LES ENTREPÔTS DE DONNÉES

PELLETIER Laurent

ANF rBDD

Du 5 au 7 novembre 2018 à Sète

ANF « SCIENCES DES DONNÉES : UN NOUVEAU CHALLENGE POUR LES MÉTIERS LIÉS AUX BASES DE DONNÉES »

LES ENTREPÔTS DE DONNÉES

LES ENTREPÔTS DE DONNÉES

Cette oeuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions 4.0 International.

Sources :

- Action Nationale de Formation RENATIS – MEDICI 2017 (3-6/07/2017) – **Sylvie Cocaud** (DIST Inra)
- Formation URFIST de Rennes 15/06/2017 - **Marie-Christine JACQUEMOT-PERBAL** (Inist-CNRS), **Violaine REBOUILLAT** (Chargée d'étude BSN10 ; Cnam-IDF)
- Colloque « Sciences Ouvertes : expériences, enjeux et perspectives » -Université de Lorraine 11/10/2018 – **Florian Mazur** (Inist-CNRS), **Anne Busin** (Inist-CNRS)

DONNÉES DE LA RECHERCHE

DONNÉES DE LA RECHERCHE
DÉFINITION (1)

Il n'existe pas de définition légale. Seulement des propositions de définition.

Définition proposée par l'OCDE :

«Les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche.

Ce terme ne s'applique pas aux éléments suivants : carnets de laboratoire, analyses préliminaires et projets de documents scientifiques, programmes de travaux futurs, examens par les pairs, communications personnelles avec des collègues et objets matériels (par exemple, les échantillons de laboratoire, les souches bactériennes et les animaux de laboratoire tels que les souris).»

Organisation de Coopération et de Développement Economiques (2007). Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics. http://www.oecd.org/fr/sti/sci-tech/38500823.pdf

DONNÉES DE LA RECHERCHE

DÉFINITION (2)

L'Association des archivistes français :

« Les données de la recherche concernent, en plus des métiers de la recherche, les métiers qui viennent en appui à celle-ci, tels que la documentation, les archives et l'informatique. Chacun de ces métiers a un rôle à jouer dans le cycle de vie des données. L'archiviste apporte son expertise pour leur gestion, leur conservation voire leur communication. Les données de la recherche sont des informations, spécimens et matériaux produits, recueillis et documentés. Elles sont collectées ou exploitées à des fins de recherche et de preuves par les chercheurs et leurs équipes. À ce titre, elles constituent une partie des archives de la recherche »

« Archivistes ! n°120 », www.archivistes.org, 30 janvier 2017

DONNÉES DE LA RECHERCHE

DANS LA PRATIQUE, UNE GRANDE DIVERSITÉ…

STATISTIQUES

RÉSULTATS D'EXPÉRIENCES

MESURES

OBJETS DE FOUILLE

IMAGES

TEXTES

OBSERVATIONS DE TERRAIN

TRANSCRIPTION D'ENTRETIENS

ARCHIVES

CODES SOURCES

DONNÉES DE LA RECHERCHE
MÉTADONNÉES

- Donnée qui renseigne sur la nature d'autres données et qui permet ainsi leur utilisation pertinente.

- Dans la perspective des entrepôts de données, les métadonnées sont un élément primordial et sont destinées à diverses catégories d'utilisateurs. Elles permettent notamment de connaître l'origine et la nature des données stockées dans l'entrepôt, de comprendre comment elles sont structurées, de savoir comment y avoir accès et comment les interpréter, de connaître les différents modèles de données en présence et les règles de gestion de ces données (Office québécois de la langue française, 2002).

- Certaines métadonnées sont générées dès la création des données, automatiquement ou manuellement.

- Fournir les métadonnées le plus rapidement possible après la création des données.

DONNÉES DE LA RECHERCHE

JEU DE DONNÉES

- « Agrégation, sous une forme lisible, de données brutes ou dérivées présentant une certaine "unité", rassemblées pour former un ensemble cohérent »
  Mémoire d'étude pour le diplôme de Conservateur de Bibliothèque (janvier 2014) de Rémi Gaillard (enssib)

- Certains entrepôts utilisent le terme «data package» plutôt que «data set» pour désigner un ensemble constitué des fichiers de données et des métadonnées associées

- Le nombre de fichiers peut être très important dans un jeu de données

- Citable comme un objet unique

DONNÉES DE LA RECHERCHE

FAIR…

« […] les données produites par la recherche publique française soient progressivement structurées en conformité avec les principes FAIR (Findable -Facile à trouver-, Accessible –Accessible-, Interoperable –Interopérable-, Reusable -Réutilisable-), préservées et, quand cela est possible, ouvertes. »

