introduction à la gestion des données en sciences exactes

Collège Doctoral de Bretagne

Damien Belvèze

damien.belveze@univ-rennes.fr

Université de Rennes

2023-07-11

- conseils sur la gestion des données
- formations (données, reproductibilité, identifiants)
- accompagnement sur les plans de gestion de données
- curation de la collection Univ-Rennes sur Recherche Data Gouv

ARDoISE

Atelier rennais de la donnée

1. données de recherche, de quoi parle t-on ?

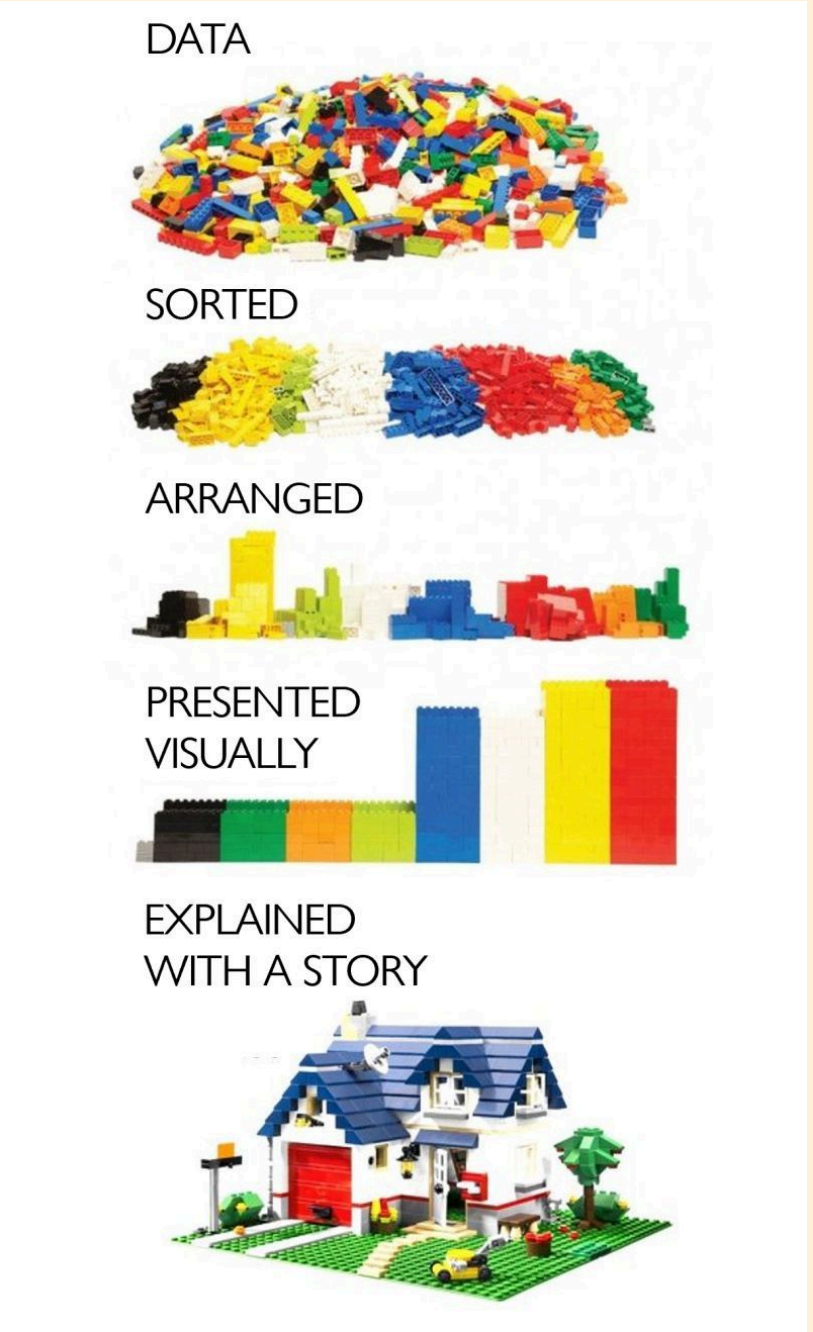


Figure 1: données brutes, données raffinées

quels fichiers sont importants à mettre à disposition ?

