Cycle de vie de la donnée et science ouverte : une approche intégrative

data2019opensci : En route vers la science ouverte : le cycle de vie des données

Esther Dzalé Yeumo

Agenda

- L’ouverture de la science : un élément de réponse à la crise de reproductibilité
- Des exemples d’impacts de l’ouverture des données
- Les enjeux en lien avec le cycle de vie de la donnée

Définition

Open Science is about **extending the principles of openness to the whole research cycle** (see figure 1), fostering sharing and collaboration as early as possible thus entailing a systemic change to the way science and research is done.

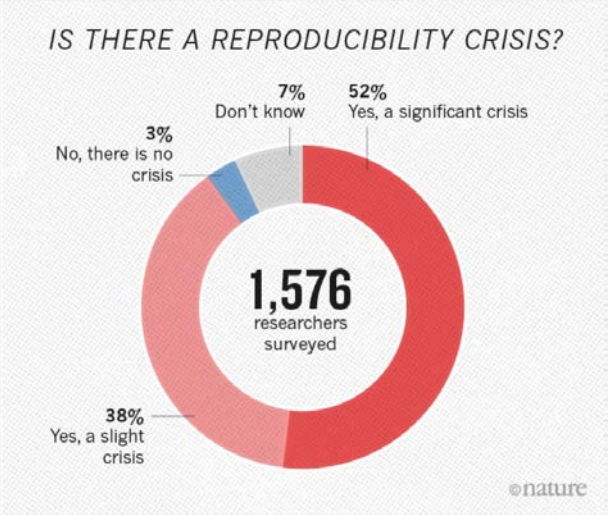
https://www.fosteropenscience.eu/content/what-open-science-introduction

Diederik Stapel

Professeur

Une crise de la reproductibilité

- Selon une étude réalisée auprès de 1 500 scientifiques et publiée par Nature en 2016 (https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970?WT.mc_id=SFB_NNEWS_1508_RHBox) :
  - plus de 70 % des chercheurs disent avoir échoué à reproduire l'expérience scientifique d'un autre chercheur
  - plus de 50% reconnaissent n'avoir pas réussi à reproduire leur propre expérience
- La crise de la reproductibilité concerne la reproduction des expériences scientifiques mais aussi la réutilisation des données brutes pour reproduire de façon indépendante des analyses statistiques



Reproducibility and the conduct of research

Data dredging
Also known as p-hacking, this involves repeatedly searching a dataset or trying alternative analyses until a 'significant' result is found.

Omitting null results
When scientists or journals decide not to publish studies unless results are statistically significant.

Underpowered study
Statistical power is the ability of an analysis to detect an effect, if the effect exists – an underpowered study is too small to reliably indicate whether or not an effect exists.

Issues

Errors
Technical errors may exist within a study, such as misidentified reagents or computational errors.

Underspecified methods
A study may be very robust, but its methods not shared with other scientists in enough detail, so others cannot precisely replicate it.

Weak experimental design
A study may have one or more methodological flaws that mean it is unlikely to produce reliable or valid results.

Improving reproducibility will ensure that research is as efficient and productive as possible. This figure summarises aspects of the conduct of research that can cause irreproducible results, and potential strategies for counteracting poor practice in these areas. Overarching factors can further contribute to the causes of irreproducibility, but can also drive the implementation of specific measures to address these causes. The culture and environment in which research takes place is an important 'top-down' overarching factor. From a 'bottom-up' perspective, continuing education and training for researchers can raise awareness and disseminate good practice.

Figure taken from the report of the symposium, 'Reproducibility and reliability of biomedical research', organised by the Academy of Medical Sciences, BBSRC, MRC and Wellcome Trust in April 2015. The full report is available from http://www.acmedsci.ac.uk/researchreproducibility.

Possible strategies

Open data
Openly sharing results and the underlying data with other scientists.

Pre-registration
Publicly registering the protocol before a study is conducted.

