Rencontres Data

COMPTE-RENDU DES SESSIONS DE JANVIER 2019 - CeRIS & DSI

Institut Pasteur

Les Rencontres Data ont été initiées par la DSI et le CeRIS pour faire remonter du terrain les problématiques liées à la gestion des données et permettre aux personnes qui manipulent et qui gèrent des données à l'Institut Pasteur de se rencontrer et de discuter de sujets d'intérêt.

Les premières Rencontres Data ont eu lieu les 17, 23 et 29 janvier 2019. Ces rencontres étaient prioritairement destinées aux chercheurs et ingénieurs qui manipulent des données scientifiques sur le campus de l'Institut Pasteur (data manager, bio-informaticiens, biostatisticiens ou encore « référents data » d'une unité…).
Au total, 36 personnes ont participé aux premières Rencontres Data : 28 scientifiques et 8 personnes des services supports.

Les objectifs étaient les suivants :

- **Donner une vue d'ensemble sur la gestion des données scientifiques** : cycle de vie des données, problématiques de la gestion des données à l'Institut Pasteur, enjeux et bénéfices de la gestion des données, exigences des financeurs...
- **Présenter le modèle de plan de gestion des données de l'Institut Pasteur** et la façon de l'utiliser en pratique via l'outil REDCap.
- **Échanger sur les pratiques de gestion des données** et sur les besoins en termes d'informations, de formations, d'accompagnement…

Ces 3 sessions ont été riches d'échanges et ont fait ressortir des questionnements et des attentes concernant la gestion des données à l'Institut Pasteur.

Des attentes de directives venant de la direction de l'Institut Pasteur

A chaque session, certains participants ont exprimé le besoin de recommandations ou même d'obligations permettant d'harmoniser et/ou de cadrer les pratiques à l'Institut Pasteur :

- **Recommandations sur les bonnes pratiques et règles de gestion des données** de manière à les rendre FAIR (Facile à trouver, Accessible, Interopérable et Réutilisable). Les projets sont de plus en plus collaboratifs avec de fréquents échanges de données notamment en interne, entre unités de recherche et plateformes. Des règles de gestion de données permettraient un gain de temps et d'efficacité, les data-managers recevant des données « propres » pour pouvoir les analyser plus efficacement.
- **Obligation de dépôt des données** (brutes et analysées) **et logiciels produits à l'Institut Pasteur dans un espace de stockage centralisé à l'Institut Pasteur**. Ces données et logiciels doivent être accompagnés de métadonnées standardisées, permettant la compréhension et la réutilisation des données. Plusieurs instituts ont mis en place une politique d'obligation de dépôt des données : l'institut Max Planck, le NIH, le CERN, la NASA… Pour rendre effective cette obligation, il a été suggéré de s'inspirer des pratiques mises en place par ces instituts, comme d'indiquer l'obligation de dépôt dans le contrat de travail de tout chercheur arrivant à l'Institut Pasteur.
- **Recommandations sur les outils et méthodes à utiliser pour une bonne sécurité des données.** Par exemple : quel téléphone portable utiliser, comment faire pour que le VPN s'active avant la wifi, comment partager ses données de façon sécurisée, quels outils et logiciels sont fiables ?

Des retours positifs sur le plan de gestion des données (PGD)

Plusieurs financeurs demandent désormais de rédiger un PGD pour chaque projet de recherche financé. Ce PGD, rédigé au début du projet, facilite la planification de la gestion des données tout au long du projet. Il permet également d'estimer les frais nécessaires à la gestion et à l'ouverture des données et qui pourront parfois être pris en charge par le financeur. Cette prise en charge par les financeurs pose question : certains participants se demandent par exemple si les frais de recrutement d'un data manager sont éligibles. Les politiques des financeurs étant différentes, le Grants Office est là pour vous aider à répondre à ce type de questions.

De nombreux participants ont vu le potentiel qu'apporterait le **PGD comme outil permettant d'optimiser la gestion et le partage de données** et préconisent de le rendre visible s'il ne contient pas d'informations confidentielles, de manière à indiquer aux autres chercheurs si les données sont partageables, et augmenter le potentiel de collaboration au sein de l'Institut Pasteur mais aussi avec l'extérieur.

La plateforme Biomics et le Hub de Bioinformatique ont pour projet de **mettre en place un PGD** commun à l'ensemble des projets de leur plateforme respective. Ce PGD serait adapté pour les plateformes, avec des questions moins spécifiques que pour un PGD de projet. Un groupe de travail sur ce sujet va être monté.

Pour la DSI, le PGD permettrait d'avoir une **visibilité sur les données nouvellement produites à l'Institut Pasteur**. En particulier pour différencier :

- Les données qui doivent être sécurisées, des données qui peuvent être stockées sur un espace moins sécurisé.
- Les données à conserver sur le long terme, des données que l'on peut ou que l'on doit supprimer (contraintes réglementaires).

Des retours positifs sur l'utilité de l'outil REDCap pour remplir le PGD

Le PGD vient d'être implémenté dans REDCap et le formulaire REDCap est amené à évoluer en fonction des retours des utilisateurs.

