Gestion de données scientifiques

Introduction

Le contexte national et européen des infrastructures de recherche favorise de plus en plus des politiques de gestions des données. En effet, pour être inscrit sur les feuilles de route nationale et européenne, les instituts sont dans l'obligation d'avoir une telle politique. Les données produites doivent répondre à des objectifs fixés dont notamment celui de l'ouverture des données.

De telles politiques sont donc mises en place au niveau institutionnel, au sein de l'Inra. La charte de l'institut repose sur les six valeurs suivantes : « ouverture, partage, transparence, accès aux données, soutenabilité financière et conformité. », (INRA, 2016a). En effet, en signant la déclaration de Berlin en 2004 qui incite à rendre l'information scientifique librement accessible, l'Inra est entré dans une démarche Open Science. Cette démarche permet d'ouvrir les données de la recherche et de les rendre accessibles à tous. L'objectif est de capitaliser les efforts expérimentaux de chacun et d'éviter les duplications. Cela permettrait d'augmenter l'efficacité de la recherche. La loi Lemaire « Pour une République Numérique » du 7 octobre 2016 incluant la problématique des données de la recherche, renforce la démarche Open Science déjà mise en place. Découle de cela une charte publiée par l'Inra pour le libre accès à ses publications et données scientifiques en février 2017. Cette charte comprend sept points de recommandations destinés aux chercheurs afin de les inciter à entrer dans la démarche Open Science (INRA, 2016b).

D'autre part, depuis quelques années, avec la révolution du numérique et l'acquisition de jeux de données toujours plus conséquents, l'Inra s'est fixée des objectifs Open Sciences. Ces objectifs comprennent : - L'organisation des infrastructures de recherche afin qu'elles soient connectés - L'organisation des données pour faciliter leur partage et leur réutilisation notamment grâce à la création d'un portail de données et de collaborations avec le CEA, le CNRS ainsi que l'Inria et l'Irstea. - Favoriser les approches prédictives en biologie et en écologie - Proposer de nouveaux modes de diffusion de la connaissance notamment en adaptant les revues propriétaires à la science ouverte mais également en encourageant des modèles alternatifs de publications. - Adapter le métier et l'environnement du chercheur au numérique en demandant des compétences en analyse de jeu de données massifs par exemple ou encore en améliorant le processus de dématérialisation. En outre, Horizon 2020, le programme de financement de la recherche et de l'innovation de l'Union européenne pour la période 2014-2020 s'inscrit lui aussi dans cette démarche de science ouverte. En effet, toutes les publications issues de projets financés par ce programme doivent être en libre accès. De plus, il existe un pilote ORD : Open Research Data qui tend à rendre accessibles le plus de données possibles acquises dans le cadre d'un projet H2020 (Ministère de l'enseignement supérieur, de la recherche et de l'innovation, 2014).

Enfin, en 2018, apparaît le plan national pour la science ouverte qui « rend obligatoire l'accès ouvert aux publications ainsi qu'aux données issues de recherches financées sur projet ». Ce plan est composé de 3 axes décrivant des mesures afin de « généraliser l'accès ouvert aux publications », « structurer et ouvrir les données de la recherche » ainsi qu'afin de « s'inscrire dans une dynamique durable, européenne et internationale » (Ministère de l'enseignement supérieur, de la recherche et de l'innovation, 2018).

Cette synthèse présentera donc les modalités liées à l'Open Science, comment s'inscrire dans cette démarche de valorisation des données. Pour cela, une première partie abordera les enjeux de l'accessibilité aux données tandis que les parties suivantes se verront plus techniques et permettront de donner un cadre de gestion des données au sens large, toujours dans un objectif d'open science. En effet, ces parties traiteront de la production de données FAIR, de la gestion des données, de leur partage et de leur archivage et enfin des différentes formes de valorisation envisageables. La sixième et dernière partie comprends trois études de cas

de data paper afin d'avoir des exemples concrets du cycle de vie des données à travers la structure d'un document de valorisation.

