introduction à la gestion des données en sciences exactes

Damien Belvèze

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

expliquer ce qu'est un atelier de la donnée au niveau national Atelier de la donnée ARDOISE = Université de Rennes 2 et Université de Rennes

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

Figure 1: données brutes, données raffinées

quels fichiers sont importants à mettre à disposition ?

- 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.pptx
- final_data.csv
- study_draft.qmd
- january_meeting_partners.docx
- fish_counter_instructions_for_use.pdf
- gantt_calendar.xlsx

Les données de recherche sont des données qui ont une valeur démonstrative et sur lesquelles s'appuient les publications. Les *inscriptions* (Bruno Latour) qui sont produites dans la vie du projet ou du labo (compte rendu de réunion, mode d'emploi d'un matériel, agenda du projet ne sont pas des données de recherche. Leur conservation peut avoir un intérêt dans certains cas, mais pas dans le cadre d'un plan de gestion de données)

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.pptx
- final_data.csv
- study_draft.qmd
- january_meeting_partners.docx
- fish_counter_instructions_for_use.pdf
- gantt_calendar.xlsx

2. pour une science cumulative, fiable et reproductible?

Figure 2: données figées, données liées

Rappeler aux doctorants, le non-sens qu'il y a à "geler" ses données dans le support en PDF de leur thèse. Lors du dépôt de la publication dans HAL, les bibliothécaires de l'Université de Rennes essaient de récupérer auprès des auteurs les données en format réutilisable, quand elles ne se trouvent que dans le PDF de l'article.

La Science Ouverte repose sur des formats libres (transparence, format accessible à tous). Un document CSV vaux mieux qu'un fichier Excel (voir plus

loin)

L'étape suivante consiste à utiliser des URI (par exemple des éléments Wikidata) dans ses tableaux de données pour relier ces données à d'autres données du même genre (RDF)

Enfin on peut utiliser des outils comme Lodex pour transformer ses données en site web dynamique permettant de visualiser ses données de plusieurs manières et d'agréger son jeu de données à d'autres jeux de données

Pourquoi conserver ces données ?

jouer la vidéo de façon synchronisée avec une classe - télécharger la vidéo depuis le mediaserver et l'enregistrer sur le bureau
- l'ouvrir sans la jouer
- partager l'écran depuis Zoom, choisir le lecteur de vidéo
- cocher les deux cases en bas de la fenêtre de partage (audio pour tous + optimisation)
- jouer la vidéo puis cesser le partage quand elle est finie

pérennité de l'accès aux données

Figure 3: lutter contre la perte des données

Gibney & Van Noorden (2013)

3. Un enjeu de Science Ouverte

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 données directement identifiantes, ou à les remplacer par des pseudonymes. Cette méthode comporte des risques de réidentification par croisement du jeu de données avec d'autres jeux portant sur la même population. L'anonymisation des données est une méthode plus subtile consistant, après suppression des données directement identifiantes, à rendre les données moins précises ou à y ajouter du bruit pour entraver les tentatives éventuelles de réidentification. Il s'agit d'un compromis à trouver entre l'intérêt des personnes dont les données sont collectées et la finesse de ces données. Des outils existent pour anonymiser les données : Arx, Amnesia. L'anonymisation des données entraîne des coûts de traitement (temps humain, prestatation) qu'il convient de prévoir en amont dans la demande de budget.

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 matériels propriétaires utilisées pour produire des données (par exemple imagerie médicale) - projets de recherche internationaux : quel droit national on va appliquer sur ces données ?

rendre ses données trouvables

Qualité d'un répertoire :

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

Zenodo : pas de versionnage des jeux de données, métadonnées moins riches que sur RDG, mais entrepôt pérenne (European Commission)

pour chercher des entrepôts et comparer leurs offres : https://re3data.org

Eviter les entrepôts propriétaires (Figshare)

critères de conformité des entrepôts : https://www.ouvrirlascience.fr/entrepots-de-donnees-de-confiance-criteres-de-conformite/

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

Sur Zenodo, on est limité à 50 gigaoctets par dataset sur RDG, on est limité à 50 gigaoctets par fichier, un dataset peut comporter plusieurs fichiers (pas vu de limitation globale), donc oui, apparemment la volumétrie est moins limitée sur RDG que sur Zenodo.

données accessibles ?

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

Figure 5: "data available upon request"

La personne inscrite dans le PGD comme personne contact pour les données peut être le corresponding author de la publication liée ou bien une autre personne, mais cette personne doit avoir conscience de son rôle si sa médiation est nécessaire pour obtenir les données. Mieux vaut renseigner ici une adresse personnelle (type gmail) qu'une adresse institutionnelle, car on peut changer fréquemment d'institution.

données interopérables ?

Table 2: 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

Dans les entrepôts on va plutôt utiliser inchi que CAS car pour accéder aux identifiants CAS il faut disposer de la base de données CAS Sci-Finder

PubchemID est considérée comme un meilleur choix qu'InCHi, car plus *machine-redeable*

formats transparents ?

Figure 6: CSV vs XLS

Ziemann et al. (2023)

Utiliser Excel pour analyser des données (comme le font 69% des chercheurs d'après une étude de 2016) est une mauvaise idée. Utiliser plutôt des outils transparents comme R. Une étude de 2021 montre qu'Excel change régulièrement des gènes en dates, et le phénomène est connu de longue date. La #reproductibilité des études passe par le maniement d'outils libres et transparents sur ce qu'ils font. Reférences ici : https://doi.org/10.1093/bib/bbad375

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

possibilité de lier le plan de gestion des données à un protocole déposé sur un autre site (par exemple https://www.protocols.io) ou Prospero

https://www.crd.york.ac.uk/prospero/ si les données sont bibliographiques et qu'il s'agit d'une revue de littérature

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.