Réponses

- 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.p
- final_data.csv
- study_draft.qmd
- january_meeting_partners.docx
- fish_counter_instructions_for_u
- gantt_calendar.xlsx

2. pour une science cumulative, fiable et reproductible?

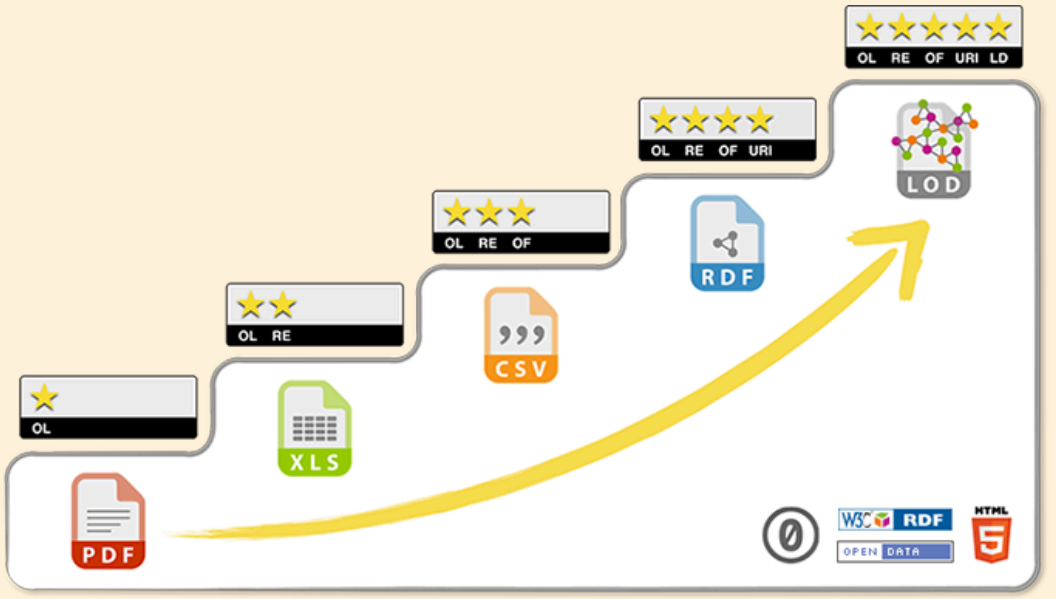


Figure 2: données figées, données liées

Pourquoi conserver ces données ?

