a separate agreement to grant a non-exclusive licence to the Licensee as detailed herein, **NOW IT IS HEREBY MUTUALLY AGREED AS FOLLOWS:**

1 Definitions

(a) The "BNC Texts" is a collection of spoken and written texts held on computer and selected for use in language based research and development (hereinafter termed the BNC Texts).

(b) The "BNC Processed Material" is the BNC Texts assembled into consistent electronic format and enhanced with syntactic and/or semantic annotations by the BNC Consortium (hereinafter termed the BNC Processed Material).

(c) "The Licensee's Results" are the results of work performed by the Licensee on the BNC Processed Material in the course of research.

OXFORD TEXT ARCHIVE

FACULTY OF LINGUISTICS, PHILOLOGY AND PHONETICS

- Browse
  - > All of the Repository
- My Account
  - Login
- General Information
  - Deposit
  - Cite
  - Oxford University users
  - Submission Lifecycle
  - FAQ
  - About
  - Help Desk
  - Privacy policy

Choose a License or make a DSA

Choose an appropriate license

A licence defines the terms of use

of your dataset.

Usually, open data = perm

BY 4.0 l

p

if

Exa terms

(f) There is no restriction on the use of the Licensee's Results except that the Licensee may not publish in print or electronic form or exploit commercially in any form whatsoever any extracts from the BNC Processed Material other than those permitted under the fair dealings provision of copyright law.

(g) The BNC Consortium does not grant to the Licensee any rights whatsoever to reproduce the BNC Texts or use all or any part of the BNC Texts in commercial products or services in any way other than would be permitted under the fair dealings provision of copyright law.

British National Cor

BNC

UK Limited, W & R

and

text providers (hereinafter termed the Text Providers) to

such texts or the categories thereof being listed in Appendix 1 to this

the end user in the manner detailed herein of the texts and/or the categories thereof listed in the said

which end user may be made up of academic researchers or researchers in commercial institutions, and

WHEREAS the Text Providers have empowered the BNC Consortium under a separate agreement to grant a non-exclusive licence to the Licensee as detailed herein, **NOW IT IS HEREBY MUTUALLY AGREED AS FOLLOWS:**

1 Definitions

(a) The "BNC Texts" is a collection of spoken and written texts held on computer and selected for use in language based research and development (hereinafter termed the BNC Texts).

(b) The "BNC Processed Material" is the BNC Texts assembled into consistent electronic format and enhanced with syntactic and/or semantic annotations by the BNC Consortium (hereinafter termed the BNC Processed Material).

(c) "The Licensee's Results" are the results of work performed by the Licensee on the BNC Processed Material in the course of research.

General Information

- Deposit
- Cite
- Oxford University users
- Submission Lifecycle
- FAQ
- About
- Help Desk
- Privacy policy

Choose a License or make a DSA

Choose an appropriate license

In some specific cases, a license can be replaced by a **Data Sharing Agreement (DSA)** or **Data Transfer Agreement (DTA)**

This specific contract defines the rights and obligations of the reuser and needs to be singed **before accessing** the data, as a **condition** to download the files for example)

It is drawn if more restrictions than a mere license is needed (personal data, confidential data, third-party data, …)

Choose a License or make a DSA

Make a Data Sharing Agreement

In some specific cases, a license can be replaced by a **Data Sharing Agreement (DSA)** or **Data Transfer Agreement (DTA)**

This specific contract defines the rights and obligations of the reuser and needs to be singed **before accessing** the data, as a **condition** to download the files for example)

It is drawn if more restrictions than a mere license is needed (personal data, confidential data, third-party data, …)

=> If re-users can sign the DSA/DTA, they still need to know that the data is out there – metadata can still be made available

