Guide de bonnes pratiques pour la gestion des données d'observation

Auteur·es	membres du réseau SIST
Date de création	07/11/2019
Date de dernière révision	07/11/2019

Ce document résulte de l'atelier du même nom réalisé lors du séminaire SIST 2019 à l'Observatoire Midi-Pyrénées.

Il doit être détaillé, étoffé (outils et technologies associées, etc) et éventuellement restructuré. Il sera révisé aussi régulièrement que possible et nécessaire.

Traitement des données

Qualité, qualification, correction, niveau de données, versionnement, …

De manière générale on préconise de conserver toute trace de ce qui a permis de générer un niveau de donnée (ou produit) ou version d'un niveau de données.

Plus précisément :

- expliciter les noms utilisés pour chaque niveau de données, si possible avec un vocabulaire contrôlé;
- fournir une description de la chaîne de traitement sous la forme de listes ou de graphes;
- conserver les algorithmes, codes, ... dans la version correspondant à un traitement donné : dépôt github + export zenodo afin de les rendre accessibles et citables facilement (DOI);
- utiliser les environnements d'exécution du type Jupyter Notebook de manière à donner accès à l'ensemble des logiciels et configurations utilisés dans un traitement.

Concernant la qualification il est conseillé :

- d'être capable d'estimer et conserver l'erreur sur la mesure et les traitements;
- d'exprimer l'intervalle de confiance, empiriquement en faisant appel à un expert du

domaine et/ou statistiquement lorsque cela est possible;
- de comparer les données par rapport à une référence ou par rapport à une autre source de données (ex : température relevé par une station météo versus carte thermique satellite);
- taguer les données mais ne pas les effacer.

Le point d'entrée de toute ces informations doit être la fiche de métadonnées : une fiche de métadonnées + un DOI par niveau (ou produit).

Utilisation de fragments pour citer les versions différentes au sein d'un niveau.

Il est préconisé de bien choisir les logiciels et d'estimer leur robustesse. De manière générale d'être capable de produire un traitement à l'identique (reproductibilité).

Pérennité des données

Comment faire en sorte que les données seront toujours existantes et utilisables dans le futur

Il nous faut définir plusieurs notions :

- c'est quoi la pérennité ? : différencier stockage et archivage (processus de plus : migration de formats, répartition de copies)
- indiquer la durée : court / long / moyen terme (à définir)
- quelles données sont susceptibles d'être pérennisées ?
- qui s'en charge ? qui finance?
- comment on réalise cet archivage pérenne ?

Pérenniser implique un archivage long terme.

Il nous faut définir de quelle missions d'organisme relève la pérennisation : qui l'opère ? les labos ? les OSU, les pôles de données ? les établissement publics comme le CINES ? externalisation vers organismes privés ?
Pérenniser les données nécessite une volonté politique et un financement associé qui peut être défini dans un PGD.

Comment ?

- Définir des formats de données ouverts, ne pas accepter tous les formats, assurer les migrations régulières de version dans le temps des formats et des supports numériques
- Mettre en place la règle 3/2/1 : 3 sauvegardes sur 2 supports différents et 1 site distant externe
- Avoir des identifiants pérennes (DOI)

Analyser les données pour identifier celles à pérenniser.

Associer les métadonnées + les référentiels et vocabulaires aux données et variables : sans les métadonnées l'archivage ne vaut rien.

La fiche de métadonnées doit indiquer le lieu d'archivage pérenne.

Accessibilité des données

Comment faire en sorte que les données soient (re)-trouvables, identifiables, accessibles, dans le respect des droits (au sens large)

(Re) Trouvables : ***rendre visibles***

Afin d'être trouvables c'est-à-dire visibles :

- les données doivent être cataloguées
- les métadonnées doivent êtres normalisées et inclure la licence d'utilisation
- Les données et les méta données doivent êtres moissonnable, donc dans un dépôt interopérable, que ce soit pour les données ou les méta données
- les données doivent être visibles sur des catalogues "officiels" ou institutionnels

Identifiables :

- le jeu de donnée est identifié de manière unique : DOI, UUID, URI…
- Identification du contenu

Accessibles : où ? quelle durée ? => Pérennité d'accès

- Infrastructure d'hébergement du SI de diffusion
- Disponibilité dans les SI et téléchargeable avec des protocoles d'accès normalisé : entrepôt de données, plateforme d'accès, API
- Répliquer pour le pérenniser
- Aperçu ou visualisation complète
- formats interopérables

Respect du droit

- RGPD
- Lieu de stockage (France ou Europe uniquement ? / Données concernant les citoyens par exemple DOIVENT êtres hébergées sur le cloud Français)
- ETALAB
- Licences d'accès ou utilisation
- Propriété intellectuelle de la donnée

==> EMBARGO technique pour des questions scientifique ou de confidentialité : Authentification par exemple

==> limitation d'accès (par exemple : taille des tuyaux) pour des raisons techniques (éviter effondrement infrastructure par exemple)

==> Contrôle de l'usage de la donnée (traçabilité, identification)

Ré-utilisabilité, interopérabilité

Comment faire en sorte que les données soient interopérables entre elles et avec les

services

Pré-requis : les données doivent être trouvables et accessibles, soit "FA" de "FAIR".

Les données peuvent être interopérables sans être réutilisables ("FAI"), et inversement ("FAR"). Idéalement, elles doivent être "FAIR".

L'interopérabilité se décline en 4 catégories :

- juridique et organisationnelle : conventions, normes, standards, plan de gestion de données (connaissance des outils permettant une mise en place aisée des protocoles d'interopérabilité)
- technique : systèmes et réseaux, protocoles d'échanges
- syntaxique : formats ouverts (txt -> tsv et binaires -> NetCDF ou HDF), outils de conversion, structure des données
- sémantique : ontologies, thésaurus, vocabulaires contrôlés (recommandations : limiter le nombre de vocabulaires, multilinguisme avec a minima l'anglais)

Des recommandations pour la réutilisabilité :

- décrire la provenance et les relations dans les métadonnées. La notion de traçabilité des données est fondamentale (identifiants pérennes, données sources)
- décrire la structure des données dans les métadonnées (catalogue d'attributs)
- garantir des données "de qualité" (cf. curation)
- privilégier l'utilisation d'environnements d'exécution des traitements (ex : Jupyter notebook)

Plus généralement, il est nécessaire de sensibiliser/former les communautés sur ces pratiques.

Notes

Guide de bonnes pratiques Observatoire de Lorraine OTELO => https://hal.archives-ouvertes.fr/hal-01275841/document