pérennité de l’accès aux données

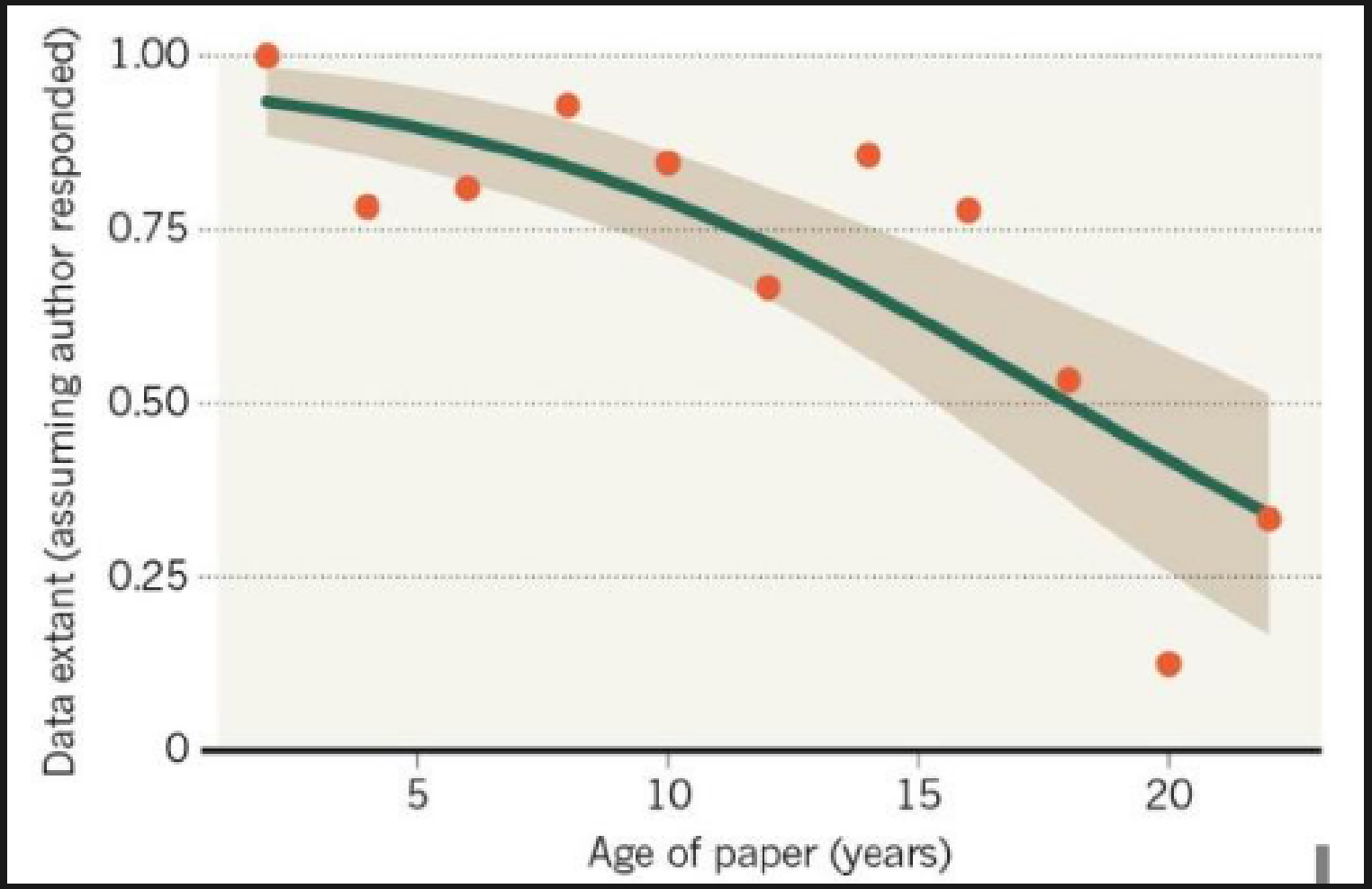
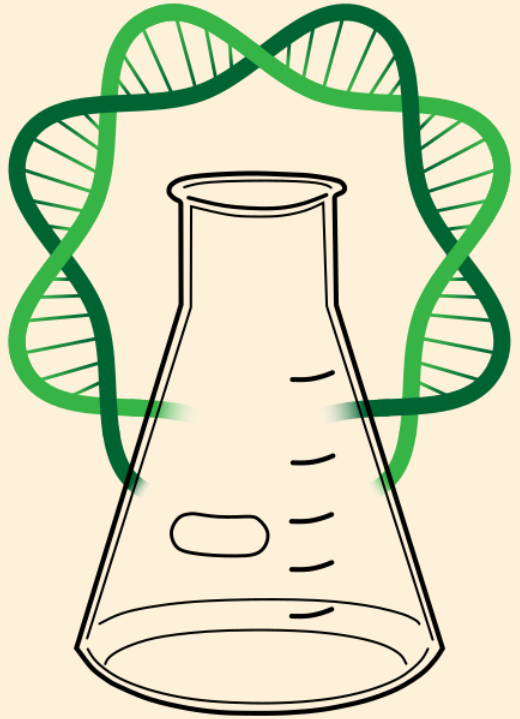


Figure 3: lutter contre la perte des données

Gibney & Van Noorden (2013)

3. Un enjeu de Science Ouverte



open science

principes FAIR

Figure 4: principes FAIR

ouverture / fermeture

- “aussi ouvert que possible, aussi fermé que nécessaire”
- Ouverture par défaut
- fermeture à justifier :
  - données personnelles
  - propriété intellectuelle

données personnelles

L’enjeu reste de donner accès dans la plus large mesure possible au jeu de données. Pour cela, on aura recours à des procédés visant à pseudonymiser ou anonymiser le jeu de données. La pseudonymisation consiste à supprimer les

propriété intellectuelle

En France, le régime du droit de la donnée fait que celles-ci par défaut sont la propriété des employeurs, et non celle des producteurs (chercheurs). Mais il peut y avoir des limites au partage de ces données : - licences contaminantes de certains

rendre ses données trouvables

Qualité d’un répertoire :

