gérer, diffuser et partager ses données de recherche

Damien Belvèze

28 février 2023

De quoi parle t-on ?

Quelques définitions concernant les données de la recherche

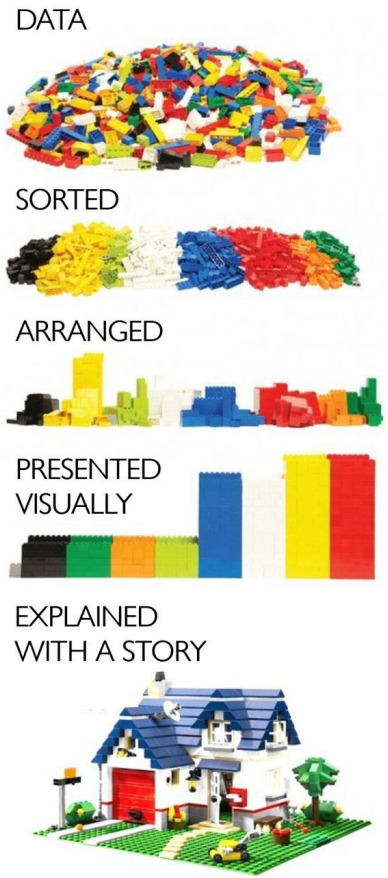
- **Données de recherche** (research data): « Enregistrements factuels (chiffres,textes, images et sons), qui sont utilisés comme sourcesprincipales pour la recherche scientifique et sont généralement reconnus par la communauté scientifiqe comme nécessaires pour valider des résultats de recherche. » [@ocdePrincipesLignesDirectrices2007]

- **Jeux de données** (datasets) « Agrégation (...) de données brutes ou dérivées présentant une certaine "unité" unité", rassemblées pour former un ensemble cohérent » (Gaillard, 2014)

- **science ouverte** (Open Science) :

> La science ouverte est la pratique de la science de manière à ce que d'autres puissent collaborer et contribuer, où les données de recherche, les notes de laboratoire et les autres processus de recherche sont librement disponibles, dans des conditions qui permettent la réutilisation, la redistribution et la reproduction de la recherche et des données et méthodes sous-jacentes.

De quelles données parle t-on ?

