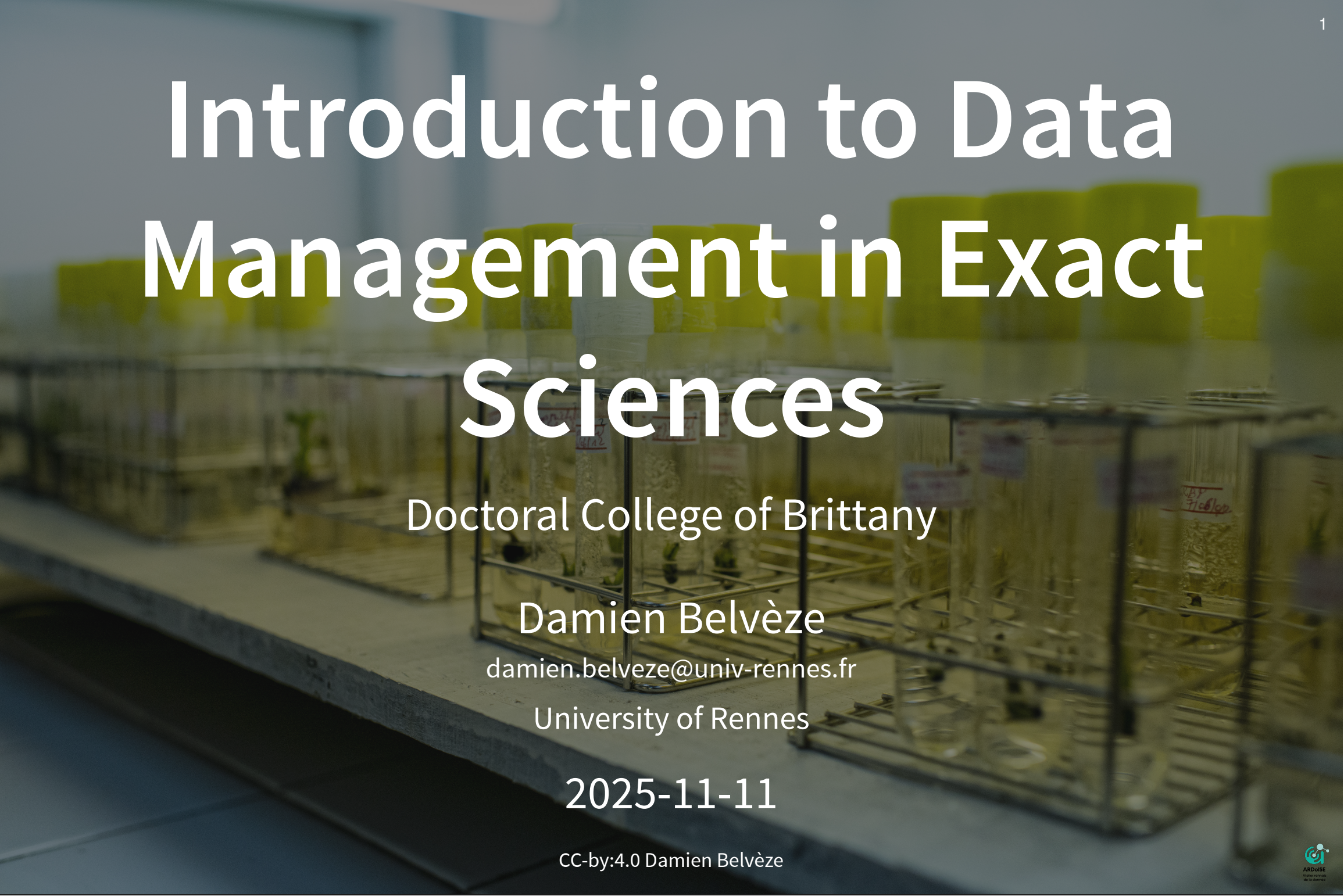


Introduction to Data Management in Exact Sciences

Doctoral College of Brittany

Damien Belvèze

damien.belveze@univ-rennes.fr

University of Rennes

2025-11-11

- Advice on data management
- Training (data, reproducibility, identifiers)
- Support for data management plans
- Curation of the Univ-Rennes collection on Recherche Data Gouv
- help for making your research source code reproducible