- renommée
- pérennité (institution support)
- licence ouverte
- identifiant pérenne
- richesse des métadonnées
- curation

discipline	entrepôt
images (SHS)	MediHal
code	Software Heritage via HAL
BioInformatique	GenOuest
Sciences Humaines	Nakala
Mathématiques	pas d'entrepôt, voir avec le groupe RNBM
environnement, hydrologie	Osuris
Sciences de la terre	data terra
Sciences de la mer	data ifremer, seanoe
sciences médicales	entrepôt INSERM sur RDG
Ecologie, environnement et société	Data.InDoRES et Cat.InDoRES

Recherche Data Gouv

- richesse des métadonnées
- curation
- référence nationale (soutenu par le Ministère)
- identifiant pérenne
- volumétrie importante
- gratuité
- génération simplifiée de datapapers
- bac à sable de RDG

données accessibles ?

Dans 93% des cas pas de réponse ou réponse négative sans justification Gabelica et al. (2022)

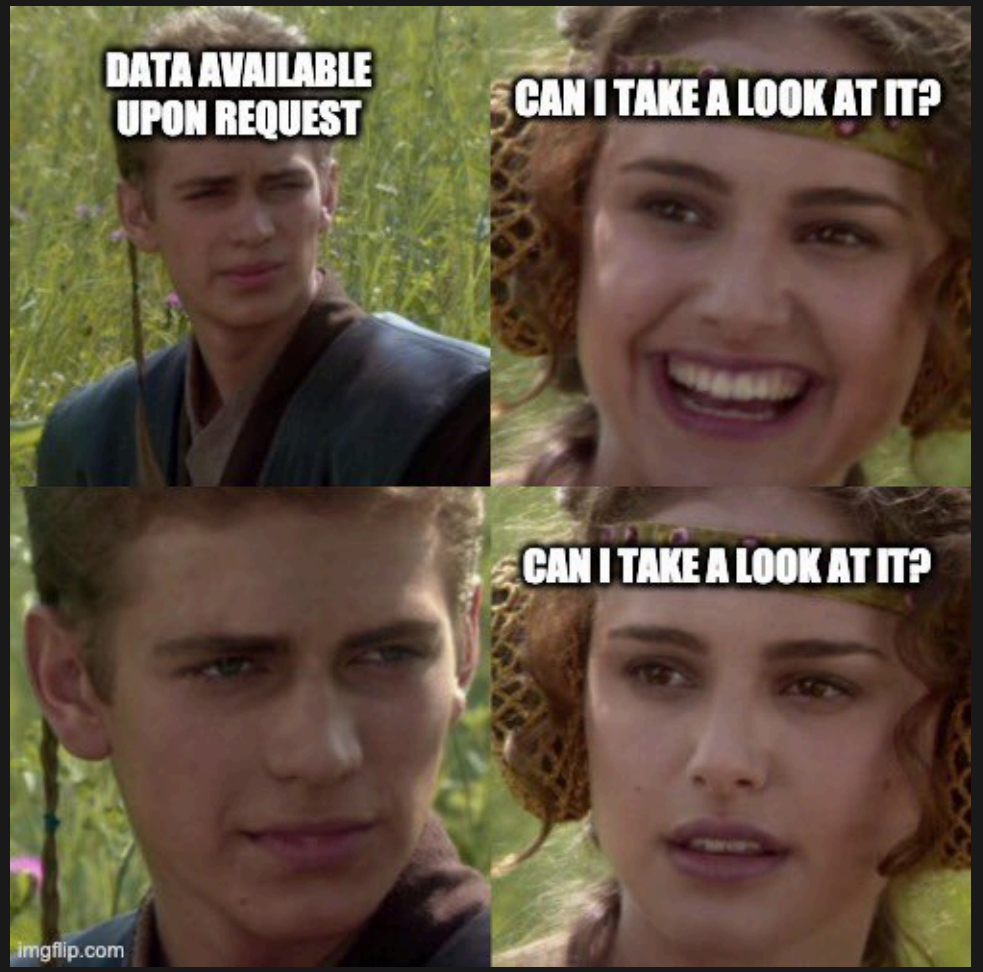


Figure 5: “data available upon request”

données interopérables ?

