

Introduction to Data Management in Exact Sciences

Doctoral College of Brittany

Damien Belvèze

damien.belveze@univ-rennes.fr

University of Rennes

2025-07-04

- Advice on data management
- Training (data, reproducibility, identifiers)
- Support for data management plans
- Curation of the Univ-Rennes collection on Recherche Data Gouv



ARDoISE

Atelier rennais
de la donnée

ARDoISE data hub

1. Research Data, What Are We Talking About?

DATA



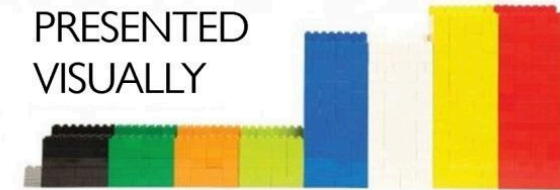
SORTED



ARRANGED



PRESENTED
VISUALLY



EXPLAINED
WITH A STORY



Figure 1: données brutes, données raffinées

Which Files Are Important to Make Available?

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.pdf
- 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?

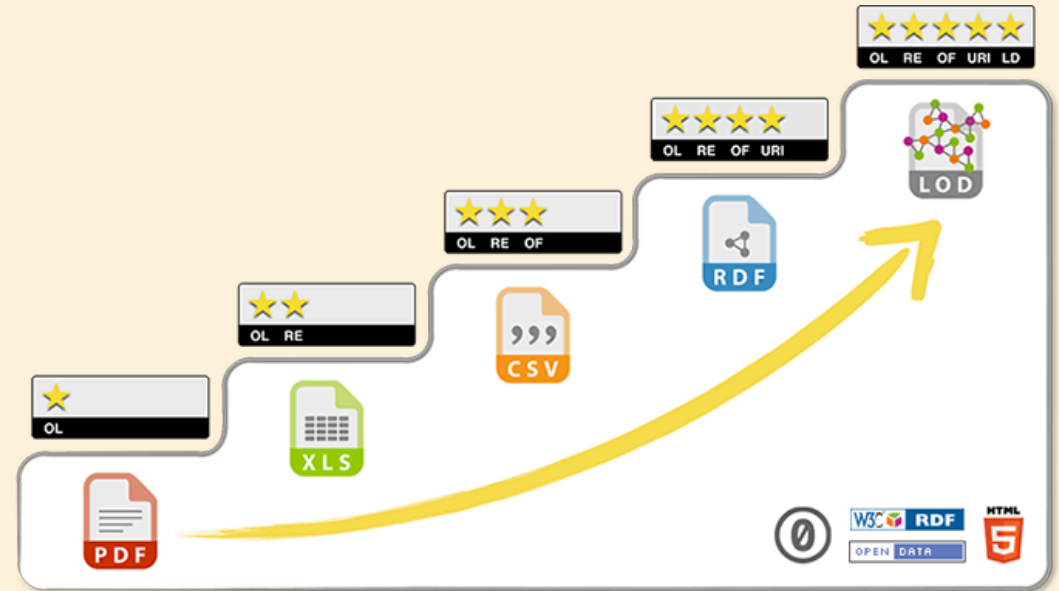
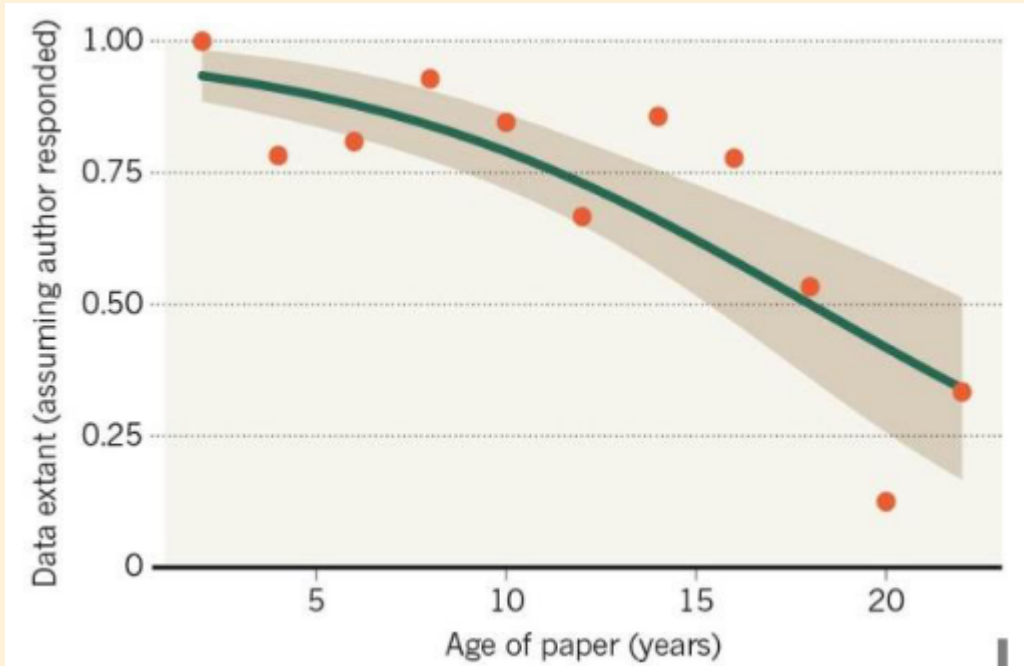


