Plan de gestion des données (PGD)
Data Management Plan (DMP)

Jacques van Helden https://orcid.org/0000-0002-8799-8584
Frédéric de Lamotte https://orcid.org/0000-0003-4234-1172
Paulette Lieby https://orcid.org/0000-0002-9289-9652

Sciences de la vie : disruption numérique et explosion des données

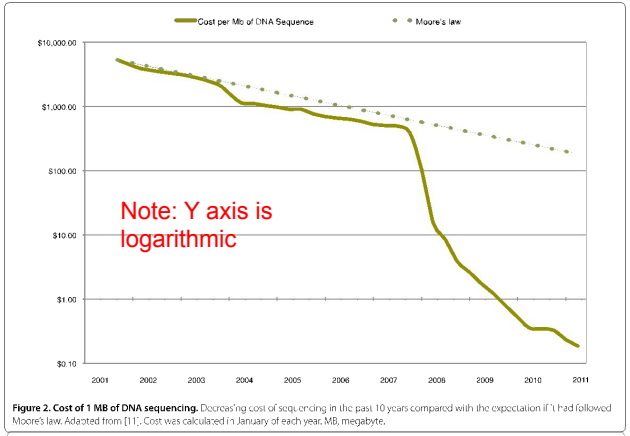


Figure 2. Cost of 1 MB of DNA sequencing. Decreasing cost of sequencing in the past 10 years compared with the expectation if it had followed Moore's law. Adapted from [11]. Cost was calculated in January of each year. MB, megabyte.

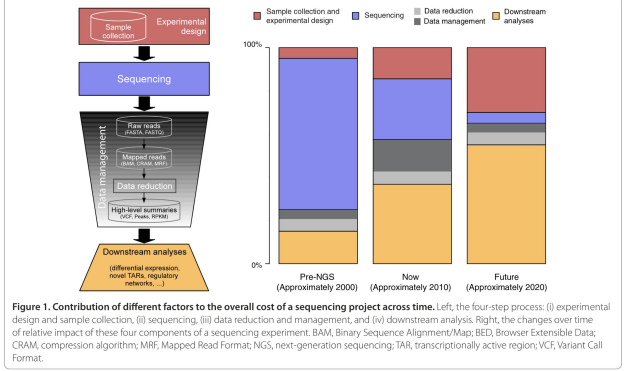


Figure 1. Contribution of different factors to the overall cost of a sequencing project across time. Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative impact of these four components of a sequencing experiment. BAM, Binary Sequence Alignment/Map; BED, Browser Extensible Data; CRAM, compression algorithm; MRF, Mapped Read Format; NGS, next-generation sequencing; TAR, transcriptionally active region; VCF, Variant Call Format.

The real cost of sequencing: higher than you think!
Sboner et al. (2011). Genome Biol 12: 125