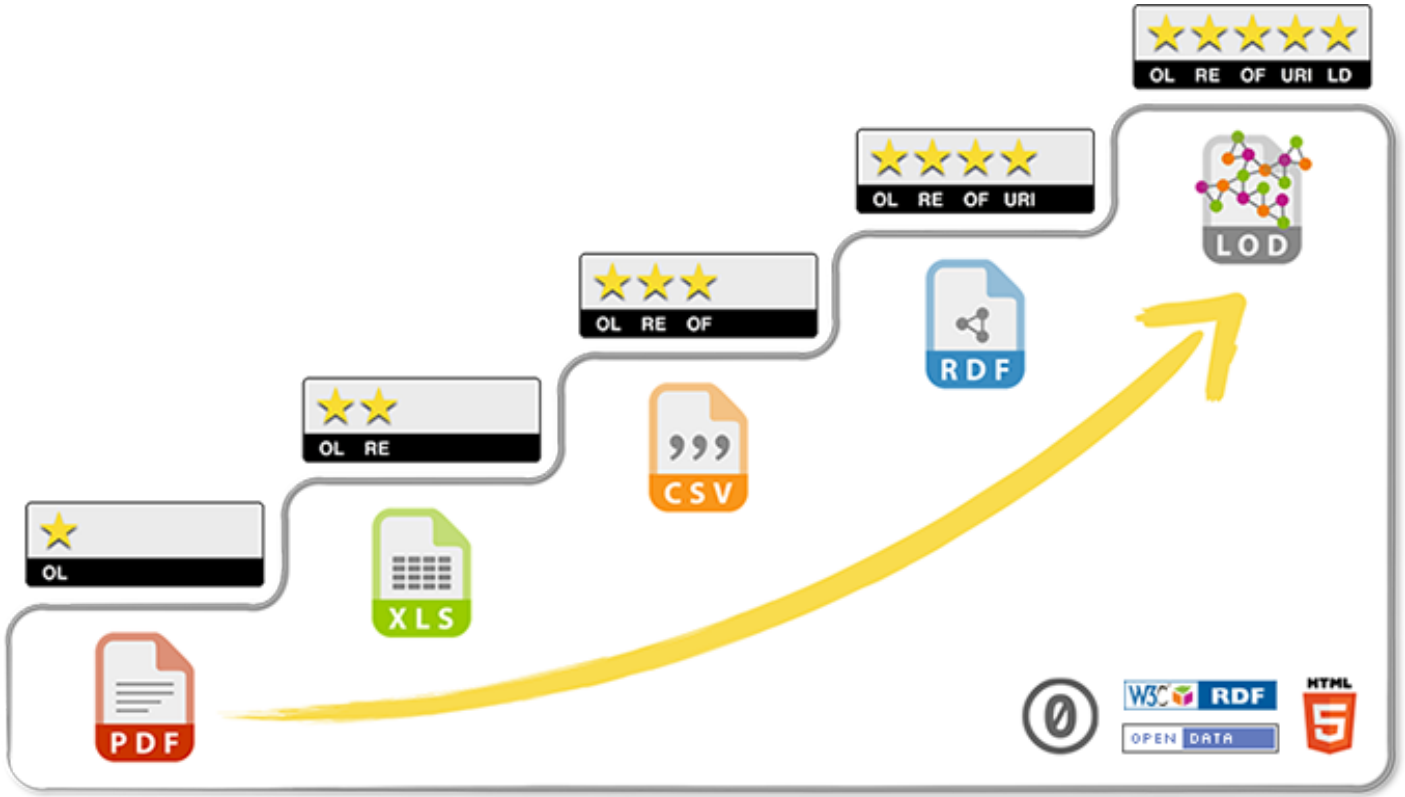


- Données brutes
- Données traitées ou dérivées
- Données analysées ou interprétées

> Data doesn’t say anything. Humans say things.
> (Andreo Jones-Rooy)

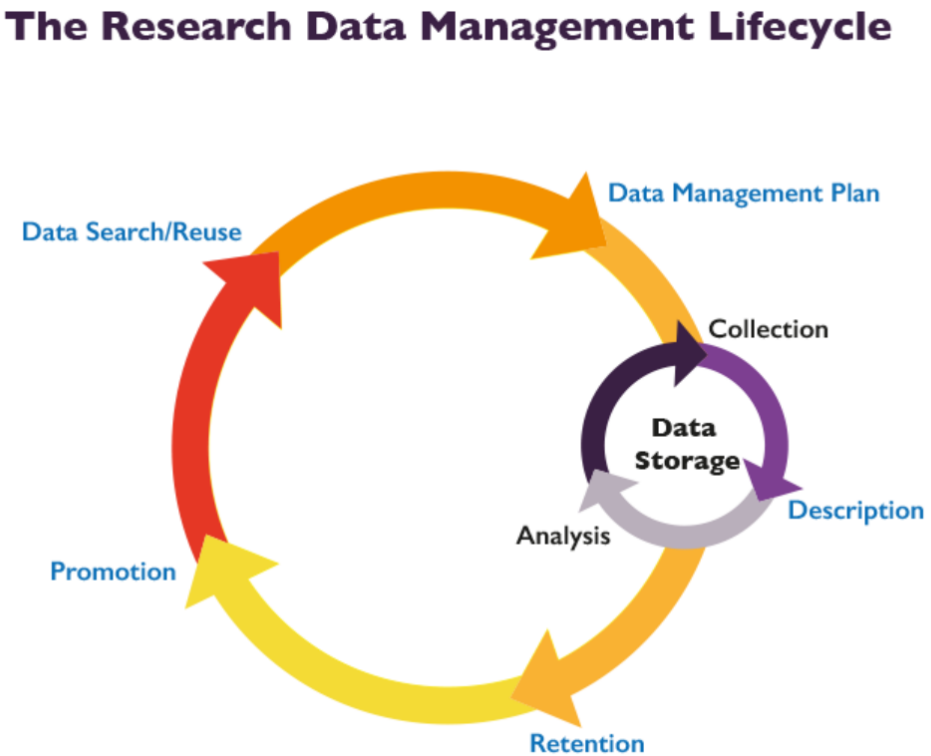
des données inégalement mises à disposition

Dès 1997, les pionniers du web s'intéressent à la manière dont les données doivent être structurées pour être réutilisables.



