Projet Open OPSE

Entrepôts de données : enjeux (I)

Données de la Recherche, en danger ?

Luc DECKER
IRD (service IST – MCST)
data@ird.fr

Démarche de préservation des données scientifiques

Sensibilisation aux risques et à leurs conséquences

Incitation (motivation) à davantage utiliser les services d'entrepôt de données à disposition

Données de la recherche

Rappel de définitions

Données primaires ou brutes : enregistrements factuels (chiffres, textes, images et sons), sources principales pour la recherche scientifique, reconnus comme nécessaires pour valider des résultats de recherche *[définition OCDE, 2007]*

Données dérivées : élaborées à partir de données primaires.

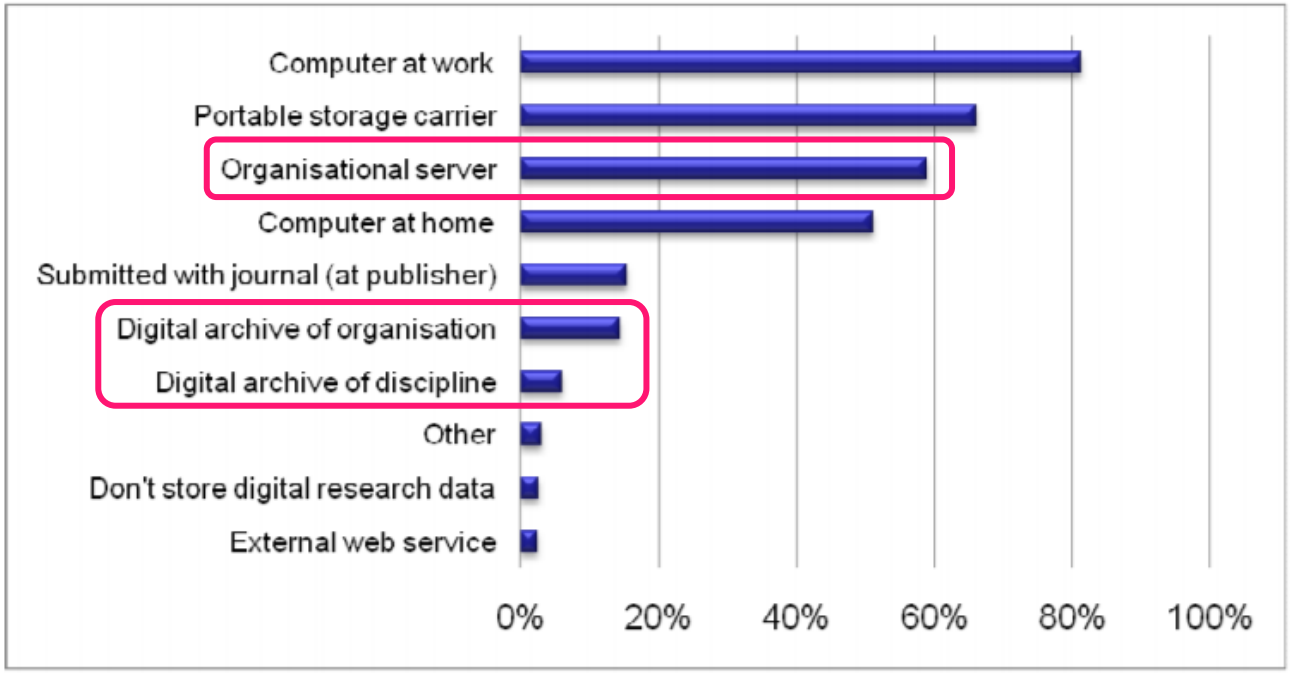
Données d'intérêt : réutilisables afin d'améliorer les connaissances par l'enrichissement, la combinaison à d'autres jeux de données.



Préservation des données scientifiques

Sondage : où les données sont-t-elles réellement stockées ?