Collaboration
Working with other research groups, both formally and informally.

Automation
Finding technological ways of standardising practices, thereby reducing the opportunity for human error.

Open methods
Publicly publishing the detail of a study protocol.

Post-publication review
Continuing discussion of a study in a public forum after it has been published (most are reviewed before publication).

Reporting guidelines
Guidelines and checklists that help researchers meet certain criteria when publishing studies.

https://acmedsci.ac.uk/file-download/38208-5631f0052511d.pdf

L’ouverture des données impacte positivement la qualité de la science

Illustration à travers quelques exemples

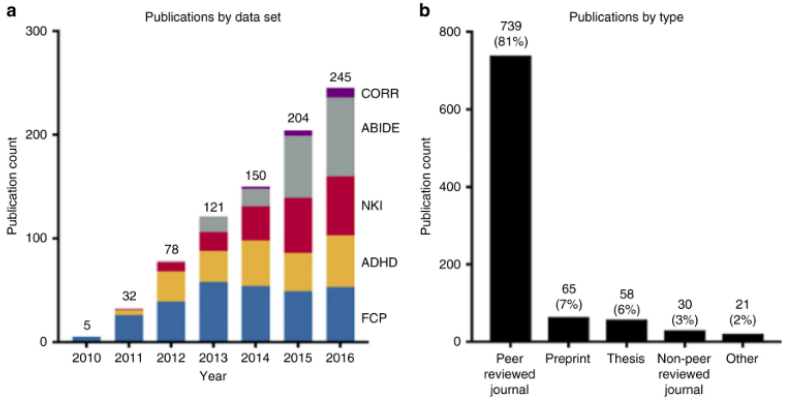
- Une question de principe: « Public money? Public data! »
- Une science de meilleure qualité
- Une science plus ouverte, plus accessible à la société

Extrait de
https://opensciencesud.sciencesconf.org/data/pages/Pourquoi_la_science_ouverte_Marin_Dacos.pdf

Impacts sur les publications

Fig. 1

From: Assessment of the impact of shared brain imaging data on the scientific literature




nature communications

Article | Open Access | Published: 19 July 2018

Assessment of the impact of shared brain imaging data on the scientific literature

Michael P. Milham, R. Cameron Craddock, Jake J. Son, Michael Fleischmann, Jon Clucas, Helen Xu, Bonhwang Koo, Anirudh Krishnakumar, Bharat B. Biswal, F. Xavier Castellanos, Stan Colcombe, Adriana Di Martino, Xi-Nian Zuo & Arno Klein

Nature Communications 9, Article number: 2818 (2018) | Cite this article

1686 Accesses | 9 Citations | 69 Altmetric | Metrics

https://www.nature.com/articles/s41467-018-04976-1#Tab2

L'ouverture des données

- De plus en plus de publications citant des données mises à disposition de manière ouverte
- Baser ses articles sur des données partagées n'empêche pas de publier dans des revues à facteur d'impact
- Le partage de données permet d'intéresser des experts d'autres domaines scientifiques
- La réutilisation de données partagées permet de réaliser des économies substantielles

Table 2 Quantifying the money saved through the reuse of data

From: Assessment of the impact of shared brain imaging data on the scientific literature

Database	Cost/subject	Phenotyping	Phenotyping	Clinical	Population	Difficulty	No. of publications	No. of scans/subject	$ Saved
		Minimal	Comprehensive	Low	Moderate	High			
FCP	$1000	x					308	1	101,003,000
ADHD-200	$2000–5000			x	x		210	1	526,275,000
NKI-RS	$3000		x				188	1	70,065,000
ABIDE	$5000–10,000				x	x	190	1	995,560,000
CoRR	$2000	x					17	2	70,065,000

INRA SCIENCE & IMPACT

La contribution inattendue des experts improbables »

Chris Raimondi was one of the early winners on Kaggle, the Australian start-up that hosts global data prediction competitions. The competition he won involved building a model from real world data that had been made openly available on Kaggle in order to optimise a predictive model. The aim of the model was to predict the rate at which HIV load would increase in patients from week to week given specific genetic markers in different patients.

