GESTION DES DONNÉES EN BIOINFORMATIQUE

Valérie Cognat

ENVIRONNEMENT

Institut de Biologie Moléculaire des Plantes

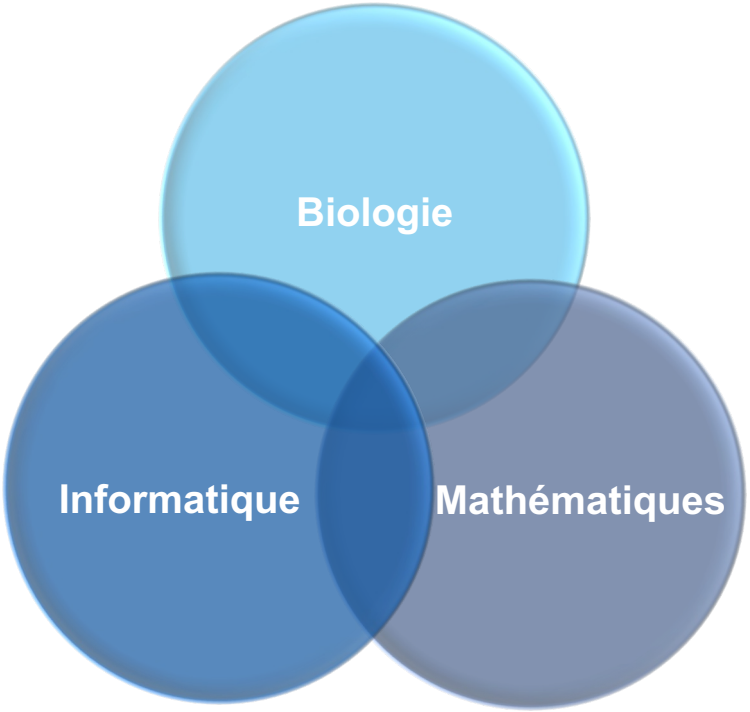
- **16 équipes de recherche**
- **5 plateformes technologiques**
  - Imagerie (PIC)
  - **Bioinformatique (BiP)**
  - Métabolomique (PIMS)
  - Séquençage (AEG)
  - Production protéines (P3P)
- **3 plateaux techniques**

ACTIVITÉS DE LA PLATEFORME

- Gestion de ressources de calcul et de stockage pour la bioinformatique
- Analyses de données biologiques
- Développement d'outils et de workflows
- Formation

1

ANALYSE DE DONNÉES : DOMAINES D'APPLICATIONS



Génomique
Transcriptomique
Protéomique
Réseaux de gènes
Réseaux métaboliques
Biologie structurale
Biologie de l'évolution
Analyse d'images

QUELS TYPES DE DONNÉES ? POUR QUOI FAIRE ?

- Essentiellement issues du séquençage haut débit
- Assemblage de génomes (plantes, bactéries)
- Identifier des gènes différentiellement exprimés dans différentes conditions (luminosité, stress hydrique, mutant, infection par des virus ou bactéries, …)
- Identifier des populations bactériennes dans les microbiotes de plantes
- …

ANALYSE DE DONNÉES

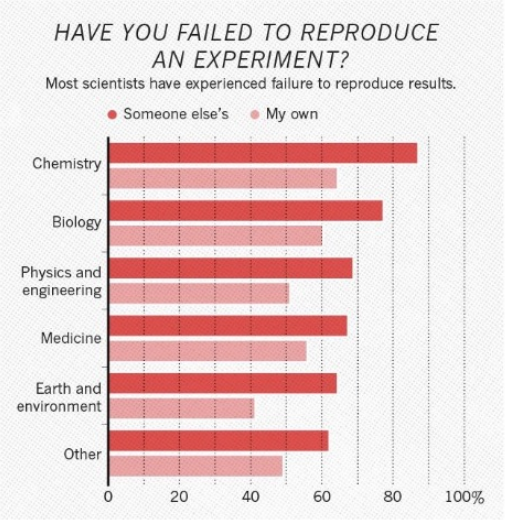
- ~ 25 projets par an – 16 équipes – publications 1 à 3 ans après l'analyse
- 3 bioinformaticiens
- Données issues d'appareils haut débit (séquenceurs, métabolomiques) : de quelques dizaines de Go à 2-3 To par projet – plusieurs échantillons par expérience
- Développement de nouveaux workflows d'analyses et d'outils de visualisation
- Utilisation d'outils Open Source
- Analyses de données publiques
- Reprise de protocoles d'analyses publiés