Figure 2: static data to linked data

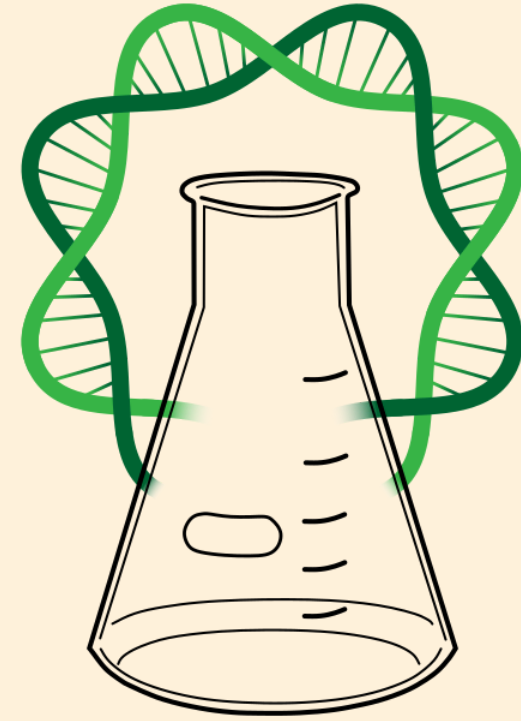
Permanence of Data Access



Gibney & Van Noorden

(2013)

3. A Challenge of Open Science



open science

FAIR Principles



Figure 3: principes FAIR

openness / closure

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

Making Your Data Findable

Quality of a directory:

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

| 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, seanoë |
| 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](#)

Are Data Accessible?

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



Figure 4: “data available upon request”

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 |


Transparent Formats?



Figure 5: CSV vs XLS

 Ziemann et al. ([2023](#))

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 6: the importance of data context

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.

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**

4. Let's Get Practical

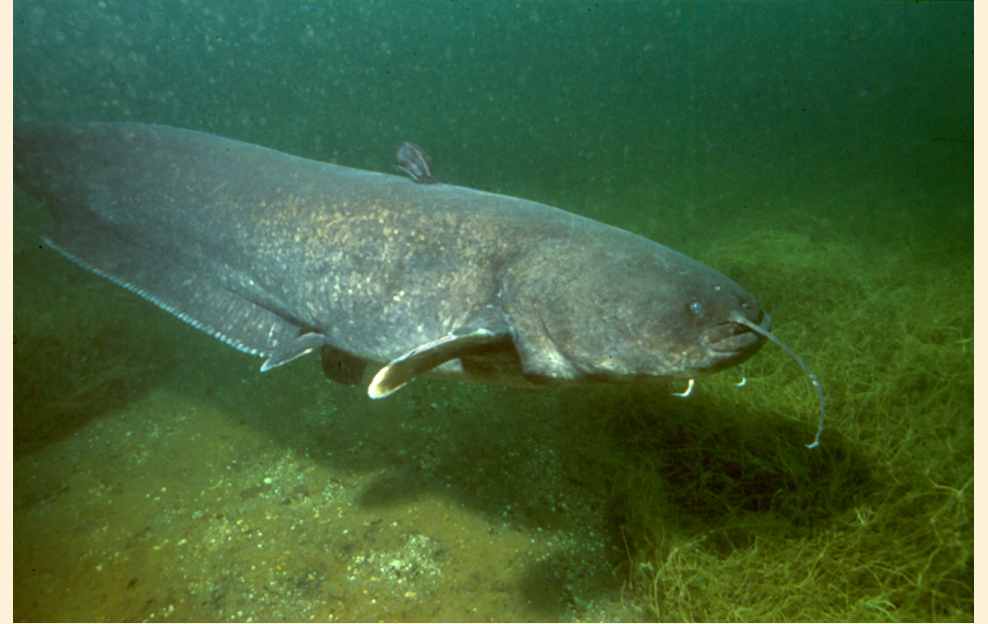


Figure 7: Silurus Glanis

fictional data (ChatGPT)