ARDoISE

Atelier rennais de la donnée

ARDoISE data hub

Speaker notes

Explain what a national data workshop is ARDOISE Data Workshop = University of Rennes 2 and University of Rennes

1. Research Data, What Are We Talking About?

Figure 1: data life cycle

Speaker notes

Plan and design

Data Management Plan is the first step, before the collecting process has begun. A data management helps to clarify to the team and makes it clear to reviewers who are responsible for the data collection porcess and the following steps : storage, sharing, archival and final destruction importance of file naming conventions (we will talk about it later)

Collect and create

source code design to process data should be also made available from a dedicated repository (forge and Software Heritage Fondation) for code source archiving Importance of Documentation to make datasets understandable and reusable by othoer researchers

analyse and collaborate

safe storage, backup routines, access rules

a typical experiment : counting the fish in a river

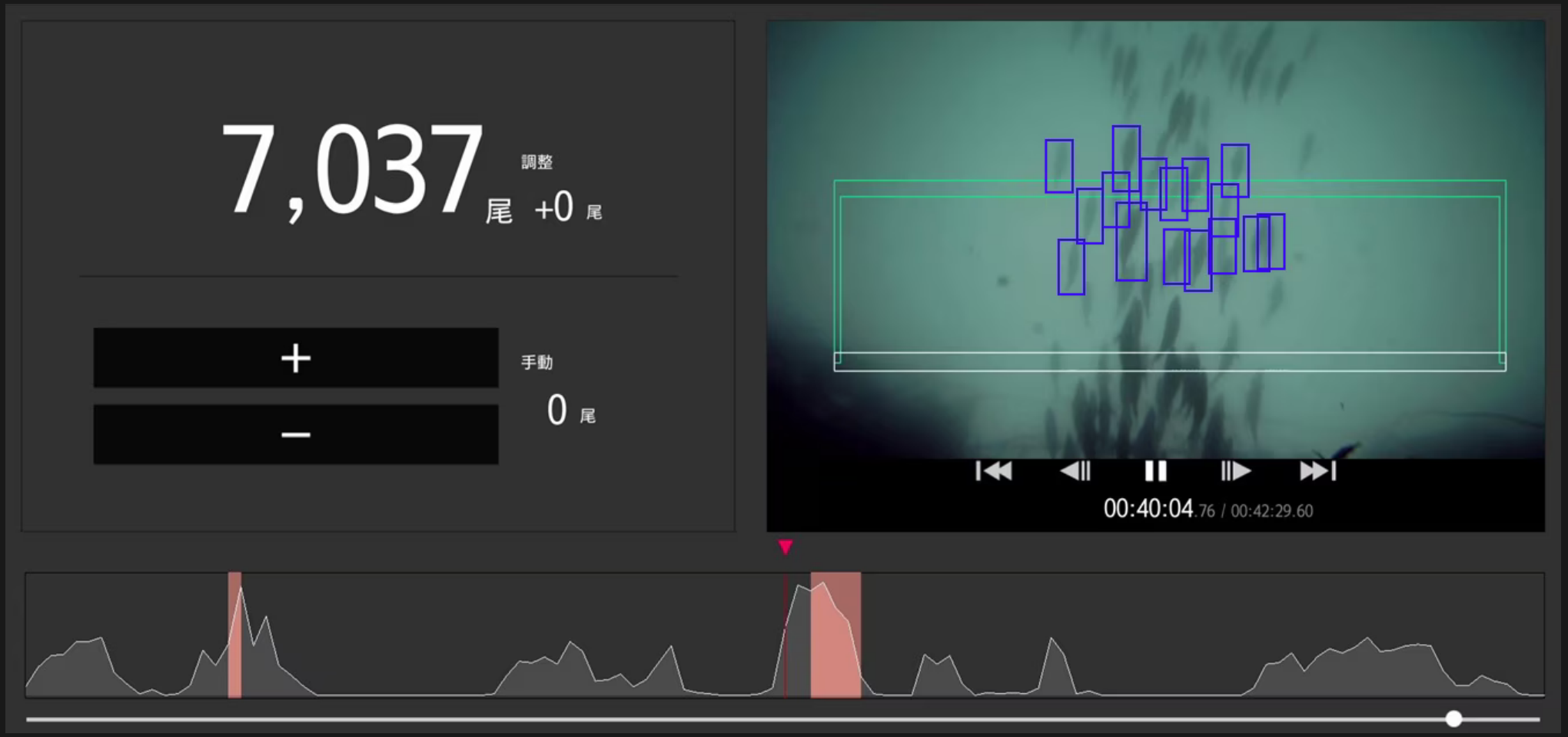


Figure 2

Speaker notes