I. Les enjeux de l'accessibilité aux données scientifiques

L'open data est un concept récent parfois encore mal connu du personnel scientifique. Les enjeux de l'ouverture des donnés sont multiples, que ce soit pour les personnes ou les instituts producteurs de données ou pour les utilisateurs des données (Pôle Données de la Recherche IST, 2018e), (Aumont, 2017).

1. Enjeux patrimoniaux

Inciter les chercheurs à rendre leurs données accessibles permet tout d'abord d'éviter de perdre des données. En effet, de nombreuses données sont encore dans des formats non pérennes voir peut-être non lisibles car trop vieux pour être remis à jour. De plus, de nombreux chercheurs n'ont plus connaissance des lieus de stockage des données acquises au long de leur carrière ou n'ont pas toujours pris le temps nécessaire pour stocker leur données correctement. En effet, il est estimé qu'environ 80% des données sont perdues 20 ans après publication (Roberge, 2015). D'autre part, le problème du stockage des données est d'autant plus important pour des chercheurs ayant travaillé dans plusieurs instituts différents au cours de leur carrière. De plus, dans le cas où les supports de stockage sont encore disponibles et lisibles par d'autres scientifiques, les données sont parfois trop peu décrites à l'aide de métadonnées pour être comprises et réutilisées par d'autres. Enfin les données apparaissant dans les publications sont généralement des données traitées et il est souvent difficile de retrouver les données brutes initiales à partir de celles-ci. Or, l'accessibilité à des données plus anciennes est impérative dans certains domaines comme la météorologie par exemple ou tous les domaines concernant l'évolution des sols.

2. Enjeux économiques

Tout d'abord, il est inutile que les fonds de l'Etat financent des expérimentations supplémentaires sur un projet alors que les données produites par d'autres instituts pourraient être en partie réutilisées car complémentaires. L'Open data (données ouvertes) permet donc « d'accélérer l'innovation et le retour sur Investissement dans la R&D » (Pôle Données de la Recherche IST, 2018e). A titre d'exemple, on peut citer l'institut européen de bio-informatique (EMBL-EBI), organisation intergouvernementale fournissant gratuitement des données dans le monde entier. Les avantages sont estimés à 1 milliards de livres sterling par an pour les utilisateurs et leurs bailleurs de fonds ce qui correspond à 20 fois le coût opérationnel de l'institut (Beagrie, Houghton, 2016).

3. Enjeux scientifiques

D'une part, l'accessibilité aux données garantit une certaine qualité. En effet, de nombreuses questions se posent actuellement sur la qualité des productions scientifiques et notamment sur le niveau de signification statistique des conclusions tirées. Il semblerait que seulement 20% seraient statistiquement fiables (Aumont, 2017). En publiant les données brutes, la recherche peut s'affranchir de ce risque, les données étant la preuve de l'exactitude de la publication. Plus les données seront vues et réutilisées par des collaborateurs et plus le risque d'erreurs ou de conclusions hâtives sera évité.

De plus, en amont, le chercheur qui publiera ses données sera plus attentif. Toutefois, publier ses données entraîne de nombreux changements techniques pour la communauté scientifique et de nouvelles responsabilités.
Pour l'institut de recherche qu'est l'INRA, l'accessibilité aux données permet d'assurer l'intégrité de ses recherches et produits. D'autre part, l'accessibilité des données peut être une passerelle entre les instituts

de recherche, producteurs de données, et les instituts techniques, utilisateurs de données. En effet, chaque année, 7% des adresses mail des chercheurs ne sont plus fonctionnelles ce qui rend plus difficile l'accès aux informations. De plus, l'Open data permettrait également de faciliter l'usage des données notamment par ces communautés techniques. Enfin, l'ouverture des données les rend citables et améliore la vitesse d'accessibilité à ces dernières. Découle de cela un taux de citation pouvant être plus important ce qui est dans l'intérêt des producteurs de données qui voient leur travail reconnu.

