Gestion de métadonnées pour les espaces de stockages de données

François Ehrenmann, Philippe Chaumeil, Daniel Jacob, Edouard Guitton

Gestion de métadonnées pour les espaces de stockages de données.

Comment mettre en œuvre de bonnes pratiques de gestion des données qui ne soient pas trop restrictives (simplicité, flexibilité, robustesse, évolutivité) ?

Table des matières

Motivations

- Répondre aux enjeux de l'organisation, de la documentation, du stockage et du partage des données d'un site, d'un projet ou d'une structure (unité, plateforme, …).
- S'inscrire dans une démarche qualité de science ouverte pour le partage et la reproductibilité.

Préambule : État du besoin

- La mise en place d'un plan de gestion de données impose des prérequis comme l'externalisation des données à préserver hors de l'espace disque des utilisateurs. Cela ne concerne pas seulement les données publiées mais toutes les données produites pendant la durée d'un projet. Cette externalisation permet surtout de rassembler les données en un même endroit et constitue déjà une sauvegarde de premier niveau. Cela s'avère encore plus nécessaire lorsque des agents temporaires (doctorants, post-docs, stagiaires, CDD) interviennent dans la production de données.
- En conséquence, se pose le souci de l'organisation de ces espaces de stockage. Faut-il les harmoniser, c.à.d. imposer des bonnes pratiques comme i) le nommage des dossiers et des fichiers, ii) une structure des dossiers (docs, data, scripts, …), iii) l'utilisation de fichiers README, iv) …
- A minima l'utilisation d'un fichier README semble la plus simple et la moins contraignante. Mais se pose ensuite la question "que mettre dedans" ? Des templates peuvent être proposés afin de simplifier leur rédaction. Mais se pose alors la question du comment les exploiter efficacement lorsque l'on souhaite retrouver de l'information ? Avec quel(s) vocabulaire(s) ?

Approche proposée

- Les deux idées principales à l'origine de l'outil sont :

    1. que l'espace de stockage soit aussi un référentiel de données, en faisant en sorte que les métadonnées aillent vers les données.
    2. de pouvoir "capturer" les métadonnées de l'utilisateur aussi facilement que possible en utilisant son vocabulaire.

- Concernant la première idée : il "suffit" de placer un fichier de métadonnées (format JSON) décrivant les données du projet dans chaque sous-répertoire, pour ensuite, trouvez les projets et/ou les données correspondant à des critères précis. Le choix s'est porté sur le format JSON, très approprié pour décrire les métadonnées, lisible à la fois par les humains et les machines.

- Concernant la deuxième idée : Compte tenu de la diversité des domaines, l'approche choisie est d'être à la fois la plus souple et la plus pragmatique possible en permettant aux utilisateurs de choisir leur propre vocabulaire (contrôlé ou non) correspondant à la réalité de leur domaine et de leurs activités. Cependant, une bonne approche consiste autant que possible à n'utiliser que du **vocabulaire contrôlé**, c'est-à-dire un vocabulaire pertinent et suffisant utilisé comme référence dans le domaine concerné pour permettre aux utilisateurs de décrire un projet et son contexte sans devoir ajouter des termes supplémentaires. Ainsi une cartographie des termes basée sur un vocabulaire contrôlé pourra être réalisée plus facilement pour générer des formats correspondant à différents standards (MIAPPE, JSON-LD, …).

- La création du fichier de métadonnées au format JSON se fait à l'aide d'une interface web. Sachant que cette interface web doit correspondre au contexte scientifique et expérimental du site (unité, projet, …), l'ensemble des métadonnées à saisir est entièrement paramétrable à l'aide d'un fichier de configuration (format TSV) pouvant être produit facilement à partir d'un tableur type MS Excel.

- Le fichier de configuration permet de définir :

    1. L'ensemble des variables (champs) à saisir (noms apparaissant dans le fichier de métadonnées),
    2. Leur label correspondant,
    3. Leur type, à savoir soit une valeur unique à saisir (textbox), soit multiple sous forme de mots-clés (checkbox), soit une valeur unique à sélectionner parmi un choix limité (dropbox). Les listes des termes prédéfinies peuvent être aussi spécifiés dans ce fichier, constituant ainsi du vocabulaire choisi et donc contrôlé par le gestionnaire de données.
    4. Une variable (champ) spéciale nommée "resources" peut être ajoutée à la liste des variables. Cette variable "resources" vous permettra d'ajouter, via l'interface web, une liste de ressources internes ou externes avec leur type, leur description et leur emplacement. Un emplacement peut être n'importe quoi : un texte, un chemin absolu dans un arbre, un lien URL, … Vous pouvez ainsi mettre par exemple un lien vers une publication (Type=article, lien=DOI).

- Le fichier de configuration est ensuite converti lui-même au format JSON et ainsi devenant complètement interprétable par la machine afin de générer l'interface web. (Cf.

dépôt GitHub du code source avec les détails techniques : https://github.com/inrae/pgd-mmdt/).