Lors des 3 sessions plusieurs fonctionnalités que propose l'outil ont été plébiscitées :

- **Possibilité de travailler à plusieurs sur le PGD**, y compris avec des partenaires extérieurs.
- **Gestion fine des droits des utilisateurs** : par exemple, pour rendre le PGD visible à son responsable.
- **Possibilité de dupliquer le PGD entier avec ses réponses,** pour ne pas avoir à ressaisir des informations proches.
- **Possibilité de versionner le PGD.**

Certains participants sont demandeurs de fonctionnalités supplémentaires qui vont être étudiées par l'équipe REDCap de l'Institut Pasteur.

Des interrogations sur la conservation des données

Constat : à l'Institut Pasteur, certaines données sont stockées sur le serveur Gaïa de la DSI (anciennement nommé Atlas) mais ne sont plus accédées depuis 2 ans (33% du volume total de données). Parmi ces données, on peut trouver :

- Des données en attentes de publication ;
- Des données qui ont un intérêt à être conservées, car uniques ou ayant coûté cher à produire (données de séquençage par exemple) ;
- Des données sur lesquelles on ne sait rien : ni leur auteur, ni leur nature…

Ces données sont toutes conservées de la même façon (avec le plus haut niveau de sécurisation possible). Pourtant, certaines données pourraient être supprimées ou stockées dans un espace de stockage moins onéreux : un stockage « froid ». Cependant, la DSI n'est pas en mesure de faire le tri parmi ces données (elle ne connait pas la nature de ces données, leur auteur ou leur niveau de sensibilité). Un projet de cartographie des données en collaboration avec les chercheurs permettrait d'optimiser la gestion du stockage.

Certains participants s'interrogent également sur la conservation à long terme :

- Quels sont les formats pérennes ?
- Quels entrepôts permettraient une conservation à long terme des données ?
- Quelles sont les contraintes réglementaires de conservation des données ?

Le service des Archives de l'Institut Pasteur est disponible pour vous aider à répondre à ces questions et propose des accompagnements personnalisés en termes de mise en place de bonnes pratiques de gestion des documents et données.

Des questions techniques liées aux infrastructures informatiques à l'Institut Pasteur

De nombreuses questions ont été posées lors des Rencontres Data sur le fonctionnement de l'infrastructure technique de la DSI :

- Dans Gaïa, peut-on gérer les droits des utilisateurs (lecture, écriture) au sein même de l'unité ?
- Y a-t-il des solutions de cloud français, que l'on pourrait utiliser à l'Institut Pasteur ?
- Comment partager des données au sein d'un département (pour partager les données produites par un instrument par exemple) ?
- Comment utiliser au quotidien le cluster de calcul dans nos analyses ?
- Sécurité des données : quelles solutions, recommandations ?
- Quand un membre d'une unité part de l'Institut Pasteur, que devient son espace Gaïa ?
- Est-il possible de centraliser les données de séquence dans un serveur à l'Institut Pasteur ? Ces données pourraient être ouvertes après un certain délai si elles n'ont pas été publiées. Peut-on faire la même chose pour les images ?

Ces questions sont prises en compte par la DSI qui apportera prochainement des réponses concrètes.

Des questions d'ordre juridique sur les données

Lors des Rencontres Data, plusieurs thématiques juridiques ont soulevé des questions :

- Anonymisation des données, en particulier dans une optique de partage de données
- Législation RGPD
- Propriété des données
- Conditions d'utilisation de données externes
- Licence à attribuer aux données et aux logiciels
- Aspects juridiques liés à la collecte des données

Il y a parfois une contradiction entre les éditeurs ou les financeurs qui demandent de mettre à disposition les données brutes et les contraintes réglementaires (données à caractère personnel par exemple). Les participants se demandent comment partager les données lorsqu'elles sont soumises à une réglementation.

Des interrogations sur la façon d'ouvrir ses données

Certains participants ont trouvé regrettable que toutes les données produites et pouvant être partagées ne sont pas ouvertes dans les faits. Sans obligation, il leur semble que les chercheurs de l'Institut Pasteur n'ouvriront pas leurs données, ce qui est dommageable pour la communauté scientifique qui ne pourra pas réutiliser ces données.

Certains participants se demandent comment rendre concrètement leurs données FAIR (Faciles à trouver, Accessibles, Interopérables, Réutilisables) et quels entrepôts de données sont considérées comme fiables pour publier ses données en dehors de l'Institut Pasteur.

Le CeRIS et la DSI ont pris en compte ces questions et vont notamment proposer prochainement une liste d'entrepôts de confiance en sciences de la vie.

Conclusion

Les rencontres Data ont été l'occasion de nombreux échanges qui ont fait émerger différentes thématiques. L'objectif à présent est d'approfondir les points soulevés lors de prochaines Rencontres Data. L'idée de créer des groupes de travail sur les différents types de données a été soulevée et sera étudiée prochainement.

Le CeRIS et la DSI sont disponibles pour répondre à vos questions :

- sur les questions informatiques liées aux données : data-dsi@pasteur.fr
- sur le plan de gestion des données : pgd@pasteur.fr
- sur REDCap : redcap@pasteur.fr

N'hésitez pas à consulter également la page webcampus du CeRIS sur la gestion des données de recherche.