Let’s think of a research team who tries to count the population of a given species in a river with the help of a underwater camera including an Artificial Intelligence powered recognition system. What information is important to share in order to keep the experiment reproducible after the publications are made ?

Which Files Are Important to Make Available?

- raw_data_fish_counter.csv
- intermediate_data.xls
- filter1.py
- first_draft_submission.pdf
- fish_counter_calibration.md
- kick_off_report.docx
- filter2.py
- notebook_experiment.ipynb
- final_data_fish_counter.xls
- project_presentation_funders.pp
- final_data.csv
- study_draft.qmd
- january_meeting_partners.docx
- fish_counter_instructions_for_u
- gantt_calendar.xlsx

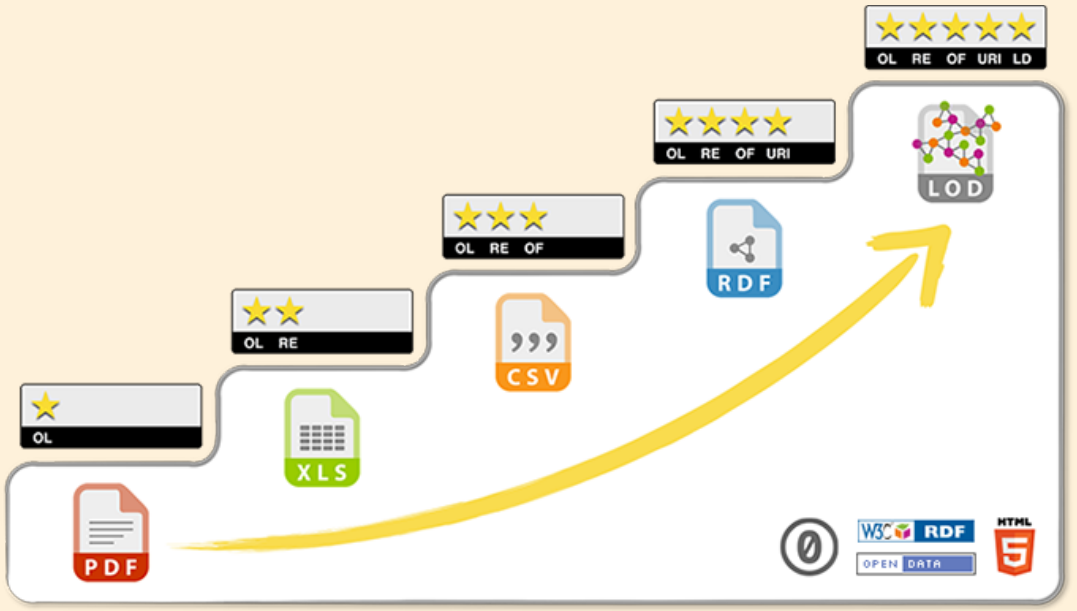
Speaker notes

Research data are data that have demonstrative value and on which publications rely. *Inscriptions* (Bruno Latour) produced in the life of the project or lab (meeting minutes, equipment manuals, project agendas) are not research data. Their preservation may be of interest in some cases, but not within the framework of a Data Management Plan (DMP).