DIFFICULTÉS RENCONTRÉES

- **Problème d'accès aux données**
- **Problèmes d'accès aux outils**
  - N'existe plus, ne sont pas maintenus
  - Difficulté d'installation
- **Problèmes de paramétrages des analyses**
  - Version des outils
  - Paramètres
- **Problème d'accès aux ressources de calcul et stockage**
- **Impossibilité de reproduire des résultats**

DIFFICULTÉS RENCONTRÉES

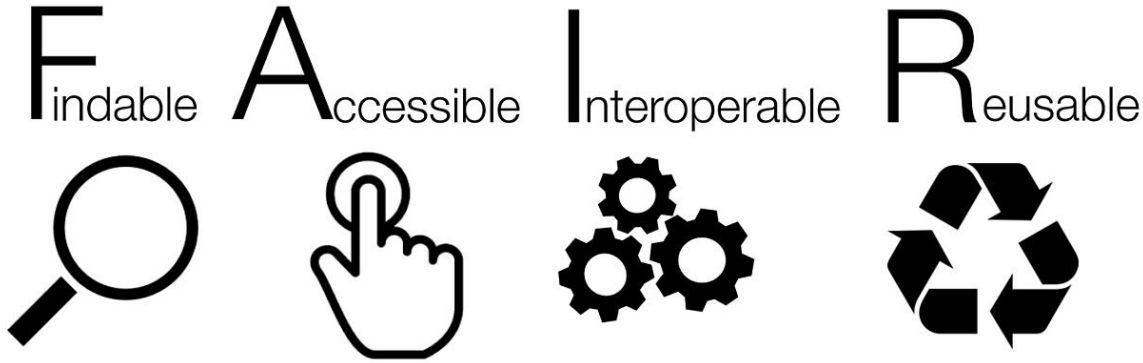
- ❑ Problème d'accès aux données
- ❑ Problèmes d'accès aux outils
  - N'existe plus, ne sont pas maintenus
  - Difficulté d'installation
- ❑ Problèmes de paramétrages des analyses
  - Version des outils
  - Paramètres
- ❑ Problème d'accès aux ressources de calcul et stockage
- ➢ **Impossibilité de reproduire des résultats**



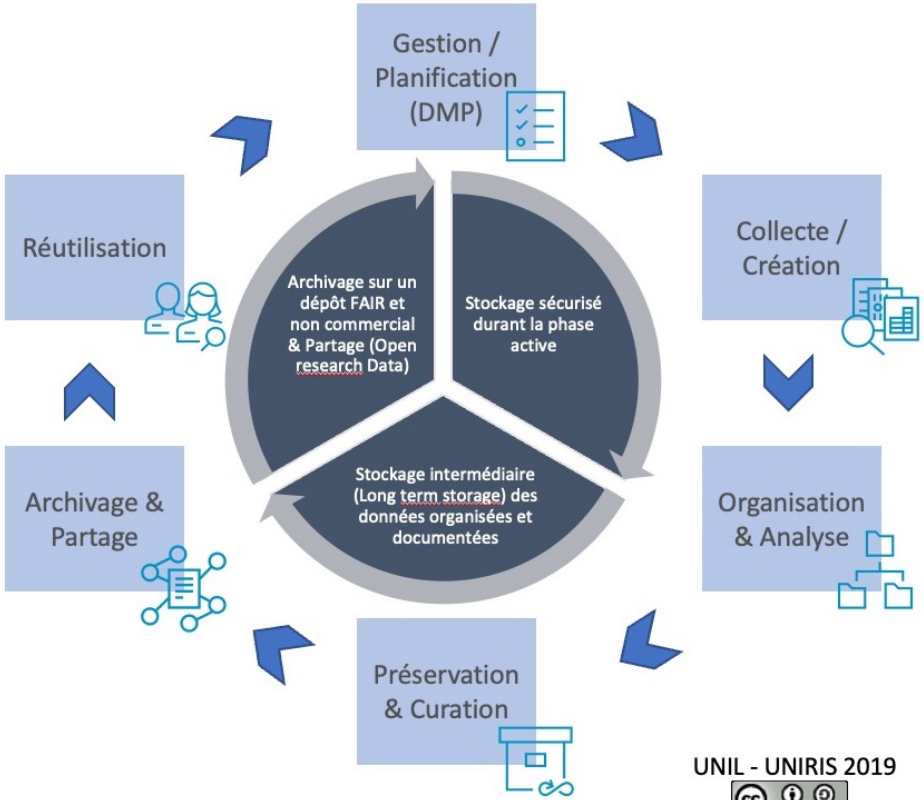
Monya Baker, 2016
https://doi.org/10.1038/533452a