Choose a License or make a DSA

Make a Data Sharing Agreement

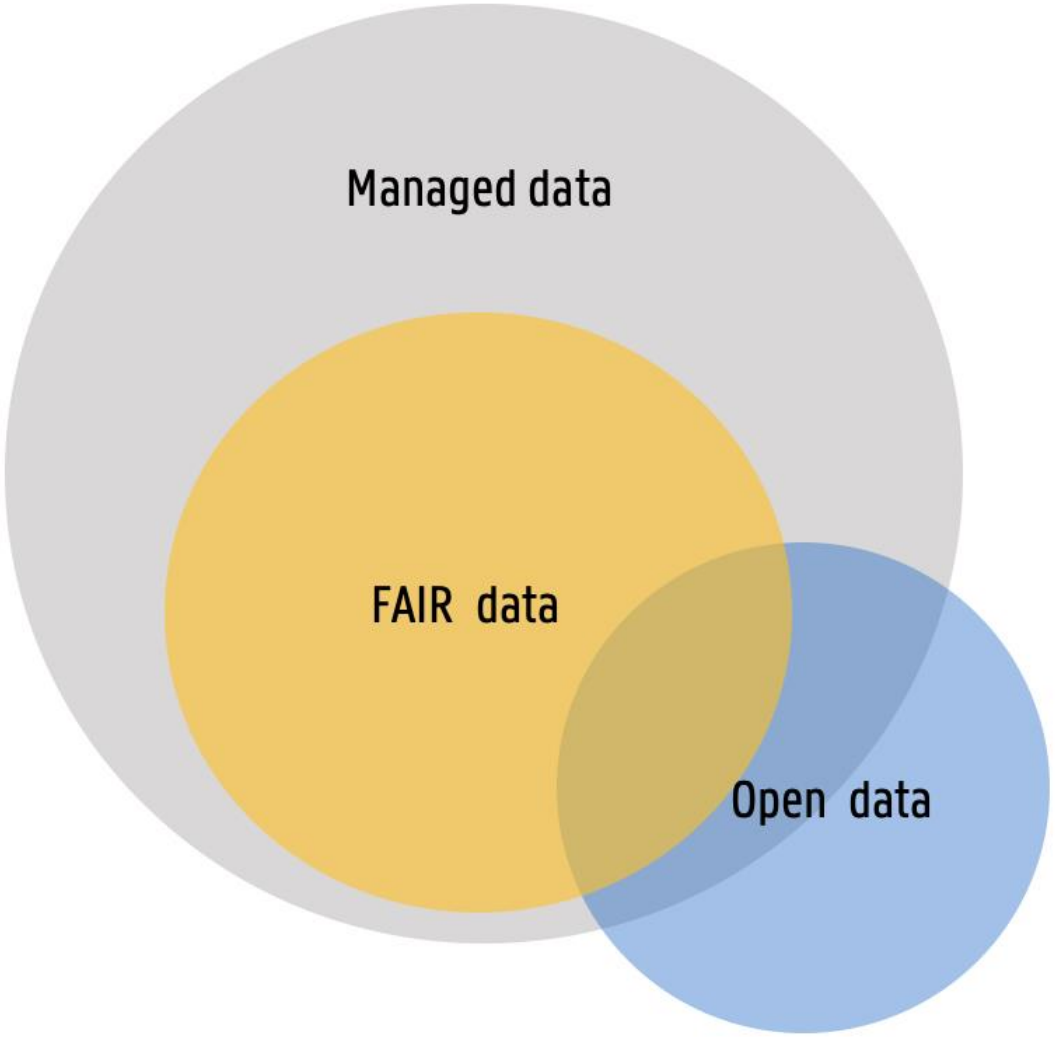
In some specific cases, a license can be replaced by a **Data Sharing Agreement (DSA)** or **Data Transfer Agreement (DTA)**

This specific contract defines the rights and obligations of the reuser and needs to be singed **before accessing** the data, as a **condition** to download the files for example)

It is drawn if more restrictions than a mere license is needed (personal data, confidential data, third-party data, …)

=> If there is a very high level of confidentiality or protection, we can still position our dataset further back in the FAIR continuum

Open Data and FAIR Data



Message 1: sharing does not necessarily mean opening up everything

FAIR is the standard, data can be FAIR and not open

Message 2: Open data does not necessarily mean that data can be re-used

Open data needs to be FAIR to be useful

Message 3: Achieving FAIRness is mostly about where to share datasets and metadata

A data repository is usually a good option

Message 4 : licenses matter, and may help the reluctant easing into data sharing

Image: UGent

Open Data and FAIR Data

Versioning

Once your data is published, it is not necessarily frozen. Most repositories allow for versioning.

High-throughput analysis of Fröhlich-type polaron models

Version 2.0

Melo, Pedro; Abreu, Joao; Verstraete, Matthieu; Guster, Bogdan; Giantomassi, Matteo; Gonze, Xavier; Zanolli, Zeila, 2023, "High-throughput analysis of Fröhlich-type polaron models", https://doi.org/10.58119/ULG/UOWRSL, ULiège Open Data Repository, V2

Cite Dataset Learn about Data Citation Standards.