- L'interface web vous permet de :
  - **Décrire** un jeu de données à l'aide de métadonnées de différents types (Description).
  - **Rechercher** des jeux de données par leurs métadonnées (Accessibilité)

- Synoptique de la démarche permettant de gérer les métadonnées de n'importe quel jeu de données.

  1. Dans un premier temps, vous saisissez les métadonnées concernant le jeu de données à l'aide de l'interface web. A des fins de sélection, un ensemble de termes peut être prédéfini, ou des termes peuvent être ajoutés à la liste et stockés pour une utilisation future.
  2. En sortie, vous récupérez le fichier JSON généré, puis le déposez dans le répertoire de données correspondant.
  3. Pour que ce fichier de métadonnées soit pris en compte, vous pouvez soit lancer manuellement le script de recherche (scan), soit attendre que la prochaine recherche automatique soit déclenchée (via cron).
  4. Enfin, vous pouvez rechercher vos jeux de données à l'aide d'une interface web dédiée pour les retrouver en spécifiant un ou plusieurs critères.

Mise en œuvre

Ouverture des données

- Il est à noter que dans l'approche proposée et décrite précédemment, il n'est nullement question d'ouvrir les données, mais de gestion de métadonnées associées aux données sur un espace de stockage avec un périmètre précis que représente le collectif (unité, équipe, projet, plateforme, …). Néanmoins, l'ouverture des données via leurs métadonnées doit être un objectif clairement affiché dans le cadre de projets financés par des institutions publiques (EU, ANR, Régions, …).

- La principale caractéristique de l'outil est, avant tout, de "capturer" les métadonnées le plus facilement possible selon un référentiel bien choisi. Ensuite comme précisé plus haut, si l'on a pris soin de définir correctement ses vocabulaires contrôlés correspondant au profil de son domaine d'application alors il devient possible de définir une correspondance entre des termes basés sur un vocabulaire contrôlé et des termes basés sur des ontologies.

  1. Formalisme FAIR
     - L'ouverture des données : « Ouvert autant que possible, fermé autant que nécessaire »
     - L'accessibilité concerne avant tout les métadonnées.

        - Les métadonnées doivent être ouvertes et disponibles même au cas où les données ne seraient pas ou plus disponibles ni accessibles.
        - Les conditions d'accès aux données doivent être spécifiées (contact & licence)
    - PGD-MMDT : Un premier accès peut se faire au niveau des métadonnées. En effet si l'interface web est accessible via l'internet alors les métadonnées elles-mêmes peuvent déjà être rendues accessibles (ex: https://pmb-bordeaux.fr/pgd-mmdt/metadata/Atacama). Même si les données ne se sont pas accessibles, ce qui compte c'est que les métadonnées le soient en y incluant les conditions d'accessibilité des données (contacts & licence).

2. Directives de reporting

    - Métadonnées : des données descriptives sur les données, qui fournissent les informations contextuelles essentielles pour interpréter et réutiliser les données.
    - Les directives de reporting (ex : MIAPPE) jouent un rôle central, car elles définissent les descripteurs clés que la communauté considère comme les informations nécessaires et suffisantes qui doivent être rapportées pour contextualiser et comprendre les ensembles de données.
        - cf https://en.wikipedia.org/wiki/Minimum_information_standard
        - Dans les sciences de la vie, par exemple, les descripteurs des étapes expérimentales (par exemple, la provenance des matériaux d'étude, les types de mesure et de technologie) et les entités moléculaires d'intérêt (par exemple, les métabolites, les protéines) sont des informations essentielles pour assurer une réutilisation efficace et significative des données.
    - PGD-MMDT : Définition des « sections » et des « champs » basés sur les directives de reporting

3. Vocabulaire contrôlé

- Un vocabulaire contrôlé est un lexique (c.à.d. Ensemble des termes utilisés dans un domaine spécifique, par une communauté) dont le but est de rendre possible l'organisation des connaissances afin d'optimiser la recherche d'information. Le vocabulaire contrôlé est utilisé dans les schémas servant à l'indexation (c.à.d. Donner à un document, un dossier, un signe distinctif, afin de le classer) comme dans les thésaurus et les taxinomies
- CV : Capture des métadonnées par les « gestionnaires de données » au sein des collectifs
- Ontologies : Mise en correspondance avec un modèle de connaissance par les « bioinformaticiens » en fonction du domaine d'application, des directives de reporting (MI), de l'entrepôt de données spécifiques, …
- PGD-MMDT : Termes prédéfinis à l'aide de CV soit par pré-sélection (dropbox, checkbox), soit par autocompletion (textbox, checkbox), soit les deux (checkbox).
    - En particulier il est possible par autocompletion de choisir un terme en provenance de BioPortal. Un exemple de javascript est fourni (cf. web/js/autocomple/bponto.js) permettant de récupérer tous les termes (classes) relatifs aux formats dans l'ontologie EDAM via l'API BioPortal.

