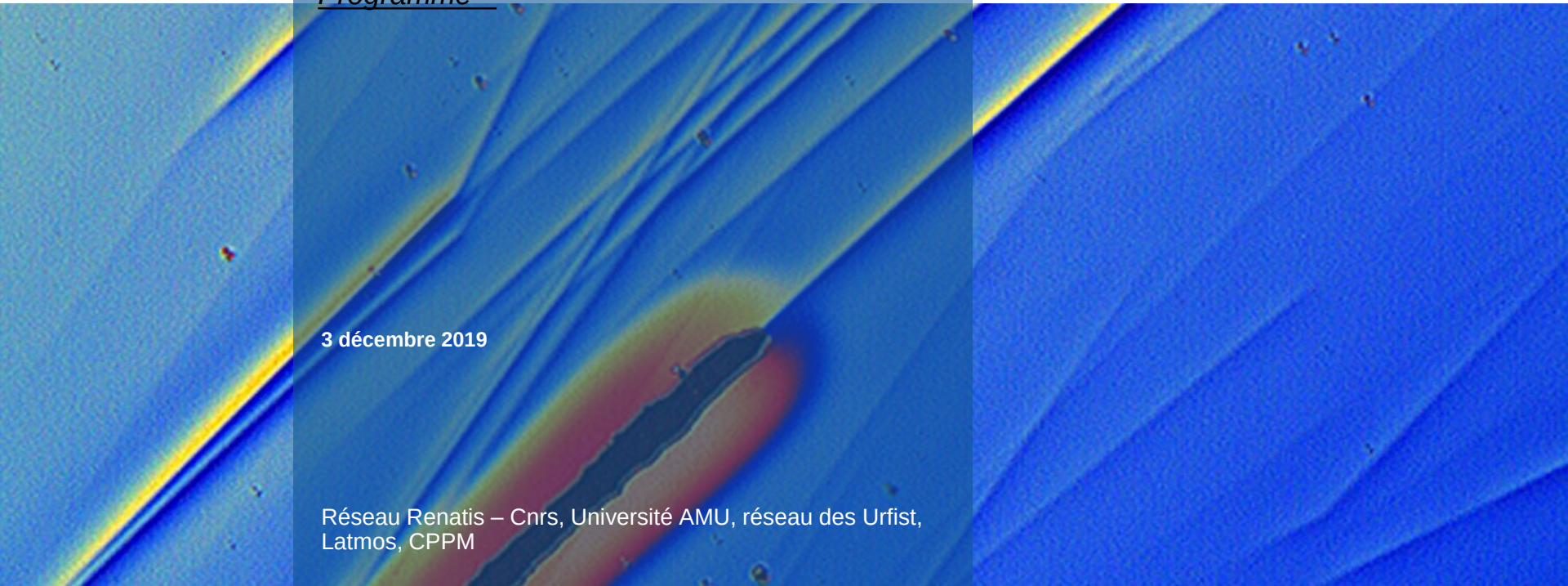
FaiR ou Fai-RR
Réutilisation versus Recherche Reproductible ?

8e atelier DIALOGU'IST

Programme ↗



3 décembre 2019

Réseau Renatis – Cnrs, Université AMU, réseau des Urfist, Latmos, CPPM

La poursuite des réflexions sur la gestion des données de recherche

En 2018,

- Un atelier sur la Fairisation* des données de recherches
- La création du Comité de la Science Ouverte

En 2019,

- Des feuilles de route Science Ouverte d'établissements de recherche
- Une sensibilisation des chercheurs à la nécessité de gérer leurs données en forte progression (DMP, …)

14 centres

De visio conférence répartis sur tout le territoire

80 participants

Des professionnels de l'IST
Des chercheurs
Des doctorants

6 interventions

Répartis en 2 thématiques
- Le contexte et les enjeux
- Les compétences et les apprentissages

7 intervenants

- Chercheurs et Ingénieurs expérimentés
- Professionnels de l'IST

Rendre les données

Trouvables, accessibles, interopérables et Réutilisables
Findable, Accessible, Interoperable and Re-usable

Programme

Partie 1 : le contexte et les enjeux

- 13h30 – 13h45 : Introduction de l'atelier et retour sur l'engagement de l'Aix-Marseille Université (AMU) par Fabien Borget, Aix-Marseille Université, Enseignant-Chercheur
- 13h45 – 14h30 : Contexte et définitions de la Recherche Reproductible par Sabrina Granger, Urfist Bordeaux, Conservateur des bibliothèques, co-responsable, référente « Information scientifique et technique »
- 14h30 – 15h00 : Le R des données FAIR, un atout en premier lieu pour les chercheurs : retour d'expérience sur les aspects à la fois techniques et psychologiques de la mise à disposition des données de la recherche par Daniel Jacob, INRA Bordeaux, Plateforme Métabolome, Ingénieur de Recherche

Partie 2 : quelles compétences et quels apprentissages ?