Answers

- raw_data_fish_counter.csv
- intermediate_data.xls
- filter1.py
- first_draft_submission.pdf
- fish_counter_calibration.md
- kick_off_report.docx
- filter2.py
- notebook_experiment.ipynb
- final_data_fish_counter.xls
- project_presentation_funders.pp
- final_data.csv
- study_draft.qmd
- january_meeting_partners.docx
- fish_counter_instructions_for_u
- gantt_calendar.xlsx

2. Towards Cumulative, Reliable, and Reproducible Science?



(berners-lee5starOpenData2015?)

Speaker notes

Remind doctoral students of the nonsense of “freezing” their data in the PDF support of their thesis. When depositing the publication in HAL, the librarians of the University of Rennes try to retrieve from the authors the data in reusable format, when they are only found in the PDF of the article.

Open Science relies on free formats (transparency, format accessible to all). A CSV document is better than an Excel file (see further)

The next step is to use URIs (for example, Wikidata elements) in your data tables to link this data to other similar data (RDF)

Finally, data should be linked together (LOD stands for Linked Open Data) with permanent identifiers :

Lowest temperature in Galway on the 24th of january may be expressed in that matter :

From buried-in-a-PDF data...

Figure 3

- Even Optical AI tools are bad at extracting data from PDFs 📓 Edwards (2025)
- Mistral failed at converting tables in PDF into markdown (figures truncated)
- PDF should be one of the layer not the only one to be communicated
- share native format and use reproducible tools such as R

Speaker notes

“Imagine a doctor who uses AI to consult a PDF table of drug protocols and prescribes 500mg of a potentially toxic chemotherapy agent instead of 50mg.”

…through Excel (but why Excel would be necessary anyway)

- better to share flat files (not workbooks)
- have you considered collecting your data in CSL rather than Excel ?

> *Unlike a script written in a programming language such as Python or R, which documents every step of the process and can be saved, versioned and rerun, an analysis that happens inside a spreadsheet using pointing and clicking is hard to follow and even harder to replicate.*

📓 Melchor (2025)

Speaker notes

why flat files are better than workbooks ? in a classic excel spreadsheet, not enough place to properly name each spreadsheet(for a NFTS partition, you ca use up to 255 characters, which would not be reasonable anyway) Most of the time, formulas link several spreadsheets together. Some cells in a spreadsheet are sums or means calculated from figures in another spreadsheet. These formulas are often difficult to manage and are written in the specific language of the software, making it challenging to reproduce them in an analysis tool. In an open science framework, spreadsheets should contain only raw data. Value formats, and calculations should be handled in a separate file or script.

Excel (as every non FOSS tool) itself is prone to errors

Figure 4: CSV vs XLS

📓 Ziemann et al. (2023)

Speaker notes

Using Excel to analyze data (as done by 69% of researchers according to a 2016 study) is a bad idea. Instead, use transparent tools like R. A 2021 study shows that Excel regularly changes genes into dates, and the phenomenon has long been known. The reproducibility of studies depends on the use of free and transparent tools on what they do. References here: https://doi.org/10.1093/bib/bbad375

besides, Excel cannot manage big dataframes with more than 24 000 records ; in that case, you have no other option than to use a CSV along with a computational tool such as R or Python.

Preserve the raw data

- The raw and final data are important to preserve and later deposit in a repository
- if your data management is well documented, you don’t need to keep intermediary versions
- keep the file that contains your raw data untouched (lock it down in a separate folder), use a duplicate as your working version

…to Linked Open Data

“Lowest temperature in Galway on the fourteenth of January 2020”

concept	ontology	expression
lowest temperature	http://inamidst.com/sw/ont/meteo	Property :temperature
temperature type	Wikipedia	https://fr.wikipedia.org/wiki/Degr%C3%A9_Celsius
Galway	DBpedia	https://dbpedia.org/page/Galway
14th of January	Unix epoch	1578984191