4. Enjeux sociétaux

Comme il a été écrit précédemment, l'Open data permet une traçabilité des données et garantit une certaine qualité. L'accessibilité aux données contribue donc à améliorer l'image de la recherche auprès des citoyens ainsi qu'à augmenter leur confiance et donc leur participation à la science (sciences participatives). En termes d'éducation, les jeux de données publiés pourraient permettre à des étudiants de les intégrer à leurs productions pour l'argumenter ou encore de s'entraîner sur de vraies données. L'Open data permettrait donc de nouveaux liens entre citoyens et instituts de recherche.

5. Enjeux éthiques

L'ouverture des données nécessite de bien réfléchir en amont à la question « Quelles données publier ? ». En effet, d'un point de vue éthique et juridique les chercheurs se doivent de respecter les droits d'auteurs, la vie privée et selon le domaine il est parfois possible de rencontrer des obligations de secret ou de sécurité. Selon le type de données, l'ouverture ne sera donc pas toujours possible notamment pour des données personnelles ou encore des données concernant le domaine de la santé par exemple.

II. Gérer des données

1. Etablir un plan de gestion des données (PGD)

Documenter et organiser les données

2. Nommage et organisation des fichiers et dossiers

3. Documenter les données

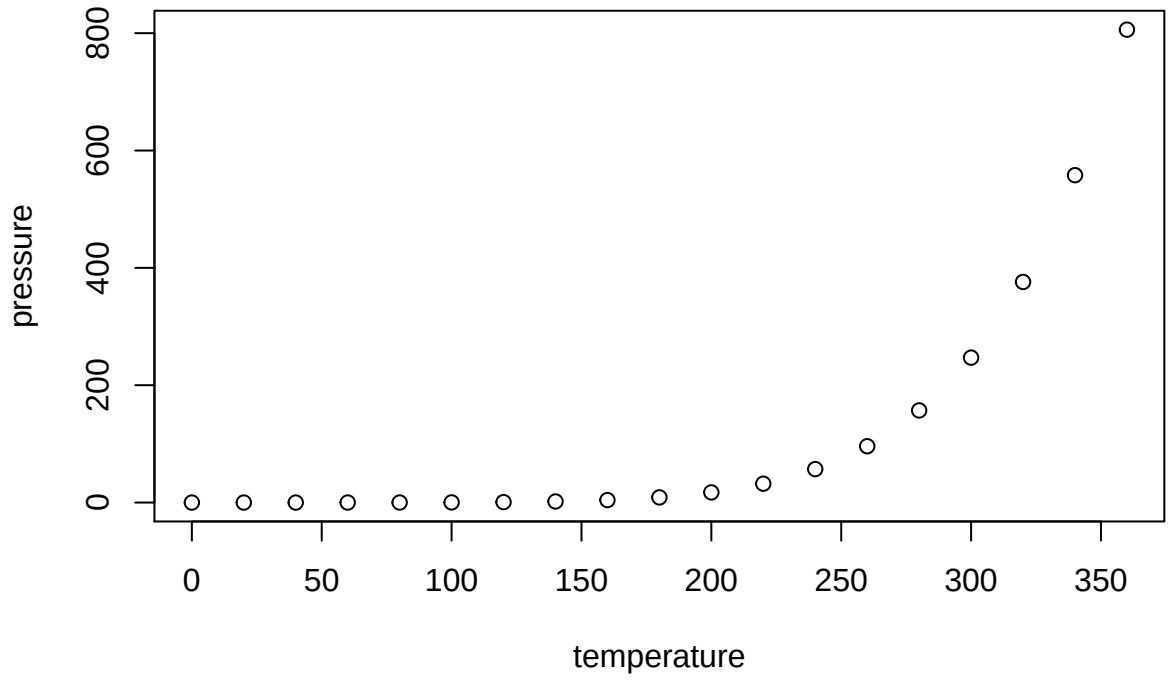
4. Identifier les données

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.