- 15h15 – 15h45 : Retour d'expérience d'une communauté, le GBIF (Global Biodiversity Information Facility, gbif.fr) par Anne-Sophie Archambeau, responsable du point nodal France, Muséum National d'Histoire Naturelle
- 15h45 – 16h00 : Comment enseigner et sensibiliser les étudiants à la gestion et réutilisation des données – Retour d'expérience par Véronique Ginouvès, AMU-CNRS, UMR 3125, Ingénieure de Recherche
- 16h00 – 16h30 : Mooc « Recherche reproductible : principes méthodologiques pour une science transparente », Retour par Laurence Farhi, Inria, Ingénieure pédagogique pour le Learning Lab et par Marie-Hélène Comte, documentaliste, Inria, Sophia-Antipolis

16h30 : Clôture de l'atelier par Fabien Borget

Un contexte politique nécessaire – La RÉVOLUTION

Une problématique

- Plus de 10 fois de plus de fraudes dans les publications scientifiques en 10 ans
- Crise de la reproductibilité de la recherche scientifique
- Crise de l' « autorat » (Publish or Perish)
- Impact de la Science ouverte dans les laboratoires (transparence, mise à disposition et accessibilité de la Recherche, collaborations scientifiques)

Des réponses ?

- Mettre en place de politique nationale (CoSO) ou d'établissements ; créer des structures de la donnée, l'exemple de l'AMU (Aix-Marseille Université)
- Rendre la recherche reproductible et les données réutilisables, mais avec respect de la loi (République Numérique, RGDP…)
- Rendre les données FAIRe-usable, les stocker, mais quelles données, quel archivage pérenne ?
- Implémenter et/ou respecter de nouveaux formats

Qu’est-ce que la Recherche Reproductible ?

Un contexte : le paradoxe de la Science Ouverte

- Premières mentions de la reproductibilité dans les années 90 avec l’Open Access
- Le numérique prend de plus en plus de place et donc porte un regain d’intérêt pour la recherche reproductible
- Ouvrir les données ne suffit pas pour aboutir à une recherche reproductible
- Des problèmes de méthodes, notamment dans le recours aux statistiques, plus que des questions de fraude
- Des outils qui évoluent, voire sont perdus
- Des inégalités des chercheurs face au numérique : il faut pouvoir s’orienter dans un contexte complexe
- Une survalorisation des résultats positifs en raison des critères d’évaluation et du cadre éditorial. Quelle est la place des résultats négatifs ?
- Quelles différences entre reproductibilité et réplicabilité/répétabilité ?
- Qu’est-ce qui doit être reproductible ? Les résultats? les méthodes? Les données?

Nécessité d’un changement culturel et d’une révolution des compétences pour les chercheurs

Quels changements pour une recherche Reproductible ? Devenir autonome ? Se former, former les doctorants …

La mise en place et la mise en œuvre de politique d'établissements

(ex de l'université de Londres)

- Favoriser la transparence
- Sensibiliser au R de FAIR
- Faciliter les dépôts de données, les annotations, …
- Se donner un seuil d'acceptation de la reproductibilité (p-value)

Un retour aux fondamentaux et de nouveaux soft skills pour les chercheurs

- Compréhension des statistiques
- Meilleure agilité
- Conception de « Design » d'études
- Langages, formats et outils Python, R, Jupyter note books …
- Évolution des pratiques éditoriales : structuration pour des études de réplication
- Identification des données à collecter
- (Ré)Appropriation du cycle de vie des données via les PGD/DMP
- Gestion des données par la qualité

Quels changements pour une recherche Reproductible ? Devenir autonome ? … mais pas que … COLLABORER …

De nouvelles collaborations pour un meilleur accompagnement

- des chercheurs avec les professionnels de l'IST, les informaticiens, les statisticiens, les datascientists

Quel niveau d'expertise ? Et pour quels acteurs ? Les professionnels de l'IST doivent-ils devenir des programmeurs ?

- Prendre en considération les 4 niveaux « novice », « initié », « apprenti » et « expert »
- Maîtriser le contexte de production des données, les méthodes à appliquer en prérequis
- Être capable d'avoir un point de vue critique

Un exemple : l'INRA(E)

- Existence de data centers (payants)
- Utilisation des solutions EOSC gratuites
- Une prise de conscience en se posant les bonnes questions sur le sens à données sur les notions de Science Ouverte, de reproductibilité de transparence, de structuration, de stockage, d'évaluation (y compris des thésards) …
- Formation FAIR le plus tôt possible (dès le Master)
- **Prendre le temps nécessaire et rassurer**

Un retour d'expérience FAIR (encore à améliorer) au GBIF (Global Biodiversity Information Facility)