Speaker notes

RDF is a common structure for data which presents date in triples : The subject denotes the resource; the predicate denotes traits or aspects of the resource, and expresses a relationship between the subject and the object

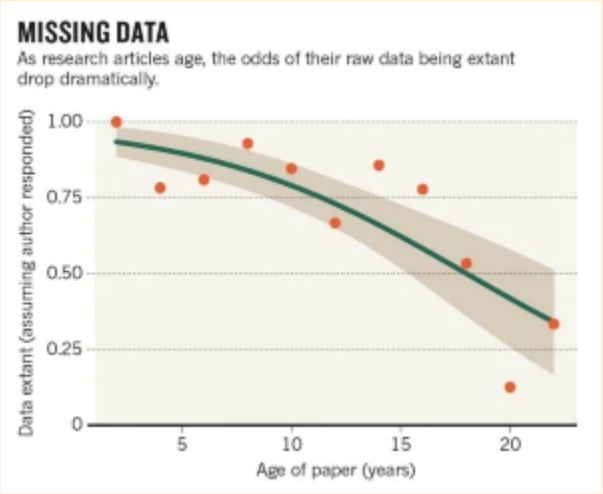
Example of RDF statement : 14th of January (subject) is written in Unix epoch (predicate) 1578984191 (object). RDF style is a necessary steps to format one’s data in LOD format.

Linking data with PID is very important to avoid ambiguities (outside Connemarra, there are a lot of places in the world that are called “Galway”)

Speaker notes

Play the video synchronously with a class - Download the video from the media server and save it to the desktop - Open it without playing it - Share the screen from Zoom, choose the video player - Check both boxes at the bottom of the sharing window (audio for everyone + optimization) - Play the video and stop sharing when it’s finished

Permanence of Data Access



Gibney & Van Noorden (2013)

Speaker notes

as time is passing by, data are more and more difficult to find if they are not properly preserved in an archive or at least a trusted data repository

3. A Challenge for Open Science

FAIR Principles

Figure 5: FAIR principles

openness / closure

- “as open as possible, as closed as necessary”
- Default openness
- Closure to justify:
  - personal data
  - intellectual property
  - industrial secret
  - defense

Speaker notes

personal data

The challenge remains to provide access to the dataset to the fullest extent possible. To do this, processes will be used to pseudonymize or anonymize the dataset. Pseudonymization involves removing directly identifying data, or replacing them with pseudonyms. This method carries risks of re-identification through cross-referencing the dataset with other datasets concerning the same population. Anonymization of data is a more subtle method involving, after removing directly identifying data, making the data less precise or adding noise to hinder potential re-identification attempts. It is a compromise between the interests of the individuals whose data are collected and the precision of this data. Tools exist to anonymize data: Arx, Amnesia. Anonymizing data incurs processing costs (human time, performance) that should be anticipated in advance in the budget request.

intellectual property

In France, the regime of data law means that data by default belongs to employers, not producers (researchers). But there may be limitations on sharing this data: - contaminating licenses of certain proprietary materials used to produce data (e.g., medical imaging) - international research projects: which national law will be applied to this data?

Making Your Data Findable

Quality of a directory:

- reputation
- sustainability (institutional support)
- open license
- persistent identifier
- richness of metadata
- curation

Speaker notes

Zenodo: no versioning of datasets, metadata less rich than on RDG, but sustainable repository (European Commission), managed by a transnational authority (see what happen to US publication repositories!)

to search for repositories and compare their offerings: https://re3data.org

Avoid proprietary repositories (Figshare)

repository compliance criteria: https://www.ouvrirlascience.fr/entrepots-de-donnees-de-confiance-criteres-de-conformite/

discipline	repository
images (SHS)	MediHal
code	Software Heritage via HAL
Bioinformatics	GenOuest
Humanities	Nakala
Mathematics	no repository, see with the RNBM group
environment, hydrology	Data Indores
Earth Sciences	data terra
Marine Sciences data	ifremer, seanoe
medical sciences	INSERM repository on RDG

discipline	repository
Ecology, Environment, and Society Data	InDoRES and Cat.InDoRES

Recherche Data Gouv

- richness of metadata
- curation
- national reference (supported by the Ministry)
- persistent identifier
- significant volume
- free of charge
- simplified generation of datapapers
- RDG sandbox

Speaker notes

On Zenodo, you are limited to 50 gigabytes per dataset on RDG, you are limited to 50 gigabytes per file, a dataset can contain several files (no global limitation seen), so yes, apparently the volume is less limited on RDG than on Zenodo.