Et vous, quelles sont vos données ?

le cycle de vie de la donnée



Source: Colorado State University Library

Les enjeux liés à la conservation et au partage des données

Enjeux économiques

> les données de la recherche sous forme numérique sont de plus en plus utilisées dans des travaux qui vont au delà du projet en vue duquel elles ont été recueillies à l'origine, ainsi que dans d'autres domaines de recherche et dans l'industrie

Rapport de l'OCDE de 2007

Enjeux de santé

le partage des données médicales fait progresser la médecine.

Importance du recueil et de la diffusion des *meta-analyses sur données individuelles*

> Sans recruter un patient de plus, grâce à une méta-analyse sur données individuelles, l'Institut Gustave-Roussy a réussi à montrer que la chimiothérapie concomitante est plus efficace que le schéma séquentiel (Florian Naudet, Claude Pellen)

Source : The conversation

Enjeux démocratiques

la donnée comme bien commun de la connaissance

Obépine (projet de recherche) = taux de Sars-Cov-2 dans les eaux usées - demande CADA

> "Vous avez formulé le 20 novembre 2020 une demande d'ouverture des données du projet de recherche Obépine."
>
> "Votre demandes d'accès s'inscrit dans le cadre de la loi 2016-1321 du 7 octobre 2016 pour une République numérique (dite "loi Lemaire") en faveur de la "circulation des données et du savoir"
>
> J'ai le plaisir de vous informer que nous rendrons disponibles et réutilisables les données de courbe de tendance dans les meilleurs délais

demande Dada

data ou capta ?

Consentement de la personne, RGPD

Les enjeux liés à la conservation et au partage des données

Enjeux scientifiques

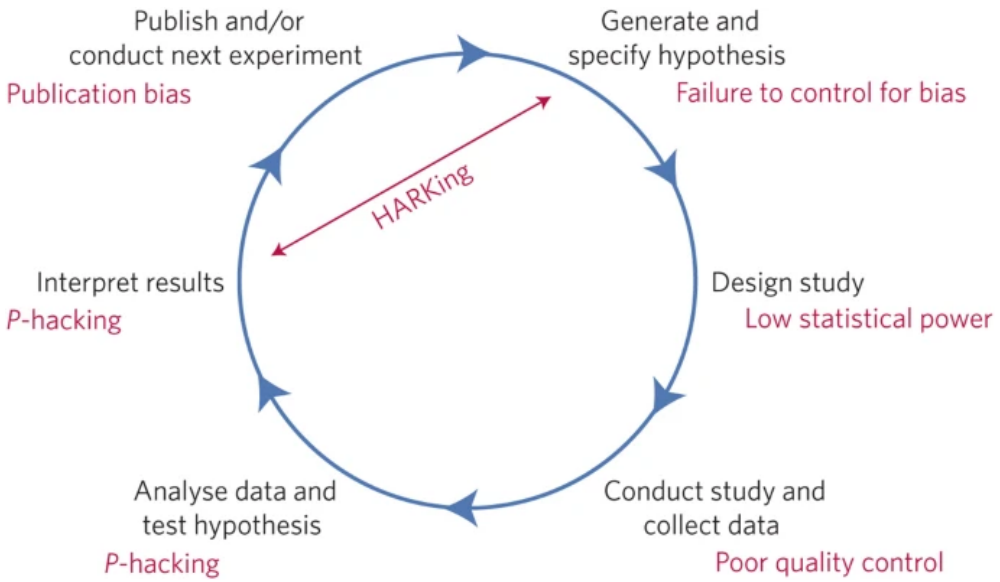
Un enjeu de reproductibilité et donc de fiabilité de la science.

ex. Rétractation en 2020 d'un article du Lancet sur l'hydroxychloroquine. Les auteurs...

> « n’ont pas été en mesure d’effectuer un audit indépendant des données qui sous-tendent leur analyse », écrit The Lancet. En conséquence, ils ont conclu qu’ils « ne peuvent plus garantir la véracité des sources de données primaires.