Raimondi built his predictive model from data that Kaggle had made open on its site. He had no prior experience in bio-infomatics and no formal training in statistics, but became interested in data science running a small search engine optimisation firm he operated in Baltimore on the other side of the world to Kaggle's then headquarters in Melbourne. He taught himself data science using YouTube videos and open source data modelling tools.

Within a week and a half of the beginning of the competition, Raimondi's work was exceeding the predictive efficiency of the model that was the state of the academic art, and by the time the competition closed two months later he had taken the state of the art from 70% accuracy to 77%.

Second place went to a team of analysts at the Thomas J. Watson Research Centre at IBM.

Les enjeux autour du cycle de vie de la donnée

UNIVERSITY
MELBOURNE AUSTRALIA

Costs and Benefits of Data Provision

Report to the Australian National Data Service

By

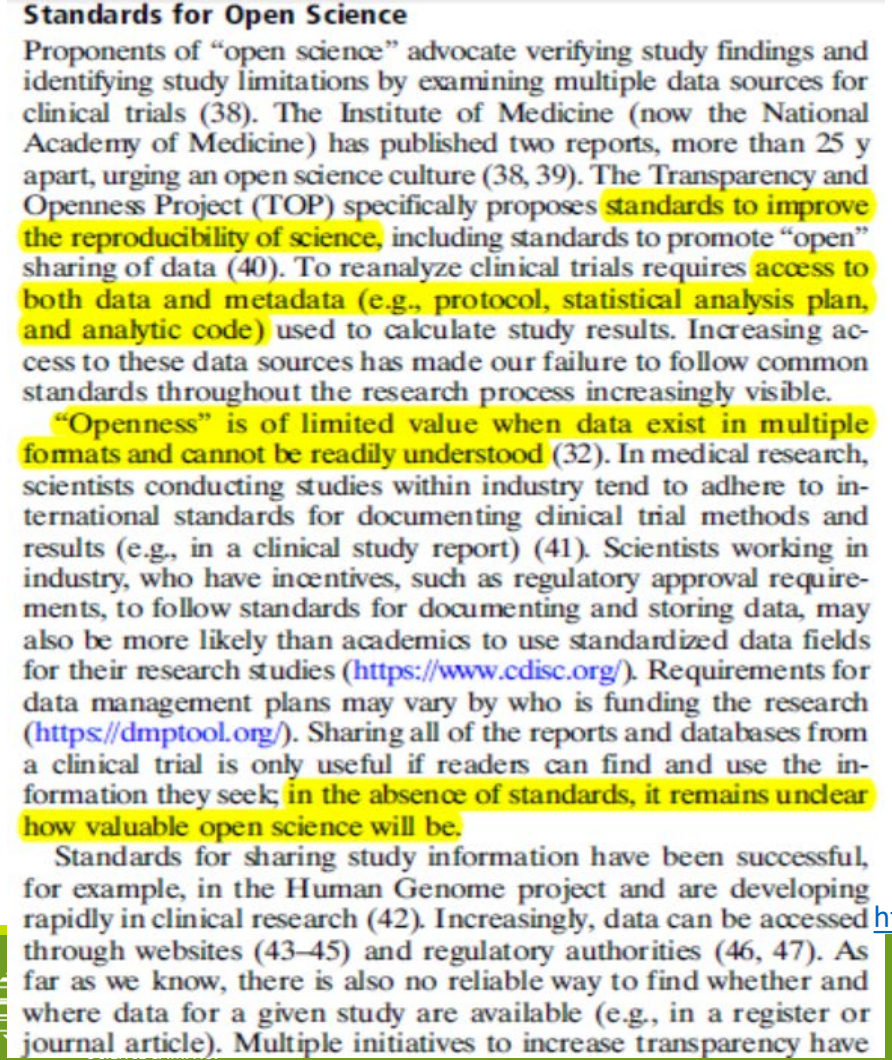
John Houghton

September 2011

Centre for Strategic Economic Studies
Victoria University

The BOM now publishes an annual *National Water Account* (NWA), which reports on the total water resource, the volume of water available for abstraction, the rights to abstract water, and the actual abstraction of water for economic, social, cultural and environmental purposes across Australia. The NWA provides information that has previously been difficult to access or unavailable to general users in a standardised form. It enables national comparability of information and highlights gaps and inconsistencies in data and knowledge, allowing improvements to be made. The information presented in the NWA strives to be: nationally consistent and comparable; publicly available; comprehensive; useful to governments, the water sector, industry and the general community; and transparent about the source and quality of the data.[55]

Extrait de
https://www.ands.org.au/__data/assets/pdf_file/0004/394285/houghton-cost-benefit-study.pdf



Standards for Open Science

Proponents of "open science" advocate verifying study findings and identifying study limitations by examining multiple data sources for clinical trials (38). The Institute of Medicine (now the National Academy of Medicine) has published two reports, more than 25 y apart, urging an open science culture (38, 39). The Transparency and Openness Project (TOP) specifically proposes standards to improve