Access Dataset
Contact Owner | Share

Dataset Metrics
771 647 Downloads

Description In this dataset we include data to allow the user to calculate polaron properties, such as zero-point renormalization energy, for a set of materials. There are scripts included that use the standard and the generalized Fröhlich models for the calculation of these properties. (2023-03-17)

Subject Chemistry; Physics

Related Publication Pedro Miguel M. C. de Melo, Joao C. de Abreu, Bogdan Guster, Matteo Giantomassi, Zeila Zanolli, Xavier Gonze, Matthieu J. Verstraete, High-throughput analysis of Fröhlich-type polaron models, arXiv:2207.00364 (2022) arXiv: arXiv:2207.00364

Notes The files in the directory "Repository/eff_masses" were downloaded from the Materials Project under the CC BY 4.0 Licence (DOI: 10.1038/sdata.2017.85 and DOI:10.1002/cpe.3698) The values inside the files in the directory "Repository/phonon" were obtained from Guido, P. et al database (DOI: 10.6084/m9.figshare.c.3938023.v1)

License/Data Use Agreement CC BY 4.0

Files | Metadata | Terms | Versions

Dataset Version	Summary	Contributors	Published on
2.0	**Files (Added: 1; Replaced: 31; Changed File Metadata: 7);** View Details	Joao Abreu, Judith Biernaux	2024-09-16
1.0	This is the first published version.	Joao Abreu, Judith Biernaux	2023-03-31

Open Data and FAIR Data

Citations

ULiège Open Data Repository > LASLA Collection >

LASLAfiles_Latin_BPNFormat_SharedwithDTA_2019

Version 1.0

Longree, Dominique; Fantoli, Margherita, 2023, "LASLAfiles_Latin_BPNFormat_SharedwithDTA_2019", https://doi.org/10.58119/ULG/49UQNU, ULiège Open Data Repository, V1

Cite Dataset

Learn about Data Citation Standards.

Access Dataset

Contact Owner | Share

Dataset Metrics

722 Downloads

Description

The folder 'LASLA_files_shared_withDTA' contains the files that were shared in BPN format in 2019 with several partners under a Data Transfer Agreement. The documentation that was included at the time is also attached, together with the model of the Agreements which were signed. This document is included to document the previous agreements, but does not apply to the present data set, which is shared under the licence CC BY-NC-SA 4.0. (2023-09-11)
As a consequence of the DTA with the LiLa ERC-team, the LASLA corpus has been linked to the LiLa knowledge base, which can be queried via the LiLa interactive search platform: (https://lila-erc.eu/LiLaLisp/). Furthermore the BPN version of LASLA files has been converted to the CoNLL-U format and enriched with the links to the LiLa Knowledge Base by the LiLa team: the files are available on Zenodo (https://doi.org/10.5281/zenodo.5961377) and Github (https://github.com/CIRCSE/LASLA).
The full list of partners with which the files were shared is: The Alpheios Project, Ltd. (https://alpheios.net/); École Nationale des Chartes, Paris, Deucalion project (https://github.com/chartes/deucalion-model-lasla); Universiteit Antwerpen (see https://github.com/emanjavacas/pie and https://github.com/hipster-philology/nlp-pie-taggers); Università degli studi di Bergamo, Università Cattolica del Sacro Cuore, Milano, Lila project (https://lila-erc.eu); UNIVERSIDAD AUTÓNOMA DE MADRID, Regla project (http://www.reglabd.org/); University of Exeter; Haverford College, project The Bridge (https://bridge.haverford.edu/).

Subject

Arts and Humanities

The creation of a DOI or other identifier for your published dataset makes it possible for it to be considered a citation

RDM is not a waste of time…!

Promote your RDM skills

They are valuable assets for employers (academics or not),
but also to describe your research environment (research proposal)

Some examples:

- Knowledge in research process (data collection, methods)
- Knowledge in data curation, coding, IT skills
- Disciplinary specificities (tools, devices, programs, etc.)
- Knowledge in ethical and/or commercial use of data in your field
- Knowledge in the repositories, websites, where you can find/share data in your field
- Knowledge in data license

RDM is not a waste of time…!