Le Monde, 4 juin 2020

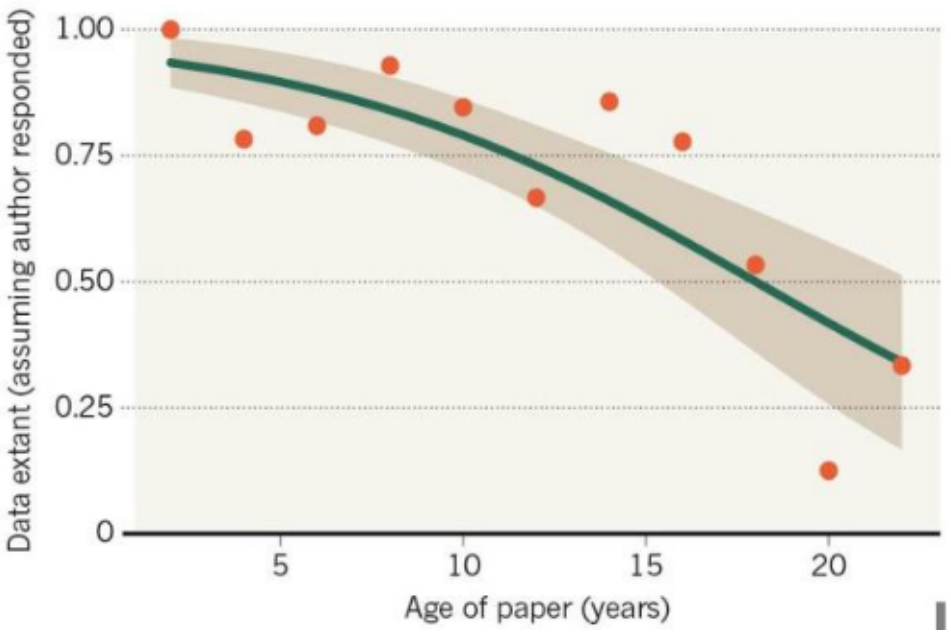
L'accès aux données et aux traitements sur ces données : une garantie de reproductibilité