Quels identifiants utiliser pour le cuivre telluride ?

registre	identifiant
CAS number	12019-52-2
PubChem CID number	6914517
PubChem SID number	24879035
openSMILES identifier	[Cu].[Cu].[Te]
InChI identifier	InChI=1/2Cu.Te
MDL number	MFCD00049727

formats transparents ?

Figure 6: CSV vs XLS

Ziemann et al. (2023)

documenter les données

> Documentation is the glue that binds a data science project together (Ziemann et al. (2023))

- Décrire avec soin les données et le contexte de leur acquisition (production, collection)
- *literate programming*
- décrire les données en utilisant des ontologies

documenter pour éviter les erreurs de contexte

être précis dans la description du contexte de la production des données

Figure 7: l'importance du contexte

ontologies

discipline	thesaurus
biodiversité	INRAE
environnement	GEMET
Biologie, Santé	MeSH
Santé mentale	ascodopsy

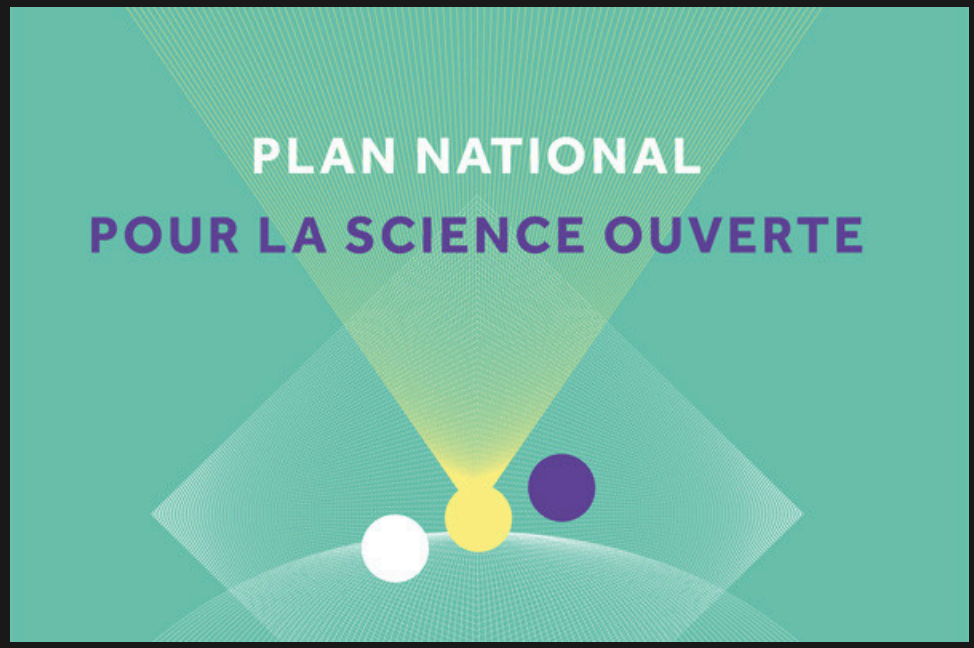
Le thésaurus Loterre (multidisciplinaire)

données réutilisables ?

- Creative Commons (CC:by)
- une licence écrite par un cabinet d'avocat expert en propriété intellectuelle et qui prévoit une multiplicité de cas d'usages autorisés ou prohibés
- ODBL
- Etalab
- pas de licence, on fait ce qu'on veut avec mon jeu de données
- CC0
- CC:by pour tous le monde sauf pour les industries fossiles, les vendeurs d'armes et Google (Thomas (2023)) texte disponible ici.