Avant

1. Concevoir l'expérimentation
2. Collecter des résultats
3. Analyser des résultats

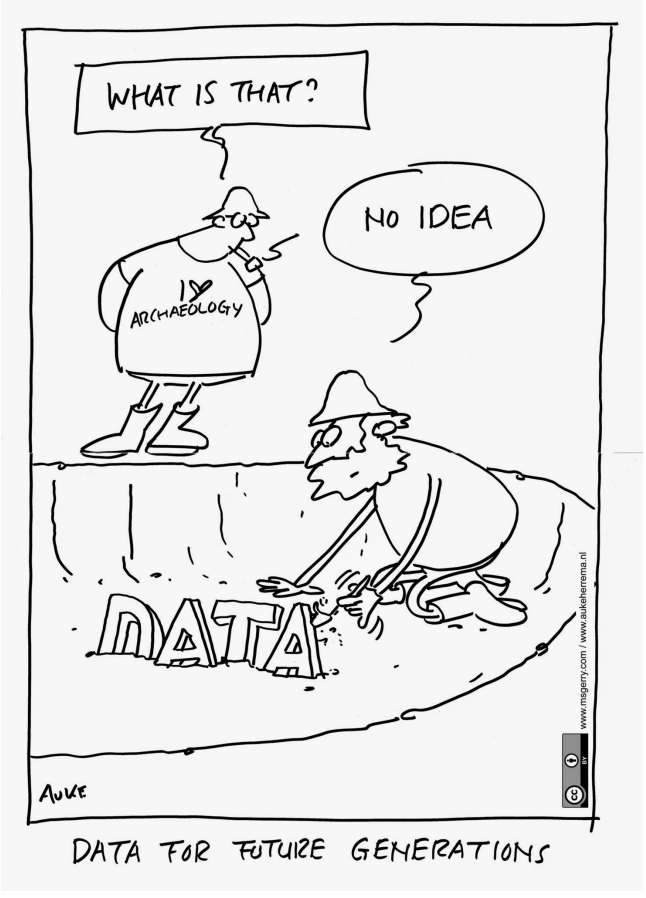
Un changement de paradigme

Maintenant

1. Générer massivement des données
2. Organiser (stocker, documenter, annoter)
3. Analyser (extraire de l'information)
4. Diffuser l'information