- Helps planning your research and saves your time and energy in the long run
- Increases the added value of all the work you put into creating your dataset (your book will not gather dust on a shelf, your data lives longer than you)
- Increase the general ethics, transparency, traceability of science…
- …and therefore, its trust from society and governing bodies
- Makes research more accessible (think of everyone who can access your data: southern universities, non-profits, …)…
- ….but lets you keep your sharing in control (legal, ethical)
- Helps YOURSELF to reuse your previously acquired data
- Have peace of mind (less risk of loss with proper storage and documentation)
- Merge datasets and start new research projects – increases collaboration
- Sometimes it is mandatory

Discussion

Anything you want to apply to your research activity right after this training session ?

Thank you for your attention!

We are happy to answer your questions and to stay in touch

The content of this presentation has been created by the authors with the help of Adeline Grard, Jérôme Eeckhout, Pierre-François Pirlet and Catherine Thiry, to whom the authors are extremely thankful.

Please feel free to contact the authors for any question relating to the issues discussed in this presentation or for further help and information.

Data storage and documentation

Data documentation

Data entry is a fraction of the work… your data needs to be sufficiently documented to be standalone

Data documentation contains everything to make someone else understand yout data = metadata

Author, **date**, **location**, **format**, **size**, **keywords**, **description**, **abstract**…

They can be found in a combination of forms:

- **Readme files**, field notes, respondent information…
- **Codebooks** (esp. in excel, these are used to explained abbreviations, variables and other conventions)
- **Data collection support** (questionnaires, guides, recording material, …)

Data storage and documentation

Data organisation

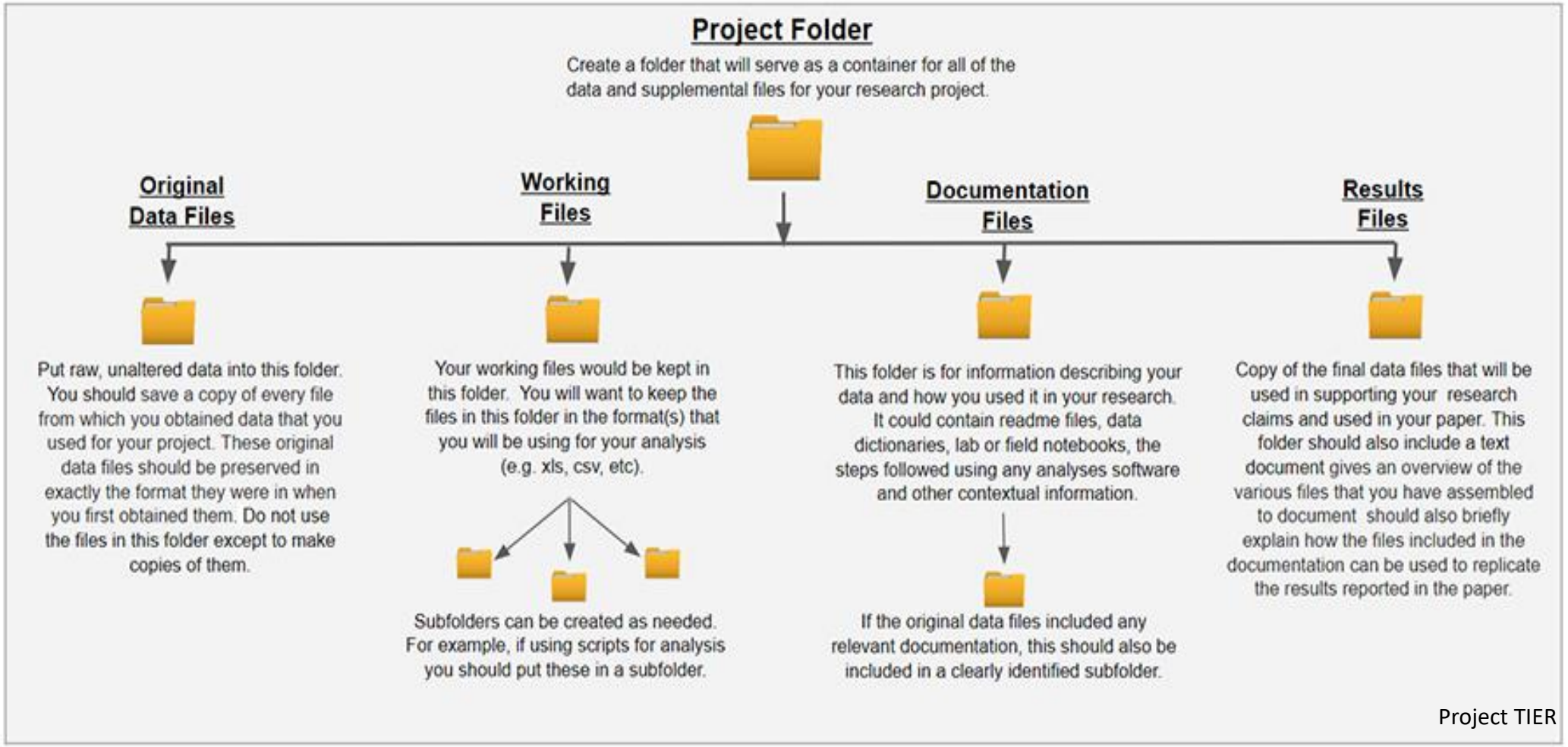
Once your data is entered, transcribed, and documented, it needs to be organised, stored and documented