76 % des chercheurs ont échoué à reproduire un résultats

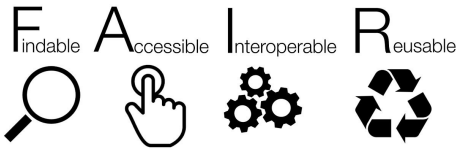
ÊTRE FAIR EN BIOINFORMATIQUE



RESPONSABILITE DANS LA GESTION DES DONNÉES



MISE EN ŒUVRE AU SEIN DE LA PLATEFORME



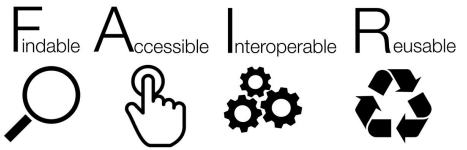
Organiser

- Espace de stockage dédié et partagé pour les projets
- Architecture des projets définie – convention de nommage
- Documentation

Stocker

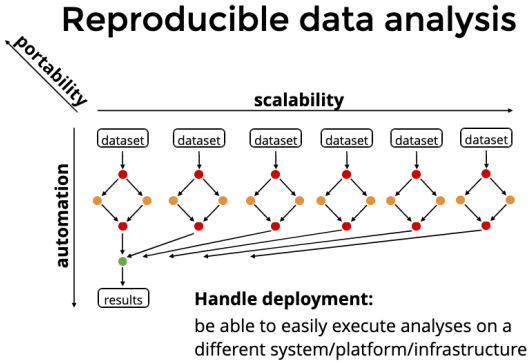
- stockage sécurisé et redondant des projets au niveau de la plateforme
- Stockage / versionning des codes sources informatiques dans un GitLab interne

MISE EN ŒUVRE AU SEIN DE LA PLATEFORME

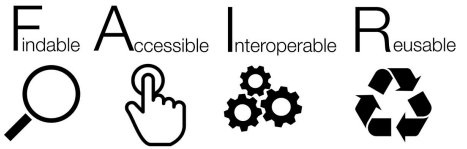


Reproductibilité et portabilité des analyses

- Gestionnaire de workflows (Snakemake, Nextflow)
- Encapsulation les environnements et les versions des outils utilisés (conda)
- Utilisation de format de données standard et des outils Open Source
- Créer des containers (docker) pour les projets plus complexes



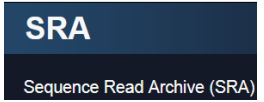
MISE EN ŒUVRE AU SEIN DE LA PLATEFORME



❑ Partage des données et codes sources pour publication

- Aide au dépôt des jeux de données dans les entrepôts dédiés [ENA / SRA] (doi) – Renseigner les métadonnées des analyses
- Dépôt de codes sources (workflows, scripts) dans GitHub

  ➢ Moissonnage dans Software Heritage
- Licence d'utilisation



EST-ON FAIR ?