Are Data Accessible?

In 93% of cases no response or negative response without justification 📓 Gabelica et al. (2022)

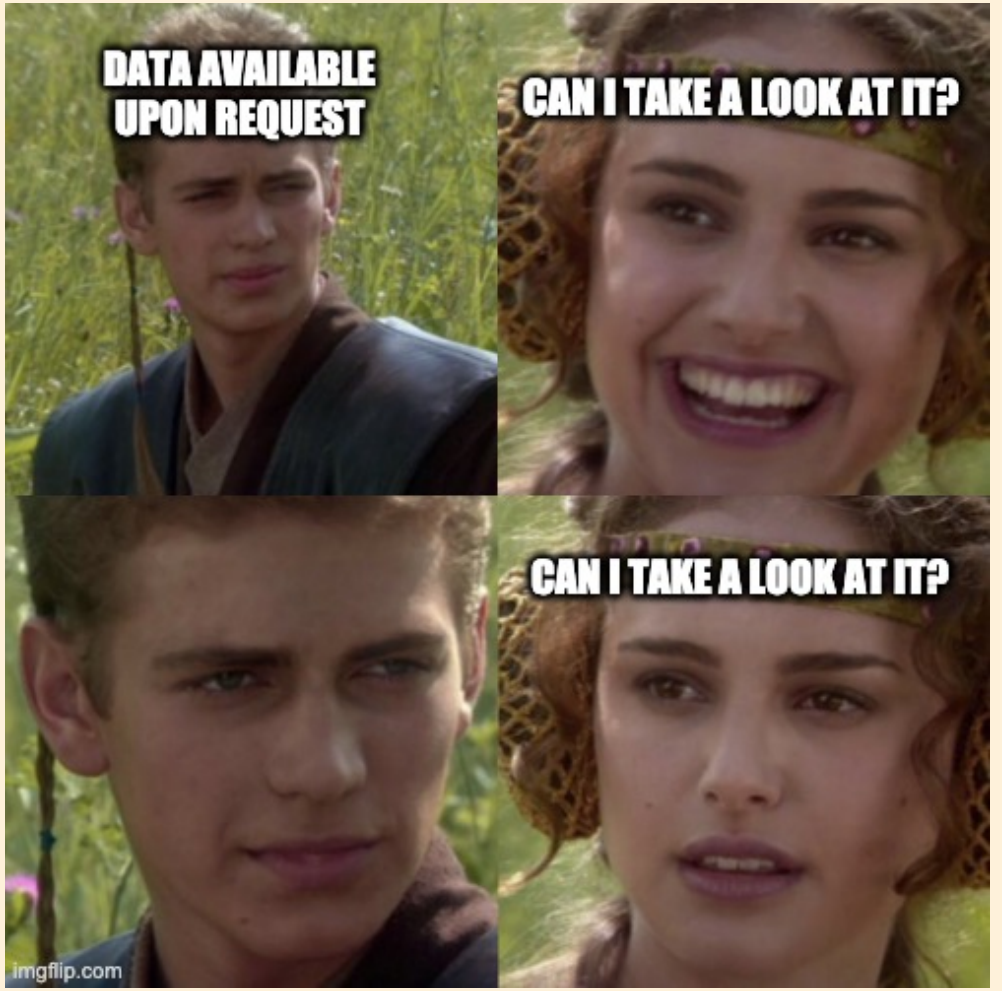


Figure 6: “data available upon request”

Speaker notes

The person listed in the DMP as the data contact person may be the corresponding author of the related publication or another person, but this person must be aware of their role if their mediation is necessary to obtain the data. It is better to provide a personal address here (e.g., gmail) than an institutional address, as one may frequently change institutions.