Demander aux élèves quelles sont les données mentionnées dans cette slide qui sont appropriées pour un jeu de données.

Réponses :

Les licences Etalab et ODBL sont recommandées en France pour permettre la réutilisation des données. Les Creative Commons sont plutôt appropriées aux résultats de la recherche (publications, preprints, posters…) la licence CC0 n'est pas autorisée en France où le droit d'auteur commande au minimum de citer l'auteur du jeu de données, ce que ne prévoit pas cette licence.

Lorsqu'aucune licence n'est indiquée, on ne peut pas réutiliser de manière sûre le jeu de données. Quand il n'y a pas de licence, par défaut, c'est le droit d'auteur qui s'applique.

Dans le cadre de la Science Ouverte, les licences sont standardisées et les usages qu'elles permettent sont facilement énonçables et compréhensibles. Si en guise de licence on a un texte qui ressemble aux CGU d'un service web, c'est que le jeu de données risque d'être difficile à réutiliser (nécessité probablement de contacter les producteurs)

La Science Ouverte peut être vue par certaines chercheurs comme une perte de contrôle sur leurs publications et données et les licences restrictives qui entravent la réutilisation à des fins jugées mauvaises par les chercheurs est encore à l'état de projet. Ces licences ne sont pas usuelles pour les données.

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	9	0	**
16	16	6	1	8	2	7	4	11	3	10	**
17	17	3	4	7	5	1	9	2#	10	6	**
18	18	5	9	6	11	3	1	10	8	2	**
19	19	7	6	0	4	10	12	5#	1	8	3
20	20	11	2	9	3	8	5	6	4	1	7
21	21	1	7	3	5	6	4	8	11	10	2
22	22	0	3	8	4	2	10	11	12	0	7
23	23	3	7	2	5	4	8	1	6	11	10
24	24	6	4	9	0	11	3	8	5	2	1

25	25	8	1	11	3	6	9	4	7	5	2
26	26	2	10	7	12	0	4	11	1	6	3
27	27	4	6	0	1	5	7	2	10	8	12
28	28	7	5	1	9	10	2	6	3	4	11
29	29	11	N/A	6	2	3	1	7	9	12	4
30	30	1	N/A	4	11	7	5	10	2	3	8
31	31	9	N/A	10	N/A	1	N/A	3	8	N/A	5

	November	December
1	6	1
2	9	12
3	10	6
4	8	4
5	0	5
6	4	9
7	1	7
8	5	2
9	11	10
10	3	1
11	6	4
12	2	8
13	4	3
14	1	5
15	0	2
16	8	11
17	3	7
18	7	1
19	2	6
20	9	10
21	4	9
22	6	1
23	9	12
24	10	6
25	8	4
26	7	5
27	4	9
28	1	7
29	5	2
30	11	10
31	N/A	1

jeu de données inspiré de Gouvernement du Royaume Uni. (2023). River Tyne Fish Count [jeu de données]. https://www.gov.uk/government/statistical-data-sets/river-tyne-fish-counts

Le jeu de données ne comporte pas d'informations sur certains signes (*) ou (#). Ces signes sont explicitées dans le mail. On insiste ici sur la nécessité de documenter correctement le jeu de données pour qu'il ne manque rien à sa

compréhension.

```
Chère Prof. Armand,

Je joins à ce mail les données recueillies cette année par nos différents capteurs installés
Les fonds de la Royal Fisheries Corporation (RFC) ont permis d'acheter 8 capteurs sous-marin
Nous avons signalé dans les données obtenues certains incidents pouvant affecter la bonne co
En février et mai, nous avons noté que certains capteurs ne fonctionnaient plus correctement
Par ailleurs, nous avons signalé que les activités nautiques sur la Tyndale qui avaient lieu

J'espère que ces chiffres vous permettront malgré tout de faire progresser votre étude et de

En vous souhaitant une très bonne fin de journée,

Mickael. J. Bernache, Biodiversity Research Institute of Portland
```

Le mail comporte des informations sur le contexte et les limites de la collecte des données. Ces informations aident à remplir les champs concernant le jeu de données.

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
cli           3.6.2   2023-12-11 [1] CRAN (R 4.3.2)
digest        0.6.34  2024-01-11 [1] CRAN (R 4.3.2)
evaluate      0.23    2023-11-01 [1] CRAN (R 4.3.2)
fastmap       1.1.1   2023-02-24 [1] CRAN (R 4.3.2)
htmltools     0.5.7   2023-11-03 [1] CRAN (R 4.3.2)
jsonlite      1.8.8   2023-12-04 [1] CRAN (R 4.3.2)
knitr         1.45    2023-10-30 [1] CRAN (R 4.3.2)
rlang         1.1.3   2024-01-10 [1] CRAN (R 4.3.2)
rmarkdown     2.25    2023-09-18 [1] CRAN (R 4.3.0)
rstudioapi    0.15.0  2023-07-07 [1] CRAN (R 4.3.2)
sessioninfo   1.2.2   2021-12-06 [1] CRAN (R 4.3.2)
xfun          0.42    2024-02-08 [1] CRAN (R 4.3.3)
yaml          2.3.8   2023-12-11 [1] CRAN (R 4.3.2)

[1] C:/Program Files/R/R-4.3.0/library
```

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