the reproducibility of science, including standards to promote "open" sharing of data (40). To reanalyze clinical trials requires access to both data and metadata (e.g., protocol, statistical analysis plan, and analytic code) used to calculate study results. Increasing access to these data sources has made our failure to follow common standards throughout the research process increasingly visible.

"Openness" is of limited value when data exist in multiple formats and cannot be readily understood (32). In medical research, scientists conducting studies within industry tend to adhere to international standards for documenting clinical trial methods and results (e.g., in a clinical study report) (41). Scientists working in industry, who have incentives, such as regulatory approval requirements, to follow standards for documenting and storing data, may also be more likely than academics to use standardized data fields for their research studies (https://www.cdisc.org/). Requirements for data management plans may vary by who is funding the research (https://dmptool.org/). Sharing all of the reports and databases from a clinical trial is only useful if readers can find and use the information they seek; in the absence of standards, it remains unclear how valuable open science will be.

Standards for sharing study information have been successful, for example, in the Human Genome project and are developing rapidly in clinical research (42). Increasingly, data can be accessed through websites (43–45) and regulatory authorities (46, 47). As far as we know, there is also no reliable way to find whether and where data for a given study are available (e.g., in a register or journal article). Multiple initiatives to increase transparency have

Pour que l'ouverture des données soit utile et valorisée, il faut

- Aussi ouvrir les métadonnées : protocoles, plans des analyses statistiques, description des instruments et des conditions de collecte, etc.
- S'appuyer sur des standards (métadonnées, formats, vocabulaires)

https://www.pnas.org/content/pnas/115/11/2590.full.pdf

FAIR : des principes pour partager utile

- Faciliter la découverte, l'accès, l'interopérabilité et la réutilisation des données
- S'applique aux données mais aussi aux protocoles, codes, workflow, etc.
- Il faut en tenir compte tout au long du cycle de vie de la donnée
  - Rédiger un PGD qui décrit la manière dont les données sont produites / obtenues, décrites, représentées, stockées, partagées