Are Data Interoperable?

Which identifiers to use for copper telluride?

registry	identifier
CAS number	12019-52-2
PubChem CID number	6914517
PubChem SID number	24879035
openSMILES identifier	CuCu.CuCu.TeTe
InChI identifier	InChI=1/2Cu.Te
MDL number	MFCD00049727

Speaker notes

In repositories, we would rather use **InCHI** than **CAS** because to access CAS identifiers, one needs to have access to the CAS Sci-Finder database.

PubchemID is considered a better choice than InCHi, as it is more machine-readable

Documenting the Data

- Documentation is the glue that binds a data science project together 📓 Ziemann et al. (2023)
- Carefully describe the data and the context of its acquisition (production, collection)
- literate programming
- describe the data using ontologies

Documenting to Avoid Context Errors

Be precise in describing the context of data production

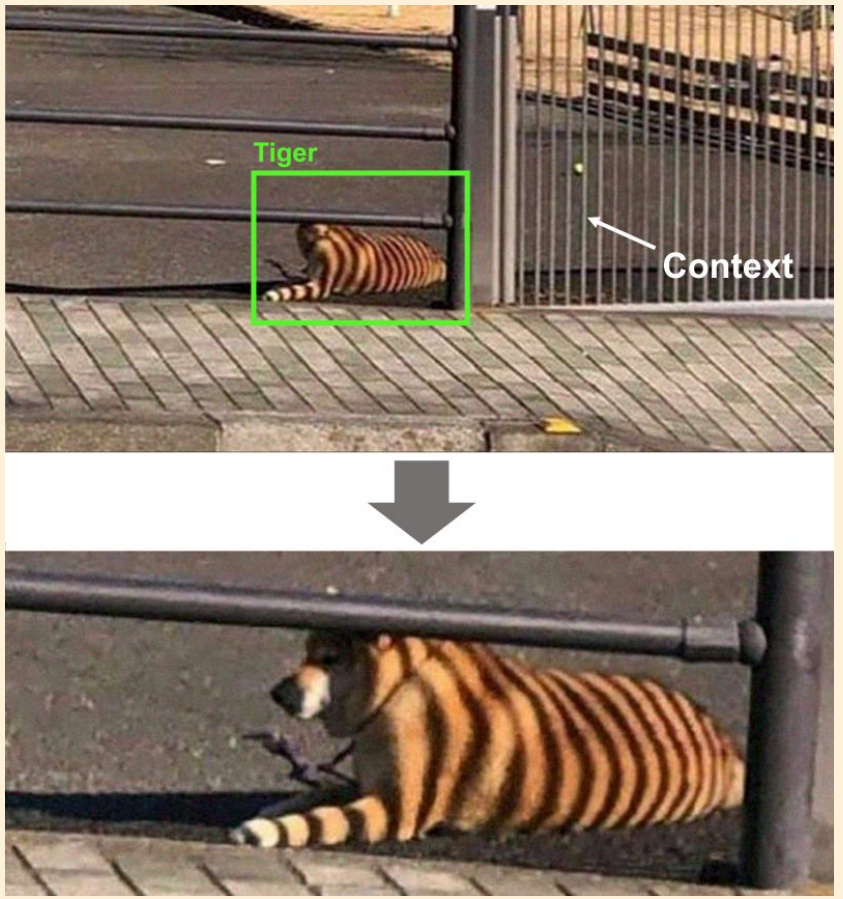


Figure 7: the importance of data context

Speaker notes

possibility to link the data management plan to a protocol deposited on another site (e.g., https://www.protocols.io) or [Prospero] (https://www.crd.york.ac.uk/prospero/) if the data are bibliographic and it is a literature review

Ontologies

discipline	thesaurus
biodiversity	INRAE
environment	GEMET
Biology, Health	MeSH
Mental Health	ascodopsy

directory of thesauri

Reusable Data?