- **En interne : oui**
  - Bonnes pratiques pour le bon fonctionnement
- **Open Science :**
  - La plateforme n'est pas propriétaire des données => Chercheurs
  - Protocole détaillé des analyses pour publications / dépôt des codes sources si nécessaire
- **Point d'amélioration**
  - Métadonnées plus riches

2

FORMATIONS EN SCIENCES OUVERTES / PRINCIPES FAIR

- **Sciences Ouvertes et DMP : 2 jours**
  - 1 session (2022)
  - 12 stagiaires – 7 formateurs
- **Les principes FAIR dans les projets de bioinformatique : 3 jours**
  - 9 au 11 avril 2024
  - URFIST
  - 14 stagiaires – 3 formateurs + 2 helpers

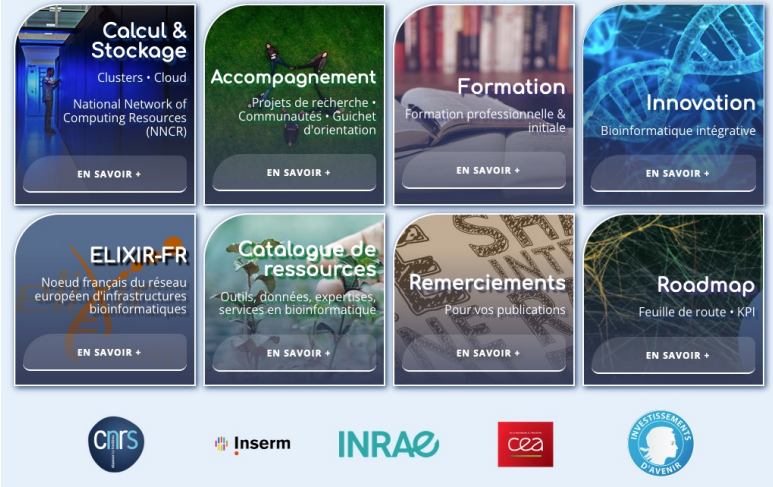
En lien avec l'Institut Français de Bioinformatique

INSTITUT FRANÇAIS DE BIOINFORMATIQUE

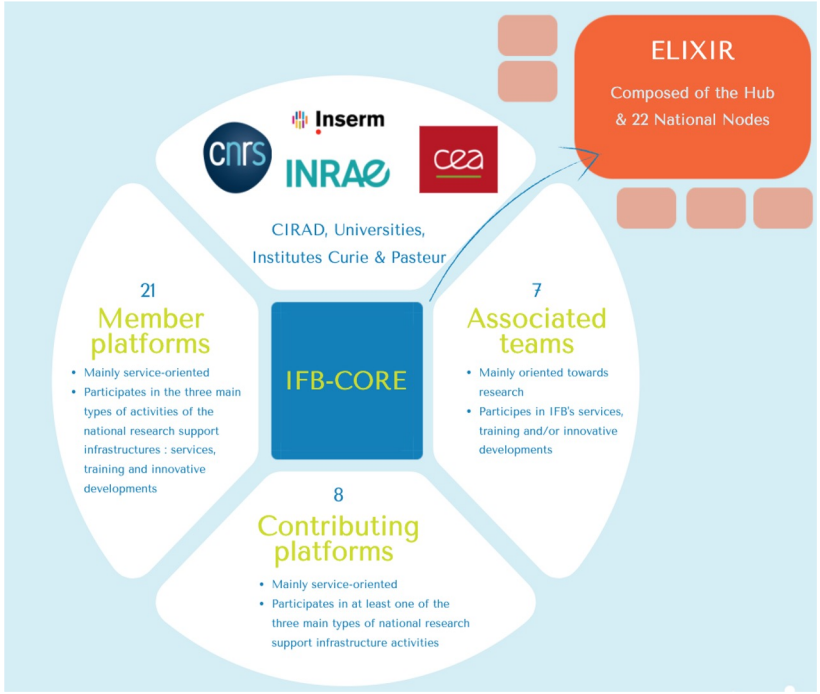
Infrastructure Nationale de Support à la Recherche créée dans le cadre de l'appel à propositions **« Infrastructure Nationale en Biologie et Santé »** du Plan Investissements d'Avenir.