poor quality control on data Manifesto for reproducible science

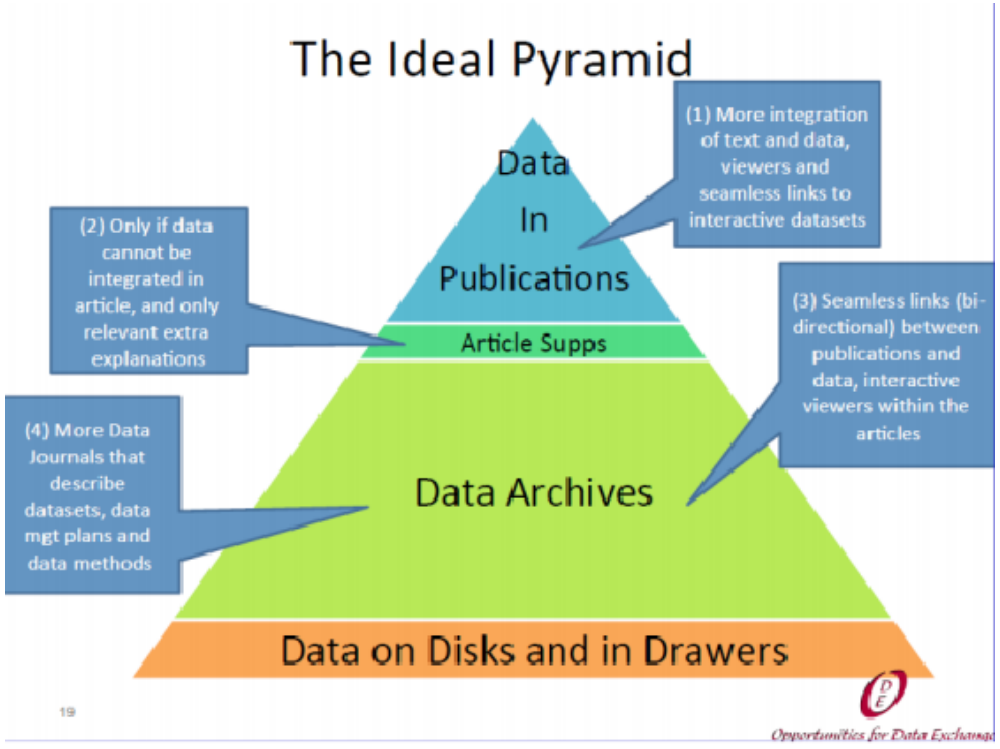
un manque de pérennité de ces données

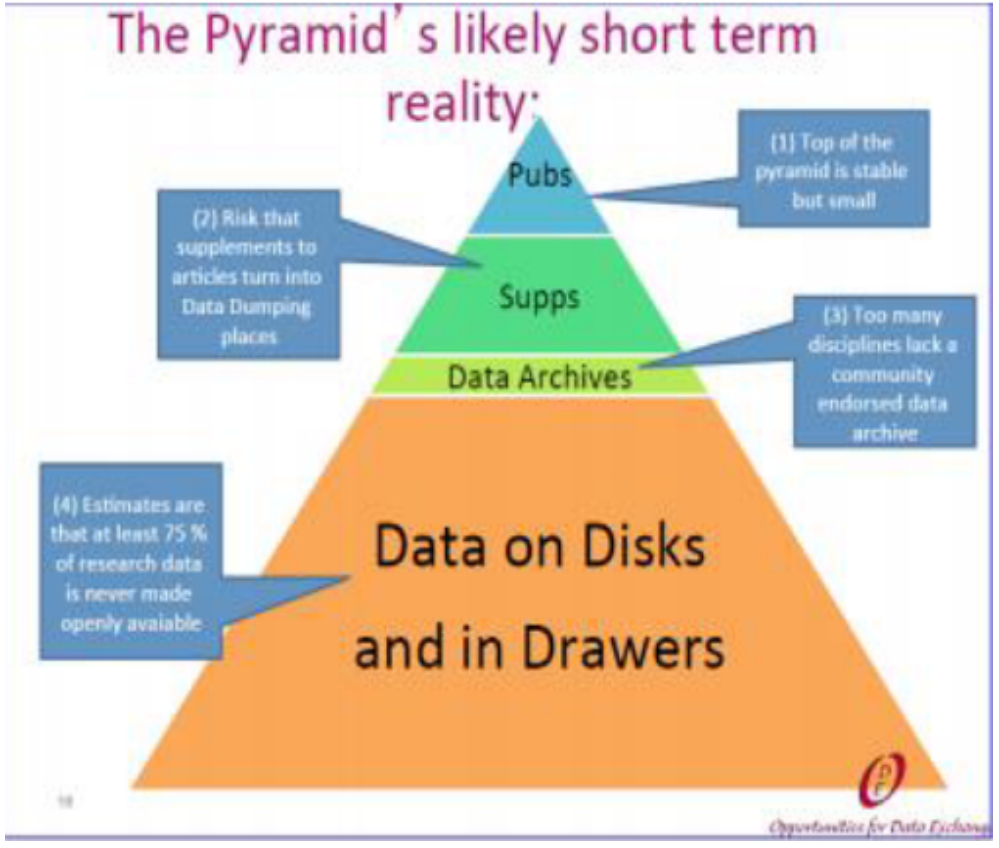
80 % des données produites ces 20 dernières années seraient perdues (Vines, Albert, 2014).



PLus le temps passe et plus les données associées à une publication disparaissent

Comment ça devrait être et comment c'est





(Marie Puren, 2021)

Le Plan National pour la Science Ouverte

- Généraliser l'accès ouvert aux publications
- **Structurer, partager et ouvrir les données de la recherche**
- Ouvrir et promouvoir les codes sources produits par la recherche
- Transformer les pratiques pour faire de la science ouverte le principe par défaut

Ministère de l'enseignement supérieur et de la recherche

Les principes FAIR

Findable / Accessible / Interoperable / Reusable

Comment rendre mes données *trouvables*

- dataset -> DOI (exemple : https://doi.org/10.5281/zenodo.5903186)
- Les répertoire de données doivent être correctement indexés par les moteurs de recherche (thesaurus -> SKOS)

comment rendre mes données accessibles ?