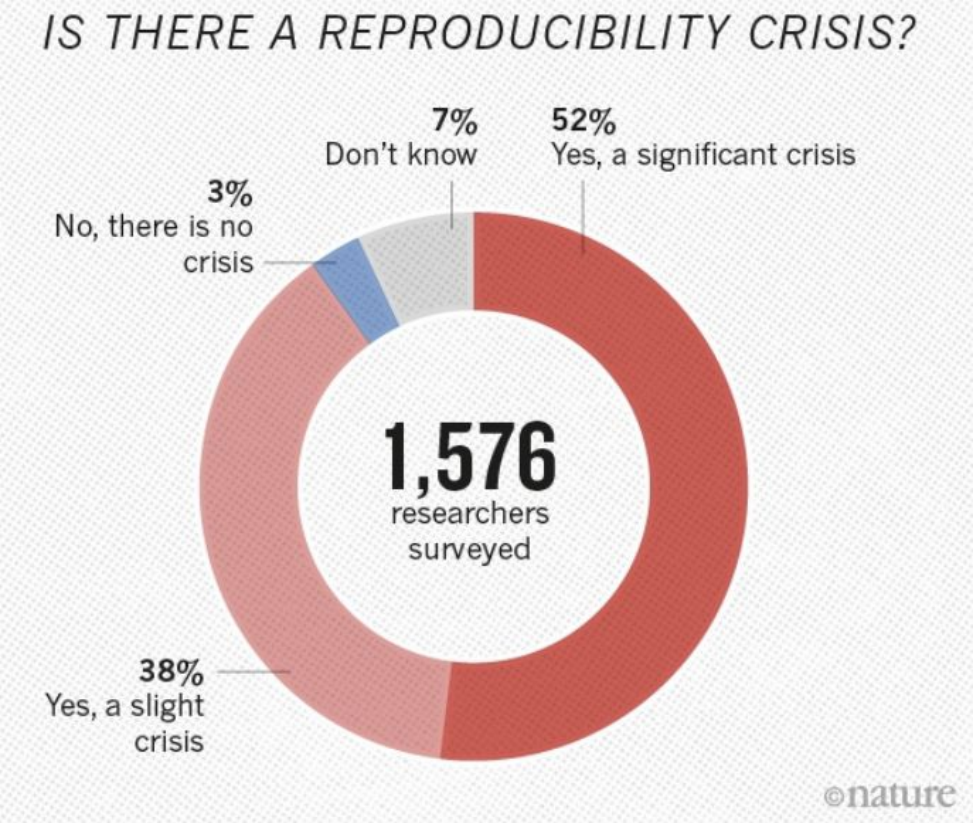
Une question culturelle

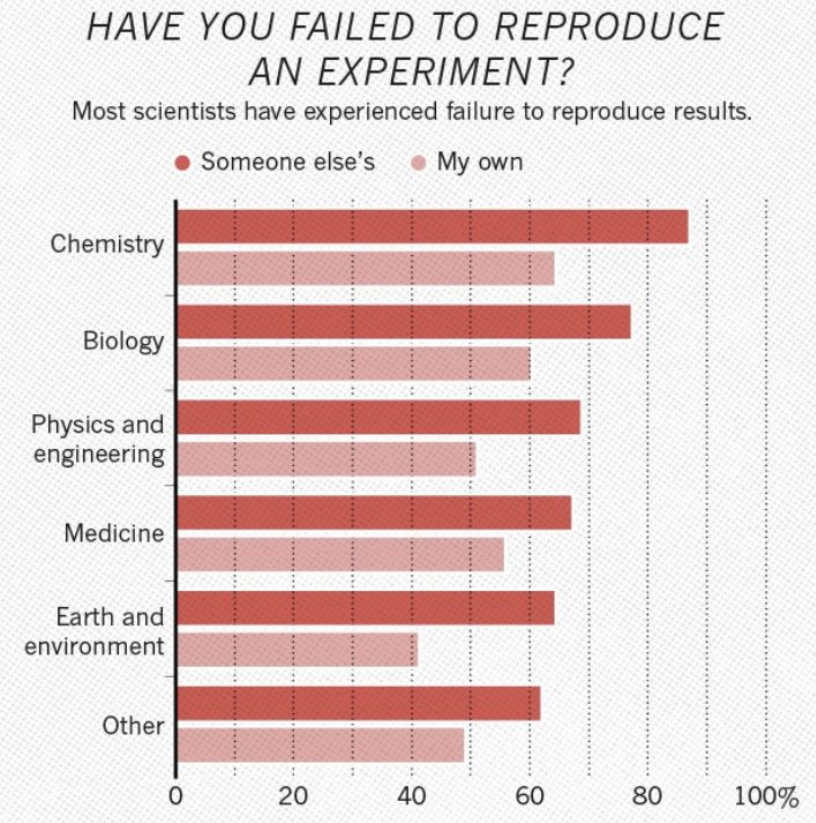
PUBLICATIONS AND DATA



DATA FOR FUTURE GENERATIONS

Les défis de la reproductibilité

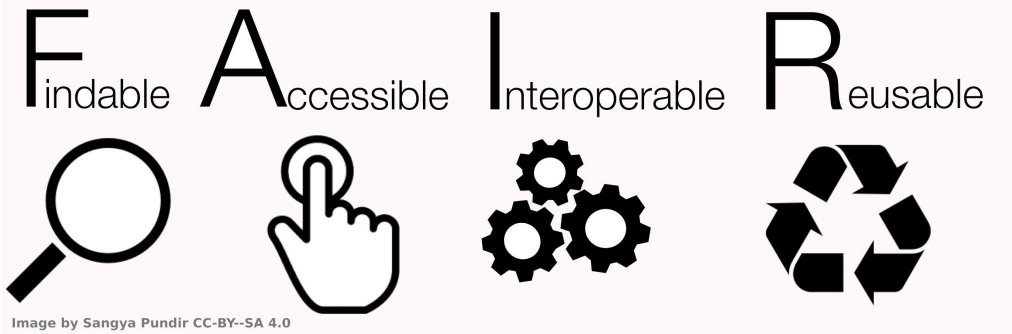




1,500 scientists lift the lid on reproducibility". Nature. 533: 452–454 - 2016

- définition «standard» : « *des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche* » (OCDE 2007)
- néanmoins, définitions moins restrictives :
  - données d'observation, données expérimentales, données computationnelles ou de simulation, données dérivées ou compilées, et données de référence (au delà du «factuel», INIST)
  - aussi, les données ne sont pas seulement des données «*nécessaires à la validation des résultats*» :
    - typiquement, en bioinformatique, les données produites sont plus nombreuses que celles strictement nécessaires pour valider un résultat
    - ces données gagnent en valeur précisément si elles peuvent être partagées

Gestion des données: les principes FAIR



Les principes FAIR Data sont un *ensemble de principes directeurs* visant à rendre les données trouvables, accessibles, interopérables et réutilisables.

Ces principes fournissent des orientations pour la gestion des données scientifiques et sont pertinents pour toutes les parties prenantes de l'écosystème numérique.

Ils s'adressent directement aux producteurs et aux éditeurs de données afin de promouvoir une utilisation maximale des données de recherche.

Vos données sont-elles FAIR ?