Plan national pour la science ouverte MESRI (juillet 2018)

DONNÉES DE LA RECHERCHE

FAIR… (PRINCIPES (1))

- Faciles à trouver
  - F1 Les données et les métadonnées sont identifiées par un identifiant global unique et pérenne.
  - F2 Les métadonnées décrivant les données sont riches.
  - F3 Les données et les métadonnées sont enregistrées et indexées dans un dispositif permettant de les rechercher.
  - F4 Les métadonnées spécifient l'identifiant de la donnée.

- Accessibles
  - A1 Les données et les métadonnées sont accessibles par leur identifiant via un protocole de communication standardisé.
    - A1.1 Le protocole utilisé est ouvert, libre et peut être implémenté de manière universelle.
    - A1.2 Le protocole utilisé permet l'authentification et l'autorisation si besoin.
  - A2 Les métadonnées sont accessibles même quand les données ne le sont plus

DONNÉES DE LA RECHERCHE

FAIR… (PRINCIPES (2))

- Interopérables
  - I1 Les données et les métadonnées utilisent un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances.
  - I2 Les données et les métadonnées utilisent des vocabulaires qui respectent les principes FAIR.
  - I3 Les données et les métadonnées incluent des liens vers d'autres (méta)données

- Réutilisables
  - R1 Les données et les métadonnées ont des attributs multiples et pertinents.
    - R1.1 Les données et les métadonnées sont mises à disposition selon une licence explicite et accessible.
    - R1.2 Les données et les métadonnées sont associées à leur provenance.
    - R1.3 Les données et les métadonnées correspondent aux standards des communautés indiquées.

DONNÉES DE LA RECHERCHE

PARTAGER… …DANS QUEL BUT ?

- Pour garantir la reproductibilité des données (vérification)
- Pour faciliter la répétabilité des données (réutilisation)
- Pour permettre l'agrégation et l'analyse de données (analyse secondaire)

⇒ Produire de nouvelles connaissances et favoriser l'innovation

ENTREPÔTS DE DONNÉES
PARTAGER… …FACILITE LA DÉCOUVERTE…

…et la réutilisation des données

- Les entrepôts sont scannés par des outils de recherche spécifiques
  - Datacite search
  - Data Citation Index (Thomson Reuters)
  - Data search (Elsevier)
  - …

- et moissonnés par des catalogues, intégrateurs, infrastructures européennes de données… de plus en plus nombreux (OpenAIRE, EOSC pilot, …)

DONNÉES DE LA RECHERCHE
LE PARTAGE… …DOIT ÊTRE PLANIFIÉ (DMP)

- Document qui définit quelles données seront créées/collectées, comment elles seront structurées, documentées, partagées et conservées pendant et après le projet
- Initié très tôt, dès le montage du projet
- Opportunité pour établir un dialogue entre les différents acteurs : effort collectif (scientifiques, ingénieurs documentalistes, informaticiens, juristes, chargés de projet…)
- Document dynamique, continuellement mis à jour tout au long du projet
- Prend en compte toutes les étapes du cycle de vie des données
- Incitation voire obligation par les : agences de financement de la recherche, organismes de recherche
- Pas seulement un formulaire administratif

H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020 (2016).

DONNÉES DE LA RECHERCHE
LE PARTAGE… …DONNÉES COMMUNICABLES

- Données communicables sous conditions
  - Confidentialité des données à caractère personnel (cf. CNIL/ CIL du CNRS) => consentement, anonymisation des données avant de les partager
  - Droits de propriété intellectuelle : droit d'auteur, droit sui generis pour les bases de données => demander l'accord auprès des détenteurs de ces droits
  - Données de santé, données provenant d'un tiers privé, données statistiques, données relevant du potentiel scientifique et technique de la nation

- Données non communicables
  - Confidentialité des informations commerciales ou industrielles,
  - Secret de la défense nationale et sécurité publique

Becardet al. (décembre 2017). Ouverture des données de la recherche : Guide d'analyse du cadre juridique en France.