> aussi ouvertes que possible, aussi fermées que nécessaires

être explicite sur le type de partage et les modalités de partage prévues pour ce jeu de données

avantage citationnel

Les articles qui donnent un accès explicite dans un entrepôt distant ont un **bénéfice de citations de 25%**

Category	Definition	Example
0	Not available	*No additional data available* (common).
1	Data available on request or similar	*Supporting information is available in the additional files and further supporting data is available from the authors on request* (DOI: 10.1186/1471-2164-14-876).
2	Data available with the paper and its supplementary files	*The authors confirm that all data underlying the findings are fully available without restriction. All data are included within the manuscript* (DOI: 10.1371/journal.pone.0098191).
3	Data available in a repository	*The authors confirm that all data underlying the findings are fully available without restriction. The transcriptome data is deposited at NCBI/Gene Bank as the TSA accession SRR1151079 and SRR1151080* (DOI: 10.1371/journal.pone.0106370).

https://doi.org/10.1371/journal.pone.0230416.t001

Comment rendre mes données interopérables

- utiliser des formats ouverts de données
- utiliser des vocabulaires qui font référence dans le domaine de recherche

et pourquoi je partagerais MES données ?

- C'est beaucoup de travail de les obtenir
- ces données confèrent à mon labo un avantage concurrentiel
- Loi Lemaire : le public préempte les données pour éviter qu'elles deviennent la propriété d'acteurs privés (enclosures)
- Les données du chercheur appartiennent en fait à son employeur (contrairement à ces oeuvres)

créer un plan de gestion de données

où trouver des exemples, des modèles ?

- *templates* et exemples disponibles sur https://dmp.opidor.fr
- Pour trouver un **plan de gestion de données structuré** (machine-readable), suivre la procédure indiquée sur le site

Obligatoire dans le cadre des financements ANR

au plus tard 6 mois après la signature du contrat

Le PGD indique comment les données vont être collectées, traitées de telle sorte qu'elles soient conformes aux principes FAIR

Où doivent être conservées les données ?

Pas chez l'éditeur

- répertoires de données = plus de latitude pour commenter les données
- l'hébergement des données par l'éditeur # ouverture des données ou réutilisations possibles
- les multinationales de l'édition scientifique peuvent avoir un usage de ces données qui est contraire à l'esprit de la Science ouverte

Où seront conservées les données alors ?

- Déposer ses données dans un entrepôt **avant de publier** (une fois l'article publié, plus possible de faire un lien de l'article vers le jeu de données sur le site de l'éditeur) comme on peut le faire en revanche sur l'entrepôt de données vers le site de l'éditeur. Le seul moyen qui nous reste est de faire un lien dans le MAA qu'on dépose sur l'archive ouverte.
- possibilité de réserver un emplacement (ID) sur un entrepôt de données avant de déposer les données et faire mention de cet ID (DOI) dans l'article qu'on rédige, même si l'emplacement est encore vide des données qu'on va y verser après publication.
- jeu de données original -> prévoir avant la bibliographie un espace "disponibilité des données",
- jeu de données réutilisé, le mentionner dans la bibliographie.

réfléchir au coût de la préservation des données

- Le coût de la préservation des données doit être pesé par rapport au coût de leur génération / collecte
- Données difficiles ou coûteuses à collecter > plus longue préservation. (exemple : le reséquençage est devenu moins coûteux que la préservation)

la conservation sous l'angle de la reproductibilité :

- conserver les données initiales et bien documenter les divers traitements
- données computationnelles : virtualisation (Docker) et versions des traitements (github)
- Ne pas conserver les données issues des traitements successifs

> tout ce qui peut être rejoué, peut être supprimé, on n’aura perdu que du temps de calcul" (Olivier Collin)

De l'importance de documenter ses données

fichier README déposé en même temps que les données qui précise ce qu'il y a dans le PGD:

- organisation et localisation des fichiers de données
- nature des observations et variables présentes dans les fichiers
- détails sur les conditions d'expérimentation qui ont permis la collecte des données
- détails sur les traitements informatiques faits sur les données

description : ontologies et thésaurus.