Faciles à (re)trouver

Accessibles

Interopérables

Réutilisables

Excellent guide sur les principes FAIR :

https://doranum.fr/enjeux-benefices/principes-fair/

Australian National Data Service (ANDS)

« aussi ouvert que possible ; aussi fermé que nécessaire »

Donc un PGD : plan de gestion des données (ou DMP : Data Management Plan)

- **Plan** : on planifie (donc on anticipe)
- **Gestion** : on gère, on fait fructifier (on commence déjà par ne plus perdre)
- **Données** : Data is the new oil, the new soil

→ Aide à la mise en place de **bonnes pratiques** de gestion du cycle de vie des données
→ **Contrat** de collaboration entre les acteurs du projet
→ Document **évolutif** au fur et à mesure de l'avancement du projet

➔ **PGD : Pour Générer du Dialogue (*Fred*)**

Les objectifs du PGD : décrire le cycle de vie des données

- Établir le rôle de chacun.e
  - Définir les responsabilités
- Assurer la reproductibilité des expériences
  - Décrire comment les données sont obtenues
- Permettre la réutilisation des données
  - Garantir la compréhension des données
- Éviter les pertes de données
  - Assurer un stockage adapté
- Respecter le droit et les personnes
  - Clarifier le cadre juridique et éthique
- Clarifier les droits de réutilisation
  - Spécifier les modalités de partage

Crédit : portail IRDdata

- Open Science
  - Plan national pour la Science Ouverte (Juillet 2018)
  - Adoption du modèle proposé par Science Europe et recommandé par le CoSO (Comité pour la Science Ouverte) (Sept 2019)
- En cours de généralisation sur les projets
  - PGD obligatoire pour les projets H2020 de l'Open Research Data Pilot (mais hors évaluation)
  - PGD obligatoire pour les projets financés ANR depuis 2019

Recommandation : les coordinateurs et partenaires de projets sont invités à utiliser l'outil de réaction DMP OPIDoR développé par l'INIST

https://www.ouvrirlascience.fr/wp-content/uploads/2019/08/Science-Europe-Guide-pratique-pour-une-harmonisation-de-gestion-données-recherche.pdf

Les données de la recherche sont des informations publiques :

- **elles sont soumises à un principe d'ouverture par défaut et de libre utilisation** *(Loi Lemaire - LPRN 2016)*
- **elles sont soumises à un principe de gratuité** *(Loi Valter 2015)*

« aussi ouvert que possible ; aussi fermé que nécessaire »

Principe : les données sont ‘ouvertes’ par défaut

Exceptions :

- La propriété intellectuelle appartenant à des tiers (droit d’auteur entre autre)
- La protection des données personnelles et de la vie privée (RGPD 2018)
  - Les secrets protégés
  - Les bases de données
  - ….

Pour en savoir plus : Ouverture des données de la recherche. Guide d'analyse du cadre juridique en France
Excellent logigramme : INRAe
Et un guide : Je publie, quels sont mes droits ? [flyer]
Aussi, ‘en construction’ : Site IFB de ressources 'Science Ouverte'

L'outil DMP OPIDoR

- Un outil de saisie et d'aide à la rédaction
- Déployé par l'Inist-CNRS pour l'Enseignement Supérieur et Recherche français
  - Gratuité pour l'ESR
  - Basé sur le code open source DMP Roadmap
  - Developpé par le Digital Curation Centre (DCC) et l'University of California Curation Center (UC3)
- Fonctionnalités
  - Basé sur la notion de **modèle (par Institut, par projet, ou autre)**
  - Partage avec **collaborateurs**
  - Gestion de la **visibilité**

https://dmp.opidor.fr

- Ouvrir un compte chez OPIDoR https://dmp.opidor.fr
- Créer un PGD pour votre projet tutoré
  - Partager votre PGD avec Hélène (helene.chiapello@inrae.fr), Fred (frederic.de-lamotte@inrae.fr), Paulette (paulette.lieby@france-bioinformatique.fr)
  - *Considérer les problèmes éventuels des données sensibles*
- Remplir le PGD au fur et à mesure de l'avancement de votre projet

Aide :

- Voir le canal slack 'dmp'