- Il est aussi possible par autocompletion de choisir un terme en provenance du thesaurus-INRA. Un exemple de javascript est fourni (cf. web/js/autocomple/VOvocab.js) permettant de récupérer tous les termes à partir de mots-clés.
- L'avantage des thésaurus, avec une hiérarchisation minimale, est qu'ils proposent des Vocab. Contrôlés rangés par thématique. C'est très exactement ce que nous avons besoin collectivement pour annoter nos données. Cela permet de bâtir des applications simples à utiliser notamment pour le choix du vocabulaire. De plus, en choisissant bien ses vocabulaires contrôlés de manière de favoriser ensuite le mapping (mise en correspondance avec une ontologie par ex.), le bioinformaticien peut facilement produire des formats standards (ex MIAPPE, JSON-LD).

- Mapping : Mise en correspondance avec un modèle de connaissance (ontologies) à partir des CV (bioportal, VOINRAE).

4. Identifiants pérennes

- Identifiants uniques et persistants : pierre angulaire des principes FAIR
- PGD-MMDT : ORCID, liens vers ressources (DOI, SWID, …)

5. "Machine-actionable metadata"

- La recherche publique française dispose d'un dépôt de données national (Recherche Data Gouv https://entrepot.recherche.data.gouv.fr/). Par une approche que l'on pourrait nommer "machine-actionable metadata", il est là aussi possible de pré-remplir un jeu de données dans le dataverse de son choix via l'API. Les tâches les plus cruciales étant 1) de bien définir son profil de métadonnées, 2) de bien choisi ses vocabulaires contrôlés, 3) de faire le mapping adéquat ; ensuite le code à développer ne représente aucune difficulté particulière. Un Use-Case est en cours de développement afin de montrer la preuve de concept.

Infrastructure

- Avant d'envisager une mise en œuvre de l'approche proposée, des points importants sont à considérer.
  - Tout d'abord, le périmètre de la gestion des données, ou autrement dit, à quel collectif est destiné l'espace de stockage (unité, équipe, projet, plateforme, …). La nature du collectif, ainsi que le volume des données envisagé conditionnera fortement le choix de l'espace de stockage.
  - Ensuite il y a deux entités à considérer sachant que le choix de la deuxième sera conditionné à la première, à savoir :
    1. L'espace de stockage des données,
    2. Le support matériel pour l'herbagement de l'interface web.
- Le choix de l'espace de stockage des données selon le collectif visé peut être de deux types :
  1. Espace de stockage en interne (type NAS)
  2. Espace de stockage en externe (type Data center, Cloud)

1. Dans le cas d'un espace de stockage en interne (type NAS), il faudra :
   - soit envisager d'installer l'interface web sur le serveur du NAS (si c'est possible)

- soit créer une machine virtuelle sur un datacenter (Service Ariane : “Serveurs virtuels nus ou packagés”) puis d’y installer le VPN GlobalProtect afin d’accéder au NAS de son unité (tests OK – oct. 2022)
- soit envisager l’achat d’un petit serveur web (à partir de 2000€HT),
- soit adopter une approche un peu différente qui serait d’avoir l’interface web installée sur le PC du gestionnaire de données d’une équipe par exemple (en mode virtualisé) afin de gérer l’espace de données.

2. Dans le cas d’un espace de stockage en externe (type Datacenter, Cloud), on peut envisager la création d’une machine virtuelle sur un datacenter (Service Ariane : “Serveurs virtuels nus ou packagés”, https://ariane.inrae.fr/) où l’on pourra installer facilement l’interface web. La connexion (c.à.d. le montage) avec l’espace de stockage pourra là aussi aisément se faire à l’aide d’outils comme rclone (https://rclone.org/). Des essais ont déjà été effectués avec succès (cf. https://pmb-bordeaux.fr/ncloud/).

- L’INRAE met à disposition des espaces de stockage sur le réseau (Cloud) via l’interface NextCloud. De même certaines universités mettent aussi des espaces de stockage partageables via NextCloud (ex Université de Bordeaux avec sa solution “cUBe”). Ces solutions de stockage peuvent être idéales pour un projet impliquant plusieurs partenaires répartis sur toute la France voire l’Europe.

Perspectives

- Mettre en place une librairie de profils (métadonnées par domaine et/ou site + CV & Mapping associés) selon plusieurs use-cases.
- Pré-remplir un jeu de données dans le dataverse INRAE DATA (via API), en établissant un mapping du fichier JSON de métadonnées vers un format JSON-LD impliquant l’utilisation d’une sémantique (schema.org)

Liens

- **Code source sur GitHub** : https://github.com/inrae/pgd-mmdt
- **Présentation** de l’outil et de sa démarche (oct 2022) : https://nextcloud.inrae.fr/s/HxEWSybeBW8rzke
- PGD-MMDT UMR BFP Equipe Meta : https://pmb-bordeaux.fr/pgd-mmdt/

Contributeurs

- François Ehrenmann (UMR BioGECO) | CATI GEDEOP
- Philippe Chaumeil (UMR BioGECO)
- Daniel Jacob (UMR BFP) | CATI PROSODIe
- Edouard Guitton (INRAE Dept. SA, Emerg’IN)