Ontologies :

- ontologies génomiques
- ontologies biologiques

thésaurus :

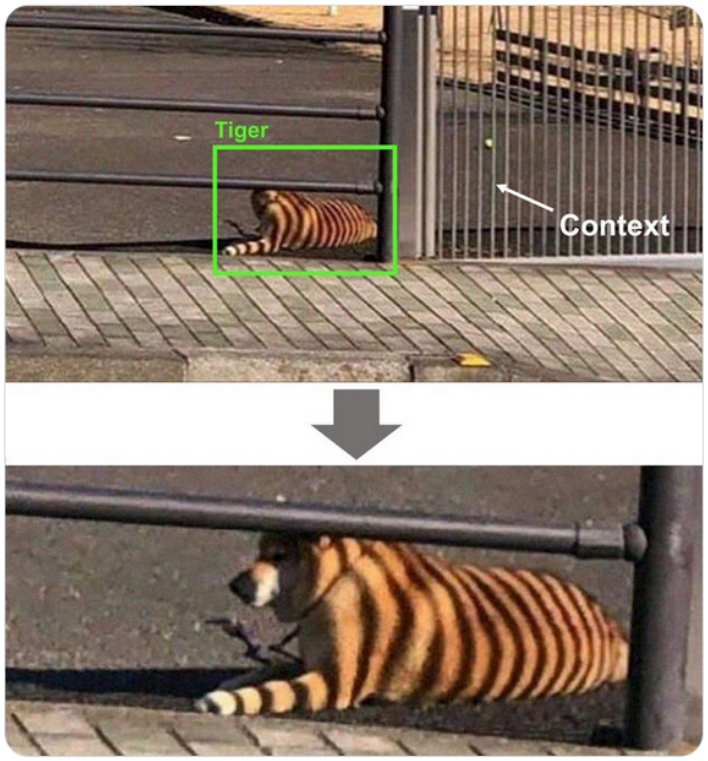
- MeSH (médecine)
- thésaurus de l'INRAE

La description montre comment interpréter les données

605 Tweets

Rodrigo C de Almeida @rodcoutalmeida · 24 mars

Without understanding the context data is just random information.
Specially in #Biology #Bioinformatics #DataScience



la publication sous l'angle de la vie privée

pseudonymiser ou anonymiser

Outils : Amnesia et Arx

- Anonymiser un jeu de données consiste non pas à pseudonymiser les enregistrements mais à faire disparaître les informations nominatives tout en essayant de rendre difficiles les opérations de réidentification.
- Plus les données personnelles contenues dans un fichier sont nombreuses et plus les risques de réidentification sont importants.

crédits

Cette présentation est très inspirée de deux autres présentations :

- Thierry Fournier, introduction aux données de la recherche en sciences exactes, 2022
- Cécile Arènes, Rédiger un plan de gestion des données, 2021

Références

Blanc, I. (2020, novembre 20). Données relatives à la concentration de SARS-CoV-2 dans les eaux usées—Une demande d'accès à l'information à Ministère de l'enseignement supérieur, de la recherche et de l'innovation. Consulté 14 avril 2022, à l'adresse Ma Dada website: https://madada.fr/demande/donnees_relatives_a_la_concentra

Brunori, G. (2020). Data Management Plan. https://doi.org/10.5281/zenodo.3664215 Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. PLOS ONE, 15(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. Nature Human Behaviour, 1(1), 1-9. https://doi.org/10.1038/s41562-016-0021 OCDE. (2007). Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics. Consulté à l'adresse https://www.oecd.org/fr/science/inno/38500823.pdf

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. PLOS Computational Biology, 9(10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285 « The Lancet » annonce le retrait de son étude sur l'hydroxychloroquine. (2020, juin 4). Le Monde.fr. Consulté à l'adresse https://www.lemonde.fr/sciences/article/2020/06/04/hydroxychloroquine-trois-auteurs-de-l-etude-du-lancet-se-retractent_6041803_1650684.html

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., … Rennison, D. J. (2014). The Availability of Research Data Declines Rapidly with Article Age. Current Biology, 24(1), 94-97. https://doi.org/10.1016/j.cub.2013.11.014