Where do you currently store your research data ? (multiple answers possible)

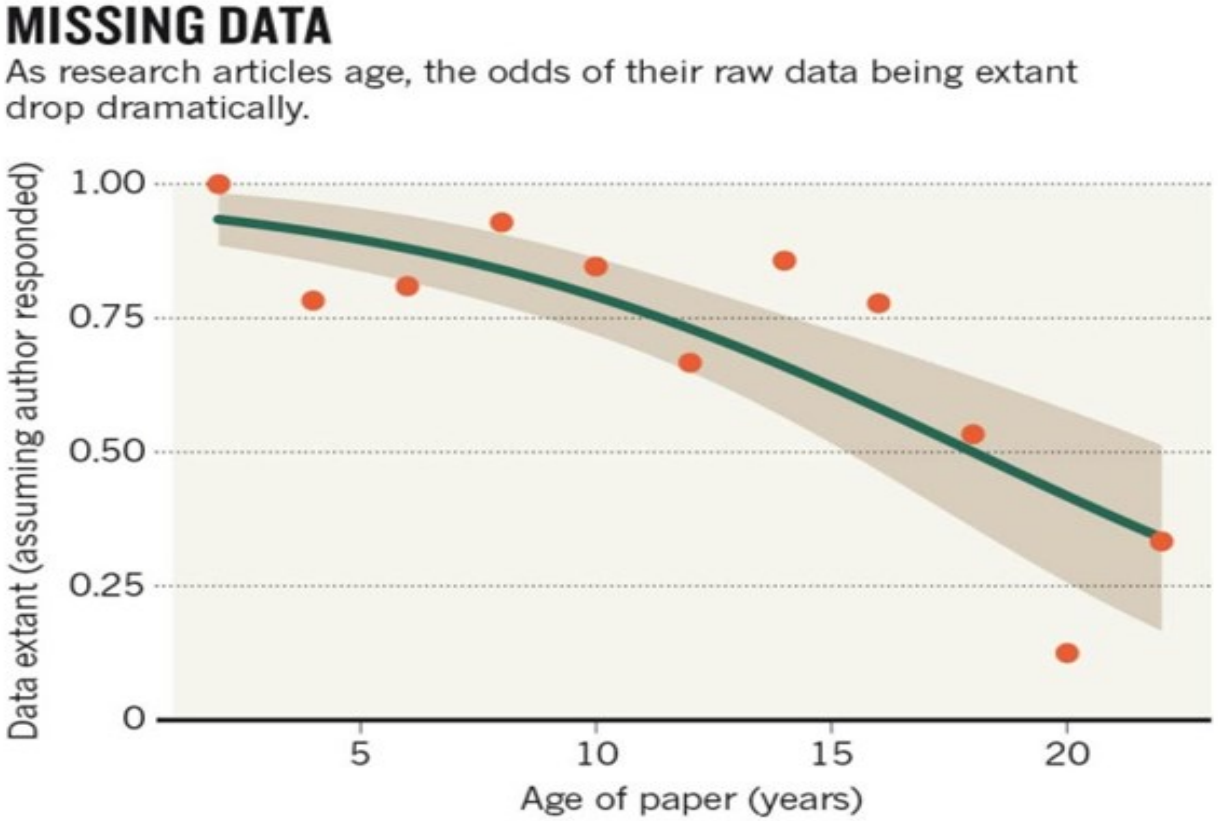


Graph 2: Source: PARSE.Insight[2] survey, held among researchers internationally, N = 1202 researchers

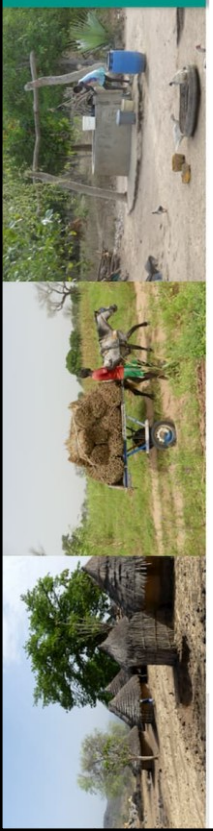
https://libereurope.eu/wp-content/uploads/PARSE-Insight_D3-5_InterimInsightReport_final.pdf

Préservation des données scientifiques

Etude : 20 ans après la publication d'articles, 80% des données brutes ont été perdues



VINES Timothy H., et al. The Availability of Research Data Declines Rapidly with Article Age, Current Biology, 2014.



Préservation des données scientifiques

Discussion

Dans votre expérience,

1) Avez-vous été témoin de pertes de données ?

2) Des données que vous auriez souhaité obtenir ont-t-elles été perdues ?