- Creative Commons (CC:by)
- a license written by a law firm expert in intellectual property that provides for a variety of authorized or prohibited use cases
- ODBL
- Etalab
- no license, do whatever you want with my dataset
- CC0
- CC:by for everyone except for fossil industries, arms sellers, and Google (📓 Thomas (2023)) text available here.

Speaker notes

Ask the students which data mentioned in this slide are appropriate for a dataset.

Answers:

Etalab and ODBL licenses are recommended in France to allow data reuse. Creative Commons licenses are more suitable for research outputs (publications, preprints, posters…) CC0 license is not allowed in France where copyright law requires at least citing the author of the dataset, which is not provided for in this license.

When no license is indicated, one cannot safely reuse the dataset. When there is no license, by default, copyright law applies.

In the context of Open Science, licenses are standardized and the uses they allow are easily stated and understandable. If instead of a license, there is a text that resembles the T&Cs of a web service, it means that the dataset is likely to be difficult to reuse (probably need to contact the producers)

Open Science can be seen by some researchers as a loss of control over their publications and data, and restrictive licenses that hinder reuse for purposes deemed bad by researchers are still in the project stage. These licenses are not usual for data.

Data Management Plan

- The DMP summarizes all the choices made for data management
- Submit an initial version of a DMP 6 months after signing a contract (ANR, European projects)
- DMP OPIDOR
- DMP Online

4. Let’s Get Practical

Figures

figure	credits
Figure 2	yanmar https://www.yanmar.com/global/news/2021/04/20/90815.html
Figure 1	Harvard Biomedical Data managemnet
?@fig-data_steps	Tim Berners-Lee
Figure 3	Jon Ippolito
?@fig-data_loss	Gibney, Van Noorden
Figure 5	Willkinson, Dumontier et al.
Figure 6	Sergio Uribe
Figure 4	meme dont l'origine se perd dans la nuit des temps
Figure 7	Ralph Aboujaoude Diez

Software Used for the Presentation

The presentation was created with free and Open Source software. Thank you to all people who make them alive ❤️ ❤️

```
Slides


[1] "Quarto version: 1.6.40"
```

```
R version 4.5.2 (2025-10-31)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 24.04.3 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.12.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0  LAPACK version
3.12.0

locale:
 [1] LC_CTYPE=fr_FR.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=fr_FR.UTF-8        LC_COLLATE=fr_FR.UTF-8
 [5] LC_MONETARY=fr_FR.UTF-8    LC_MESSAGES=fr_FR.UTF-8
 [7] LC_PAPER=fr_FR.UTF-8       LC_NAME=C
```

quiz : H5P

repository : Framagit owned and managed by the Fnch pro-FOSS association Framasoft

References

Edwards, B. (2025). Why extracting data from PDFs is still a nightmare for data experts. In *Ars Technica*. https://arstechnica.com/ai/2025/03/why-extracting-data-from-pdfs-is-still-a-nightmare-for-data-experts/

Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: Mixed-methods study. *Journal of Clinical Epidemiology*, *0*(0). https://doi.org/10.1016/j.jclinepi.2022.05.019

Gibney, E., & Van Noorden, R. (2013). Scientists losing data at a rapid rate. *Nature*. https://doi.org/10.1038/nature.2013.14416

Melchor, S. (2025). Six questions to ask before jumping into a spreadsheet. *Nature*, *644*(8076), 569–570. https://doi.org/10.1038/d41586-025-02511-z

Thomas, M., Éric Tannier. (2023, May 17). *Se réapproprier la production de connaissance - AOC media*. AOC media - Analyse Opinion Critique. https://aoc.media/opinion/2023/05/17/se-reapproprier-la-production-de-connaissance/

Ziemann, M., Poulain, P., & Bora, A. (2023). The five pillars of computational reproducibility: Bioinformatics and beyond. *Briefings in Bioinformatics*, *24*(6), bbad375. https://doi.org/10.1093/bib/bbad375