Here are a few tips, but any recipe can work if it aims at re-useability

Imagine the data needs to be re-used by an incoming PhD student in ten years without them having to call you

Data storage and documentation

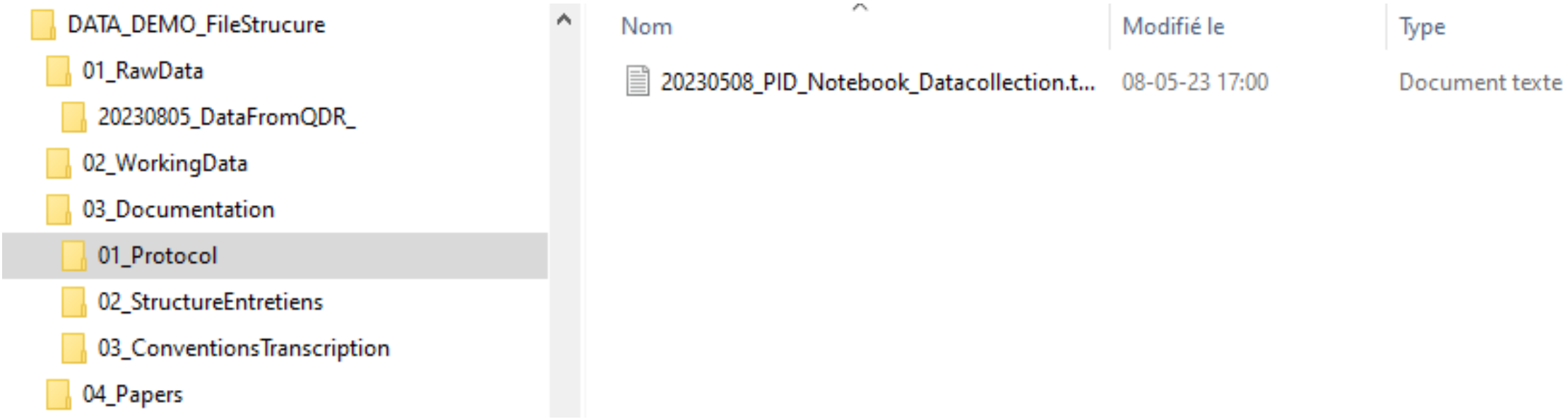
Data organisation: files



Data storage and documentation

Data organisation: files

Use a hierarchic file structure



Data storage and documentation

Data organisation: files

All your data have to be stored -> rule of thumb: never erase anything

Raw materials (interviews, focus groups, pictures, stories)
Information about your cases (demographics of participant, ...)
Field notes, contextualisation notes (atmosphere, location description,...everything that helps understanding the data collection)

Tips:

Make a conceptual map between your materials (“mind map”)
Create Personal IDs for each case (per project, lab member,