Global Biodiversity Information Facility
Système Mondial d'Information sur la Biodiversité
» Accès libre et gratuit aux données de la biodiversité
www.gbif.fr

Programme intergouvernemental lancé par l'OCDE qui permet à près de 60 pays participants et de 40 organisations associées de récolter des données brutes en biodiversité et d'y donner un accès libre et ouvert

- Données d'observation, des séquences génétiques, des référentiels taxonomiques, issues de machine (télédétection, drone, TDM …) …

Utiles pour divers domaines d'application

- Niches écologiques; agrobiodiversité, santé, suivi d'espèces invasives, médecine légale, changement climatique, conservation…

Pas de PGD/DMP

- Mais un outil d'auto-évaluation. Se base sur le cycle de vie
- Obligation d'appliquer une licence CC-0 ou CC-BY ou CC-BY-NC au choix du producteur
  ➔ facilite le R de FAIR

Une standardisation ➔ facilite le I de FAIR

- Sur les métadonnées (EML), sur les données (DarwinCore, ABCD)
- Un DOI sur chaque jeu, chaque téléchargement et citation et à chaque datapaper
- Des mise à disposition par api

Des ressources FAIR (goFAIR initiative)

Plus d'un milliard de données téléchargeables (dont données sensibles floutées)

Plus de 51 000 jeux de données issus de plus de 1 500 institutions

80 milliards de données téléchargées par mois (2018-2019)

Des nœuds nationaux

France : plus de 65 millions de données

Reconnu par le CoreTrust Seal comme entrepôt de données fiables

Entre-aide entre pays

Une veille sur la réutilisation au Danemark➔ Revue Science Review

Sensibiliser et former
Les étudiants, les doctorants, les chercheurs, tout le monde …

Exemple du Master « métiers des archives et des bibliothèques – humanités numériques » à l'AMU

- Implémentation d'un tronc commun pour le système d'information, la structure des données, les réusages et les aspects juridiques
- Appropriation par les étudiants : voir les restitutions des étudiants via quelques billets, par exemple sur les données de recherche et journée thématique (obligatoire) – Visual Studies

Exemple du DU « Scientific Data Management (Gestion des données scientifiques) » à Montpellier

- Opérationnel depuis janvier 2020
- 3 axes : analyse, sécurité et ouverture des données

Sensibiliser et former
Les doctorants, post-doc, chercheurs, ingénieurs, masters, tout le monde

Le mooc « Recherche reproductible: principes méthodologiques pour une science transparente ↗ » proposé par l'INRIA disponible sur la plateforme FUN

- 2 sessions en 2018 et 2019; 3e à venir (mars 2020)
- Pris en considération dans les plans de formation des établissements
- Accessible à tous domaines
- Comporte 5 modules, "La reproductibilité, en crise ? Reproductibilité et transparence", "Cahier de notes, cahier de laboratoire", "Le document computationnel", "L'analyse réplicable", "La réalité du terrain"

A venir fin 2020 un nouveau Mooc plus technique sur les outils de gestion des environnements logiciel, l'automatisation avec un workflow, les techniques de gestion des données de recherche

Voir le support ici …

Attestation de suivi à 50%

ATTENTION : notion sur Python et R requise dès le Module 3

Intégré à certaines écoles doctorales

FAIR ET FAI-RR : pas d'opposition mais un principe de fonctionnement méthodologique

- Les données doivent être publiables, y compris les données « invisibles » (non publiées) qui représentent 90% des données produites par les chercheurs
- Il faut faire attention à la récupération « marketing » et à la novlangue
- Il peut exister des données de qualité sans qu'elles soient estampiller « FAIR » (exemple des cahiers de laboratoire)
- Il faut se donner le temps nécessaire pour bien partager les données, l'agenda du chercheur doit en être le moteur
- Il est nécessaire de structurer les données pour les rendre interopérables et développer les outils permettant la garantie FAIR
- Il faut former les chercheurs le plus tôt possible au cycle de vie des données et aux outils numériques (participer à l'Opidor Tour)

Le « R » de FAIR rend le chercheur 2.0

Pour en savoir plus *(cliquez sur la flèche)* ...

Les supports↗

- Introduction et conclusion de Fabien Borget,
- Des interventions de Sabrina Granger, Laurence Farhi,
- Illustration de l'intervention de Véronique Ginouvès

Le livre↗

- Vers une recherche reproductible ; Faire évoluer ses pratiques
  Loïc Desquilbet, Sabrina Granger, Boris Hejblum, Arnaud Legrand, Pascal Pernot, Nicolas Rougier
  Facilitatrice : Elisa de Castro Guerra
  2019-07-23

Le MOOC↗

- Recherche reproductible : principes méthodologiques pour une science transparente

- Christophe Pouzat, Arnaud Legrand, Konrad Hinsen

Un rapport↗

- Les usages des données du GBIF (2019)