Le plan de gestion de données

- Le PGD résume tous les choix faits pour la gestion des données
- Déposer une première version d'un PGD 6 mois après la signature d'un contrat (ANR, Projets européens)
- DMP OPIDOR



4. Place à la pratique

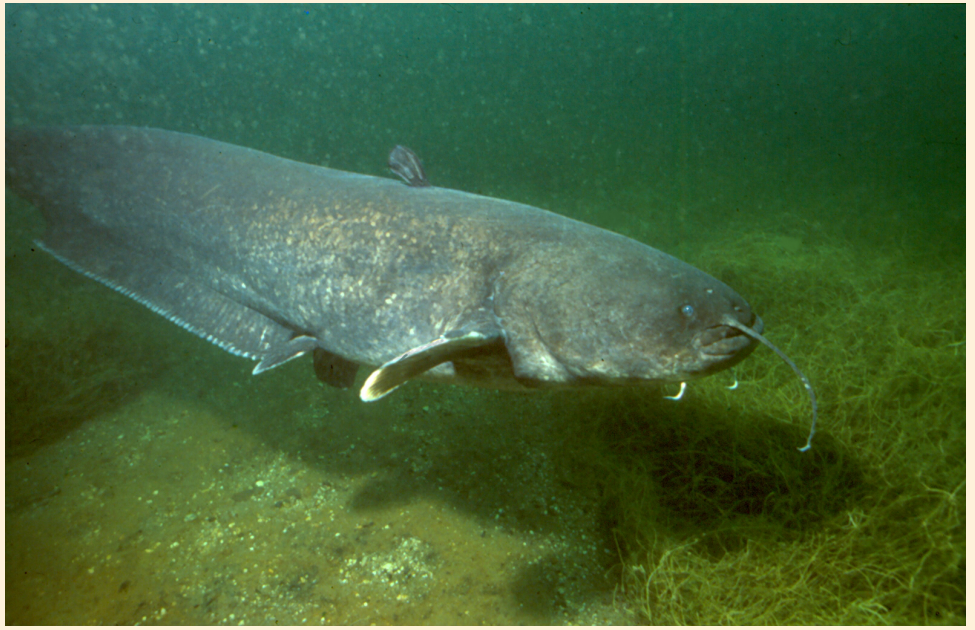


Figure 8: Silurus Glanis

données fictives (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	2	0*	10	12	4	8	6	0	0	**

Chère Prof. Armand,

Je joins à ce mail les données recueillies cette année par nos différents capteurs installés sur la Tydale dans le cadre de votre étude "Growth of glane silure catfish *silurus glanis* in european river, the case of the Tydale river".
Les fonds de la Royal Fisheries Corporation (RFC) ont permis d'acheter 8 capteurs sous-marins le long de la Tydale qui bien paramétrées étaient capables de ne compter que des poissons dont la masse était supérieure à 10 kg. Grâce à l'intelligence artificielle équipant la caméra, les silures ont été comptés avec une marge d'erreurs d'environ 3%.
Nous avons signalé dans les données obtenues certains incidents pouvant affecter la bonne conduite de l'étude.
En février et mai, nous avons noté que certains capteurs ne fonctionnaient plus correctement et avons du les réparer. En octobre, c'est le système centralisé de collecte des résultats qui est tombé en panne pendant 11 jours

figures

figure	source et crédits
Figure 1	Maricx
Figure 2	Tim Berners-Lee
Figure 3	Gibney, Van Noorden
Figure 4	Willkinson, Dumontier et al.
Figure 5	Sergio Uribe
Figure 6	meme dont l’origine se perd dans la nuit des temps
Figure 7	Ralph Aboujaoude Diaz
Figure 8	Dieter Florian

logiciels utilisés pour la présentation

Toute la présentation a été conçue avec du **logiciel libre** (merci Richard M. Stallman), et notamment les logiciels suivants :

- 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     2024-04-16
```

```
 pandoc   3.1.11 @ C:/PROGRA~1/Pandoc/ (via rmarkdown)

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

l’éditeur de texte est VScode. VScode rend l’utilisation de Quarto plus facile (Quarto est un logiciel dépourvu d’interface graphique, de même que Pandoc qui est intégré à Quarto). VScode permet d’encapsuler également des chuncks de R (si R est chargé sur la machine avec les packages correspondants)

Le package sessioninfo de R permet de montrer comment on peut documenter automatiquement le code utilisé pour produire une figure à partir d’un jeu de données par exemple. La reproductibilité du code est un enjeu essentiel et connexe à l’accès aux données.

Références

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 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