Son rôle :

- assurer un support
- déployer des services
- organiser des formations
- réaliser des développements innovants pour les communautés des sciences du vivant



UNE FEDERATION DE PLATEFORMES



MISSIONS ET ACTIONS DE L'IFB

Infrastructure numérique, NNCR

Développement logiciel

Orientation et Accompagnement des usagers

Mutualised Task Forces

Bases de connaissances

Prospective & innovation

Formation

Science ouverte & Interopérabilité

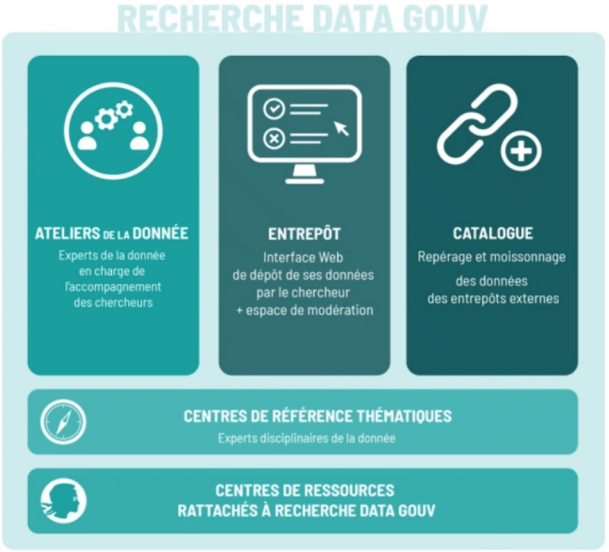
SCIENCES OUVERTES ET INTEROPÉRABILITÉ

Centre de Référence Thématique des données en Biologie-Santé

- Juillet 2022
- Ateliers de la donnée
- Interactions avec le “**Réseau des INBS**”

Liens avec la communauté internationale

- Elixir Interop Platform, CONVERGE
- Bioschemas, EDAM
- Elixir RDMkit et RDM Community
- RDA, EOSC, etc.



RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

recherche.data.gouv.fr

SCIENCES OUVERTES ET INTEROPÉRABILITÉ

❑ Développement d'outils

- FAIR-checker : évaluation FAIR de ressources web (*J Biomed Semantics, 2023*)
- Metark : courtage de données pour l'ENA
- OpenLink : Interopérabilité PGD, CLE, entrepôts de données et publication avec transfert automatique de métadonnées
- PGD modulaires, PGD multi-omiques, PGD entités
- Interopérabilité entre DSW et DMP-Opidor
- maDMP : machine actionable DMP en collaboration avec DMP-Opidor

SCIENCES OUVERTES ET INTEROPÉRABILITÉ

❑ Formation

➢ Science Ouverte et PGD

- 2 sessions en distanciels en 2020-2021
- Sessions en présentiel : Paris – Strasbourg en 2022
- Session dédiée aux plateformes IBISA en 2023
- Session dédiée aux données de phénotypages des plantes en 2023

➢ Les principes FAIR en Bioinformatique

- 4 sessions en présentiel à Paris depuis 2020
- Sessions en région : Nantes, Clermont, Jouy-en-josas, Montpellier, Strasbourg

SCIENCES OUVERTES ET INTEROPÉRABILITÉ

❑ Formation

➢ FAIR-ification du matériel pédagogique (Elixir)

- Métadonnées (bioschemas)
- Identifiants pérennes (orcid, doi)
- Dépôt des formations sur le catalogue de l'IFB, TESS, Goblet
- Dépôts du matériel pédagogique : Zenodo, GitHub, GitLab
- Formats interopérables
- Licence

https://elixir-europe-training.github.io/ELIXIR-TrP-FAIR-training-handbook/

LIENS

https://www.france-bioinformatique.fr/

https://ifb-elixirfr.github.io/IFB-FAIR-data-training/

https://ifb-elixirfr.github.io/IFB-FAIR-bioinfo-training/