ENTREPÔTS DE DONNÉES

ENTREPÔTS DE DONNÉES
DÉFINITION

« Un entrepôt de données (Data repository, digital repository) est un réservoir constitué majoritairement de données de recherche, brutes ou élaborées, qui sont décrites par des métadonnées de façon à pouvoir être retrouvées. »
Données ouvertes de la recherche : nouvelles pratiques de publication et de partage. Cours Enssib (Aventurier P., 2013)

« Service en ligne permettant la collecte, la description, la conservation, la recherche et la diffusion des jeux de données »

ENTREPÔTS DE DONNÉES
REPÉRAGE… …DES LISTES ET DES ANNUAIRES

- Des annuaires d'archives en libre accès, contenant aussi des jeux de données
  - ROAR (Registry of Open Access Repositories)
  - OpenDOAR (Directory of Open Access Repositories) (178 entrepôts de données sur 3344 répertoriés)

- Des annuaires d'entrepôts multidisciplinaires
  - re3data (Registry of Research Data Repositories) (1859 entrepôts répertoriés) tous domaines
  - Open Access Directory (repris dans re3data)

- Des annuaires et listes spécialisés en sciences de la vie et sciences biomédicales
  - BioSharing (917 bases de données répertoriées)
  - NIH Data Sharing Repositories

- Des listes d'entrepôts recommandés par des éditeurs
  - Elsevier
  - Springer Nature
  - PLOS One

ENTREPÔTS DE DONNÉES
CLASSIFICATION DES ENTREPÔTS

- Disciplinaire: SEANOE, PANGAEA, Réseau Quetelet
- Multidisciplinaire : Zenodo (Europe), Figshare, Dryad
- Institutionnel : Edinburgh DataShare, Merritt, Enlighten: ResearchData, Open Data LMU (Allemagne)
- Spécifique d'un projet de recherche: Scientific Drilling Database
- Banque de données de référence, thématique : GenBank, UniProt

ENTREPÔTS DE DONNÉES
FONCTIONNALITÉS (1)

dépôt et conservation des données

- Téléchargement / import de fichiers :
  - Données + documents complémentaires
  - Limites : taille des fichiers, Formats acceptés, données publiées…
  - Via IU ou API
- Organisation des données en collections
- Conservation (stockage sécurisé + archivage à long terme)
  - Intégrité
    - sauvegarde
    - empreinte numérique
  - Lisibilité
    - veille sur l'évolution des technologies,
    - utilisation de formats durables (ou conversion), standards
  - Intelligibilité
    - métadonnées riches, standards
    - identification pérenne
    - contrôle des versions de jeux de données.

ENTREPÔTS DE DONNÉES
FONCTIONNALITÉ (1) : LE DÉPÔT DE DONNÉES…

- Dépôt métadonnées + données (+ si besoin : documentation, code…)

- Modalités de dépôt
  - Manuel (par le biais d'un formulaire en ligne, par mail…)
  - Par API
  - Echanges avec une équipe de curateurs, si elle existe

- Exemples
  - Dépôt manuel de données dans Zenodo
  - Par API
  - Partage de codes dans Zenodo
  - Tutoriel pour TreeBase

ENTREPÔTS DE DONNÉES
FONCTIONNALITÉS (2)

identification pérenne des données

- Attribution d'un identifiant au moment du dépôt
  - Au jeu dans son ensemble,
  - +/- à chaque version du jeu de données

- Intégration possible d'identifiant préexistant chez certains entrepôts (ex : Zenodo)

ENTREPÔTS DE DONNÉES
FONCTIONNALITÉS (3)

description des données

- Métadonnées
  - Schémas de métadonnées
    - génériques
    - spécifiques à un domaine
  - Qualité des métadonnées
    - vocabulaires contrôlés (import ou accès à des vocabulaires externes)
  - Saisie import via IU vs API
- Documentation complémentaire

ENTREPÔTS DE DONNÉES
FONCTIONNALITÉS (4)

contrôle des droits d'accès aux données, conditions d'utilisation et licences

- Droit d'accès aux données :
  - téléchargement libre (open)
  - embargo
  - demande d'accès (restricted), guestbook
  - pas de téléchargement possible (closed)
    - Génération d'URL privé parfois proposé

- Attribution d'une licence à chaque jeu de données
  - Saisie libre vs liste fermée
    - attention aux entrepôts qui imposent une licence unique

ENTREPÔTS DE DONNÉES
FONCTIONNALITÉS (5)

recherche, affichage, export des (méta)données

- Recherche
  - simple et/ou avancée : dans les métadonnées, parfois dans les données,
  - Graphique (zone géographique, structure molécule…)
  - Navigation et affinage par facettes

- Affichage
  - des métadonnées
    - Liens avec d'autres datasets, avec des publications
    - Génération de la citation
  - des données (prévisualisation, selon format)