Data storage and documentation

Data organisation: files

Use naming conventions

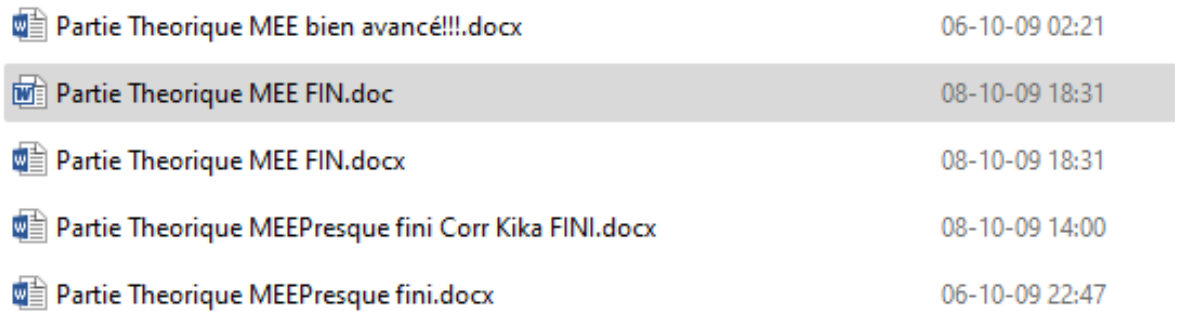
Useful file names are consistent, meaningful to your team, and allows you to find the file easily

- Ideally, have a **list of conventions** available in the lab/institute/project team and have an agreed upon format
  Example: DATE_Description_Version_Author.format
- **Dates** – agree on a logical use of dates so that they display chronologically (YYYY_MM_DD)
- Use **versioning suffixes** (e.g. _V01_AG)

change version number for big changes

- **Agree on who** is responsible

for using "final" _final



Partie Theorique MEE bien avancé!!!.docx	06-10-09 02:21
Partie Theorique MEE FIN.doc	08-10-09 18:31
Partie Theorique MEE FIN.docx	08-10-09 18:31
Partie Theorique MEEPresque fini Corr Kika FINI.docx	08-10-09 14:00
Partie Theorique MEEPresque fini.docx	06-10-09 22:47

Data storage and documentation

Data

	Risque de panne, perte ou dommage physique du support	Risque d'absence de backups	Risque de volume inadapté	Structure inadaptée, pas de métadonnées ni de documentation	Solution externe, hors infrastructure de l'ULiège (confidentialité ...)	Ne garantit pas d'utiliser des droits d'accès (identifiants)
Clés USB, disques durs externes, ...	X	X		X	X	X
Serveurs appartenant à l'UR	X	X				
Portails en ligne institutionnels (DoX, EDC, ...)			X	X		
Clouds externes (Google Drive, Dropbox, ...)			X	X	X	

Data storage and documentation

Data organisation: storage support

Backup and security best practices (to the best of your ability)

- **Apply the 3-2-1 rule**: 3 copies, 2 different support, 1 off-site (example: 1 on my laptop for work, 1 on the university storage solution, 1 on the university cloud)

- **Avoid** portable storage solutions as only support (that means USB drives, portable hard-drives, **laptops**, smartphones…)

  Use **centralised** solution (uni-managed or lab-managed storage spaces, institutional servers, uni-managed clouds, … do you have them?)

- If human data, **absolutely no outside clouds** like DropBox, GoogleDrive, or even gmail
- **Password** protection is always a good idea – talk to your IT service

Data storage and documentation

Data organisation: storage support

Backup and security best practices (to the best of your ability)

- Store **all your data and its story**: documentation, annexes, field notes, codebooks …
- If backup space is limited, have a **strategy**
  For example: backup only the raw data + the documentation to process it + its latest version (the raw data is usually very important to backup)
- If you want to transfer data to a coworker, use Belnet FileSender (ok GDPR)

Data storage and documentation

Data organisation: storage support

Data preservation and longevity

There is no universal data conservation rule, but there is a destruction rule in the GDPR recommendations. 20 years is a good recommandation.

If you choose to publish data, things get easier ☺

Plan ahead and think long-term (that is why we avoid portable storage and favour uni-managed solutions)

Ask your IT service (most have magnetic tape backup service, cold storage vs hot storage)

Data storage and documentation

Preservation

Besides publication, preserving (parts of) a dataset can be of added value to you and your peers

- It may act as a back up, esp. if your dataset is published extra muros
- It may enable further verification (reproducibility)
- It may end up contributing to a larger archive
- It may be useful for teaching and learning purposes
- It is especially important for data that is difficult to replicate (specific conditions, period, …)

As a **rule of thumb**, unless it is explicitly meant to be destroyed, datasets that underpin an article should be archived for verification purposes for around 15 years
But maybe **not everything needs to be kept** (maybe only the raw data, maybe the final version, …
-> data curation process and archival strategy as part of the DMP

https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep#4

Focus on GDPR

Personal and sensitive data : Steps towards privacy compliance

For more information about personal data, see the "Data Privacy Handbook" written by Utrecht University.