| | Day | January | February | March | April | May | June | July | August | September | October |
|----|-----|---------|----------|-------|-------|-----|------|------|--------|-----------|---------|
| 1 | 1 | 5 | 3 | 8 | 4 | 2 | 10 | 11 | 12 | 9 | 7 |
| 2 | 2 | 3 | 7 | 2 | 5 | 4 | 8 | 1 | 6 | 11 | 10 |
| 3 | 3 | 6 | 4 | 9 | 7 | 11 | 3 | 8 | 5# | 2 | 1 |
| 4 | 4 | 8 | 1 | 0 | 3 | 0 | 9 | 4# | 7 | 5 | 2 |
| 5 | 5 | 2 | 10 | 7 | 12 | 8 | 4 | 11 | 1 | 6 | 3 |
| 6 | 6 | 4 | 0 | 3 | 1 | 5 | 7 | 2 | 10 | 0 | 12 |
| 7 | 7 | 7 | 5 | 1 | 9 | 10 | 2 | 6 | 3 | 4 | 11 |
| 8 | 8 | 11 | 0* | 6 | 2 | 3 | 1 | 7 | 9 | 12 | ** |
| 9 | 9 | 1 | 9* | 4 | 11 | 7* | 5 | 10 | 2# | 3 | ** |
| 10 | 10 | 9 | 2* | 10 | 6 | 1* | 11 | 3# | 8 | 7 | ** |
| 11 | 11 | 7 | 6* | 5 | 1 | 9* | 2# | 4# | 11 | 8 | ** |
| 12 | 12 | 4 | 11* | 1 | 8 | 5* | 6 | 9 | 7 | 10 | ** |
| 13 | 13 | 12 | 7* | 2 | 4 | 11* | 3 | 5 | 6 | 9 | ** |
| 14 | 14 | 8 | 5* | 3 | 7 | 6* | 10 | 1# | 2# | 4 | ** |
| 15 | 15 | 0 | 0+ | 10 | 10 | 4 | 0 | 0 | 0 | 0 | ++ |

Dear Prof. Armand,

I'm enclosing the data collected this year by our various sensors installed on the Tydale as part of your study "Growth of glane silure catfish *silurus glanis* in european river, the case of the Tydale river".

Funds from the Royal Fisheries Corporation (RFC) enabled the purchase of 8 underwater sensors along the Tydale, which, when properly parameterized, were able to count only fish weighing over 10 kg. Thanks to the camera's artificial intelligence, catfish were counted with a margin of error of around 3%.

We have noted certain incidents in the data obtained that could affect the proper conduct of the study.

In February and May, we noted that some sensors were no longer working properly and had to be repaired. In October, the centralized results collection system broke down for 11 days, causing us to lose data.

In addition, we pointed out that boating activities on the Tyndale River on certain days in June and July could disturb the catfish, which were therefore

Figures

| figure | credits |
|-------------|--|
| Figure 1 | Maricx |
| Figure 2 | Tim Berners-Lee |
| ?@fig-perte | Gibney, Van Noorden |
| Figure 3 | Willkinson, Dumontier et al. |
| Figure 4 | Sergio Uribe |
| Figure 5 | meme dont l'origine se perd dans la nuit des temps |
| Figure 6 | Ralph Aboujaoude Diaz |
| Figure 7 | Dieter Florian |

Software Used for the Presentation

Except for its translation from french to english, which was made with the help of ChatGPT, the presentation was created with free software (thank you Richard M. Stallman), including the following software:

Quarto 1.3.450

VScode 1.8.0

R :

```
— Session info —  
setting  value  
version  R version 4.3.0 (2023-04-21 ucrt)  
os       Windows 11 x64 (build 22631)  
system   x86_64, mingw32  
ui       RTerm  
language (EN)
```

```
collate   French_France.utf8
ctype     French_France.utf8
tz        Europe/Paris
date      2024-04-18
pandoc    3.1.11 @ C:/PROGRA~1/Pandoc/ (via rmarkdown)
```

– Packages

```
package      * version date (UTC) lib source
```

the text editor is VScode. VScode makes using Quarto easier (Quarto is a software without a graphical interface, as is Pandoc, which is integrated into Quarto). VScode also allows encapsulation of R chunks (if R is loaded on the machine with the corresponding packages)

The session info package in R allows showing how data manipulation code can be automatically documented to produce a figure from a dataset, for example. Code reproducibility is an essential issue related to data access.

References

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