Conseils pour la rédaction des PGD

1. Commencer tôt
2. Réfléchir à la réutilisation
3. Vérifier les politiques en vigueur
4. Se faire aider
5. Penser large
6. S'inspirer du travail d'autrui
7. Être précis là où c'est nécessaire
8. Être concret
9. Oser avouer qu'on ne sait pas (encore)
10. Mettre à jour

Source : 10 Tips for Writing a Data Management Plan, Edugroepen, 2018

Le PGD et l'automatisation des flux de données : maDMP comme machine-actionable DMP

- Du PGD textuel (aujourd'hui) au PGD lisible et actionnable par une machine :
  - Permettre une mise en correspondance des métadonnées (qui décrivent les données) avec les ressources informatiques oú ces données résident
  - Induit un flux d'information automatisé

PGD  ressources informatiques

Connecter les ressources informatiques aux PGD

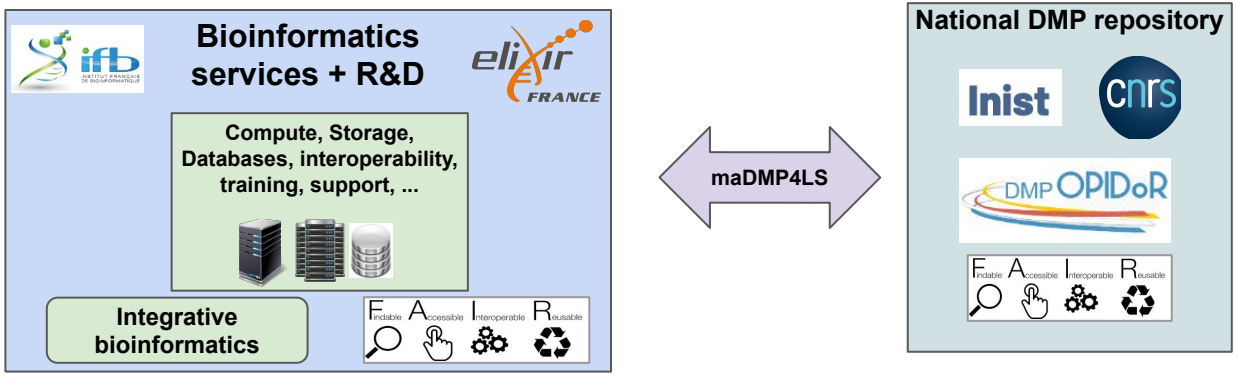
IFB National Network of Computing Resources (NNCR)

- **NNCR**
  - france-bioinformatique.fr/infrastructure
  - Serveurs HPC cluster + cloud
    - 21 052 CPUs
    - 9 552 To stockage « chaud »
    - 109 868 Go RAM
  - Répartis sur 8 villes
- **NNCR Task Force**
  - Ressources humaines mutualisées par les plateformes
  - Gestion des comptes et des espaces projets
  - DevOps
    - Intégration continue
    - Packaging
    - Virtualisation
    - Containerisation
  - Support communautaire
- **CesGO** (www.cesgo.org)
  - Espace collaboratif
  - Gestion des comptes et groupes (« my »)
  - Partage de données, projets, groupes de discussion

Ressources INIST pour le plan de gestion des données (PGD) = data management plan (DMP)

- **OPIDoR DMP** (dmp.opidor.fr/plans)
  - Dépôt de PGD
  - Modèles institutionnels de PGD
- **DORANum** (doranum.fr)
  - Des ressources pour accompagner la communauté scientifique dans la gestion et le partage des données

- Projet-flash ANR **machine-actionable DMP for Life Sciences** (**maDMP4LS**),
  - Collaboration IFB - INIST
  - allocation des ressources (espace-projet, volume, ressource calcul, accès collaborateurs) par accès programmatique aux DMP déposés sur OPIDoR (INIST)
  - mise à jour conjointe du DMP et des ressources au fil de l'évolution du projet
- **Gestion à chaque phase** des données et des métadonnées : production (collaboration avec autres infrastructures), analyse (serveurs IFB), conservation (centres de stockage), accès (dépôts nationaux et internationaux).
- Programmation des flux de données via le maDMP (extension du projet maDMP4LS) : des sites de production aux serveurs d'analyse, puis de ceux-ci vers les centres de stockage et vers les dépôts.



At the cross-road of data flows – from project conception to results delivery