Mise en œuvre des principes FAIR

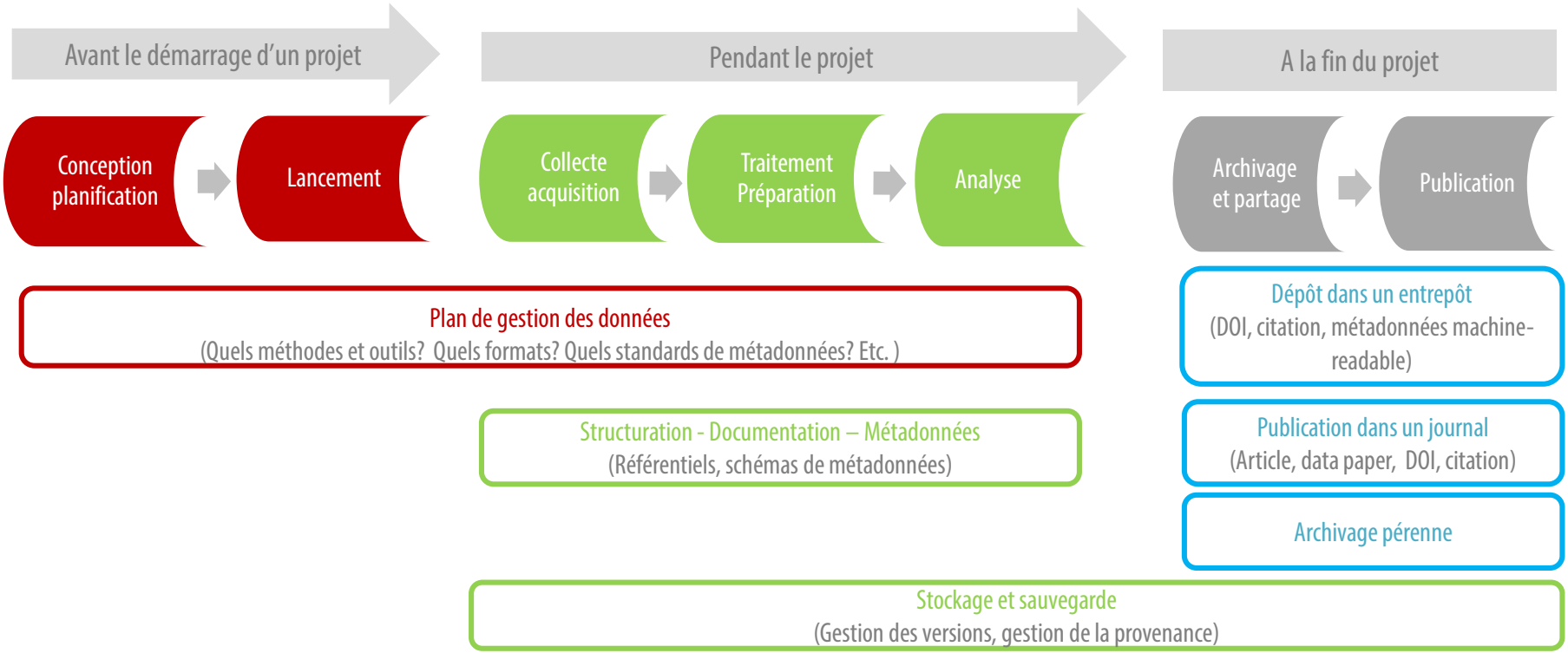
Des principes directeurs pour mettre à disposition des données

- Documentées de manière suffisante et standardisée
- Structurées et dans des formats ouverts et standards
- Identifiées et citables
- Des conditions d'accès et de réutilisation clairement explicites
- Dans une infrastructure qui favorise la préservation, la découverte et l'accès, par les humains et les machines

« FAIR data is not a platitude and is not a goal; it is a process»

- Commencer tôt, dès la conception du projet de recherche
- « FAIR » progressivement
  - Plusieurs outils d'évaluation existent pour évaluer le niveau de compatibilité avec les principes FAIR et avoir des indications sur ce qui peut être amélioré
- Être un acteur
  - Participer à la définition et aux évolution des standards
  - Contribuer à la traduction des principes FAIR dans votre domaine
- Ne pas rester isolé

FAIR BY DESIGN



En résumé

L’ouverture de la science contribue à la rendre plus transparente, reproductible, cumulative, économique, accessible aux citoyens

L’application des principes FAIR, tout au long du cycle de vie de la données, permet de d’accroître la valeur et le potentiel de réutilisation des données ouvertes

La « FAIRisation » des données est un processus d’amélioration continue

Merci pour votre écoute