Objectif **Conserver les données** de la Recherche **sur le long terme**

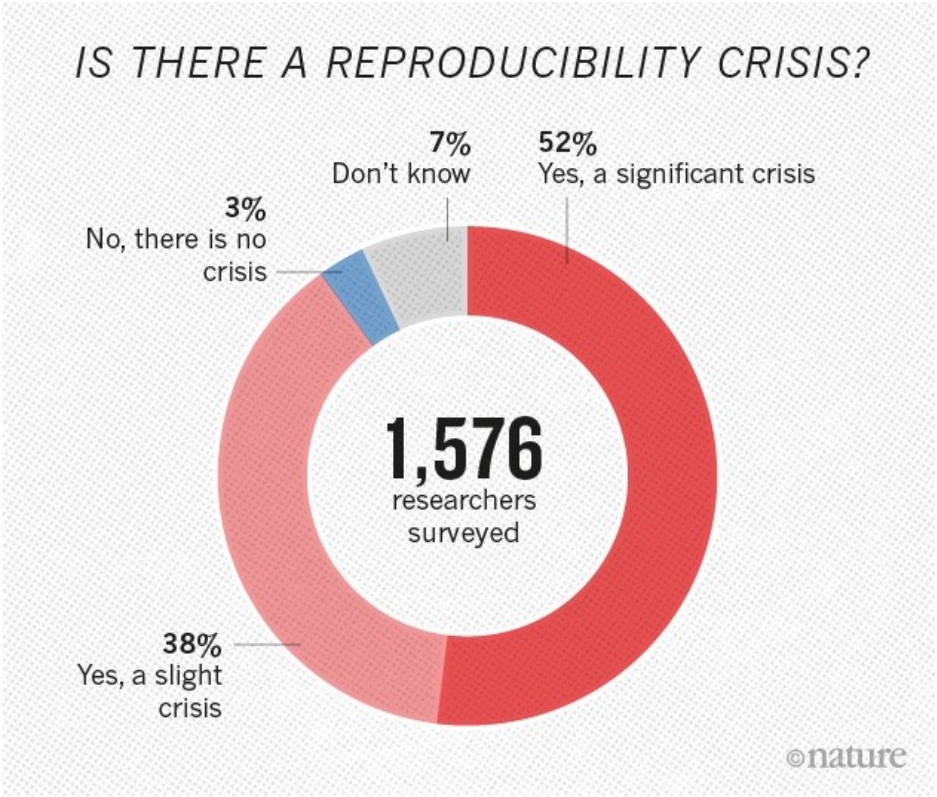
Données **perdues**
ou **inexploitables** :

- **Quels sont les risques ?**
- **Quelles conséquences ?**

Objectif **Conserver les données** de la Recherche **sur le long terme**

Reproductibilité des expériences

1500 chercheurs répondent à un sondage du journal *Nature*



Nature May 2016 : https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

« Plus de 70 % des chercheurs ont essayé et échoué à reproduire des expériences d'un autre scientifique,

et plus de la moitié n'ont pas réussi [au moins une fois] à reproduire leurs propres expériences »

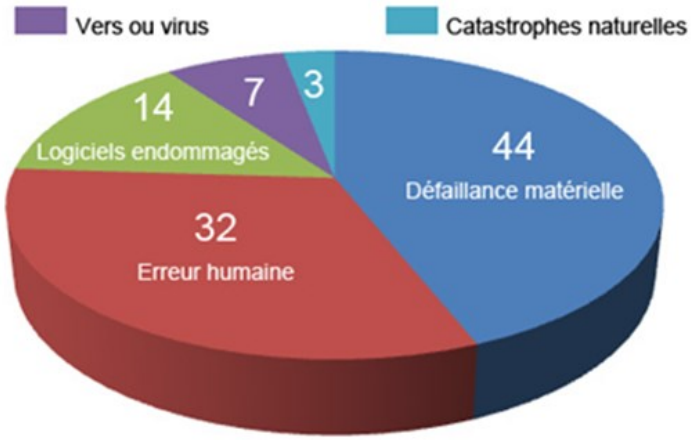
Quel lien entre reproductibilité et préservation des données ?

Préservation des données scientifiques

Causes possibles des pertes de données

(synthèse des risques)

- **Défaillance** des supports
- **Obsolescence** matérielle
- Obsolescence logicielle
- Virus informatiques
- **Destruction** des supports
- Lieu de stockage indéfini
- **Erreurs** humaines



Projet Open OPSE

Préservation des données scientifiques

Conséquences des pertes de données

(synthèse)

- Pertes de temps : travail à recommencer
- Pertes budgétaires, en particulier de fonds publics
- **Réduit les possibilités** de…
  - **vérifier les résultats**
  - **comparer des résultats dans le temps ou l'espace**
  - **réutiliser les données**, en particulier par d'autres scientifiques ou/et pour d'autres finalités