➔ Research Data Management Support et al., *Data Privacy Handbook,* Utrecht University, 2023. https://doi.org/10.5281/zenodo.8005847.

Research Data Management Support, *10 steps towards privacy compliance in research* (v2023.12.21), Utrecht University, 2023. https://doi.org/10.5281/zenodo.10417514

Focus on GDPR

When dealing with **personal data**, there are **three main obligations :**

INFORMATION
Subjects need to know what their data will be used for, so as to be able to give **informed consent**

REGISTRATION
Each personal data processing need to be recorded in a register, with its justification and legal basis ("why do you need this data and what have you done with it?")

PROTECTION
A risk analysis or impact assessment needs to be conducted, and appropriate protection actions need to be taken (storage, access, preservation, …)

GDPR is a **legal obligation**, but it derives from general **ethical principles** to do no significant harm and to be mindful of human rights.

Data management is **not only about compliance**: it is about ethics, for the subject, for you, for the community.

Data collection

Legal framework: GDPR

General Data Protection Regulations (GDPR - 2016) is a set of regulations protecting the **privacy** of humans. It addresses the collection, processing, storage, transfer of personal data inside and outside EU.

Personal data = **any information relating to an identified or identifiable natural person**, directly or indirectly, in particular by reference to an identifier, such as a name, an identification number, location data, an online identifier etc

Sensitive data = data that in other contexts could contain risks for the subject (political opinions, health and sex life, genetics, ethnicity criminal records, biometric data, religion, children data…

Data collection

Legal framework: GDPR

General Data Protection Regulations (GDPR - 2016) is a set of regulations protecting the **privacy** of humans. It addresses the collection, processing, storage, transfer of personal data inside and outside EU.

When applied to research, it results in a few obligations:

- **Informing** subjects of the use of their data and leaving them the right to withdraw (with some limitations) -> that means consent forms
- **Protecting** their data from transferring, leaking, publishing... -> that means choosing proper storage solutions
- Building their **study protocol** and data collection on a **legal basis** ("why do you need personal data") -> this means **thinking ahead** and anonymising

Data collection

Legal framework: GDPR

We only collect data if we can justify its:

Lawfulness – legitimate basis must be clarified. For research these are most often ‘legitimate interest’, ‘public interest’ along with ‘consent’.

Fairness – towards the data subject.

Transparency – data subjects should be aware of the processing of their personal data.

Purpose limitation – purpose must be specified, explicit and legitimate. Personal data collected for one purpose should not be used for another purpose unless it is compatible with original purpose.

Data collection

Legal framework: GDPR

Before you start collecting, make sure to apply these principles:

Data minimization and proportionality – only collect the data you need.

Accuracy – keep records up to date.

Storage limitation – assess the purpose and reasoning for storing the data for lengthy periods of time. Never store data outside of your control (no non-uni cloud!)

Integrity and confidentiality – protect data from damage and unlawful processing. Information security, encryption, pseudonymisation.

Accountability – demonstrate responsibility and compliance through documentation

Data collection

Legal framework: GDPR

Anonymisation is not…

- Blurring data: adding noise, re-sampling, blurring images and beeping words
- Pseudonymizing (still identifying, especially since you must keep a pseudonyms key)
- Erasing some parts of the dataset

But it is a **combination** of these techniques and more…

See Data Amb training video!

No panic: your **DPO** is there to help you (consent forms, protocol review, storage review, …)

Rule of thumb: collect what you need, store in a protected way, call your DPO and don't share openly

Data collection

Legal framework: beyond GDPR

GDPR is a **legal obligation**, but it derives from general **ethical principles** to do no significant harm and to be mindful of human rights.

Data management is **not only about compliance**: it is about ethics, for the subject, for you, for the community.

In general, think of your proposed protocol in terms of these principles:

Data collection

Legal framework: beyond GDPR

Beneficence (do good): research should be conducted for the benefit of individuals taking part in your research, and for broader society and the natural environment.

Non-maleficence (cause no harm): research should not increase discrimination or expose people to risk - therefore their identity should be protected as far as possible.

Accountability: An accountable person must be assigned for each research project. This person is answerable to research participants and others, regarding the research conduct. Accountable researchers establish processes and documentation to ensure privacy and confidentiality for research participants.

Transparency: Research should be conducted in a transparent manner. Research participants should be aware of their participation and how their data is used within it.