- Exports
  - métadonnées (≠ formats),
  - téléchargement des données

- Génération de template de data papers à partir des métadonnées

ENTREPÔTS DE DONNÉES
FONCTIONNALITÉS (6)

exploration et visualisation des données

- Outils d'analyse ou de visualisation
  - TwoRavens,
  - WorldMap,
  - Mercury
  - …

ENTREPÔTS DE DONNÉES
FONCTIONNALITÉS (7)

découverte, visibilité des données

- Exposition des métadonnées via OAI-PMH ou dans un triple store RDF (Linkeddata)
- API pour la recherche et l'accès
- Liaison avec d'autres ressources :
  - Ex : IsCitedBy, Cites, … qui sont les valeurs possibles de la propriété relationType de la métadonnée relatedIdentifier
- Statistiques d'usage de ses données
  - nombre d'affichages et de téléchargements, datasets les + téléchargés, altmetrics, …

ENTREPÔTS DE DONNÉES
FONCTIONNALITÉS (8)

gestion des utilisateurs et des droits

- Authentification
  - Intégration de systèmes externe (fédération d'identifié, comptes ORCID, GitHub…)

- Gestion des utilisateurs et des droits
  - Gestion de rôles (administrateur de collection, curateur…)
  - Attribution de droits d'accès à des utilisateurs, à des groupes d'utilisateurs (collection, communautés, …)
  - Workflow de publication des données

ENTREPÔTS DE DONNÉES
CHOIX… …TENIR COMPTE DES RECOMMANDATIONS

- de la discipline dont relève la recherche
  - Genbank, UniProtKB, ProteinData Bank in Europe (PDBe)…etc

- des tutelles
  - Inra : code sources : forge logicielle SourceSup; données n'ayant pas vocation à aller dans un entrepôt disciplinaire : portail Data Inra

- des financeurs
  - Wellcome Trust Data repositories (13 entrepôts recommandés)
  - H2020 : entrepôt garantissant la gratuité de l'accès, de l'extraction, de l'exploitation, de la reproduction et de la dissémination.

- des revues pour les données qui seront publiées
  - Recommandent de + en + le dépôt dans un entrepôt disciplinaire (ou à défaut généraliste), voire intègrent un entrepôt dans le processus de publication (Nature + Figshare, Open Journal System + Dataverse…)
    - Nature : +90 entrepôts recommandés
    - CellPress: liste par type de données

…EN RÉSUMÉ

EN RÉSUMÉ
L'ENTREPÔT ET LE CYCLE DE VIE DU PROJET

- Avant le projet (DMP Opidor)
  - Réfléchir aux données produites, à leur diffusion et aux entrepôts concernés, aux coûts de préservation des données
  - Commencer à rédiger un plan de gestion des données

- Pendant le projet (Zenodo, Dataverse)
  - Déposer dans un entrepôt les données à partager avec des tiers identifiés (si peer review nécessaire par ex.)

- A la fin du projet (Pangaea, Cines)
  - Déposer en open access, dans un entrepôt, toutes les données qui peuvent l'être
  - Déposer les données à archiver sur une plateforme d'archivage

=> Le dépôt dans les entrepôt devient obligatoire / la norme

EN RÉSUMÉ
INFORMATIONS UTILES

- Des ressources INIST existent pour faciliter cette mise en place
  - Gestion et diffusion des données de la recherche (MC. Jacquemot-Perbal, A. CIOLEK-FIGIEL mars 2016 Formation URFIST Lyon)

  - Ressources DoRANum
    - Fiche synthétique sur les dépôts et les entrepôts
    - La minute entrepôt

  - DMP Opidor
  - Dataverse => Expérimentations, contacter l'INIST

  - Ingénierie documentaire

EN RÉSUMÉ
LIENS UTILES

- Validité du format pérenne d'un document :
  - FACILE -Service de validation de formats (CINES) https://facile.cines.fr/

- Licences :
  - Licence Ouverte etalabversion 2.0 https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf
  - Open Data Commons Open Database License (ODbL) https://opendatacommons.org/licenses/odbl/

- …et d'autres :
  - Data Seal of Approval
  - Data Literature Interlinking Service
  - Data2paper
  - Disciplinary Metadata (DCC)
  - Nakala https://www.nakala.fr/
  - NSF Arctic Data Center
  - European Framework for Audit and Certification of Digital Repositories
  - Sustainability of Digital Formats: Planning for Library of Congress Collections