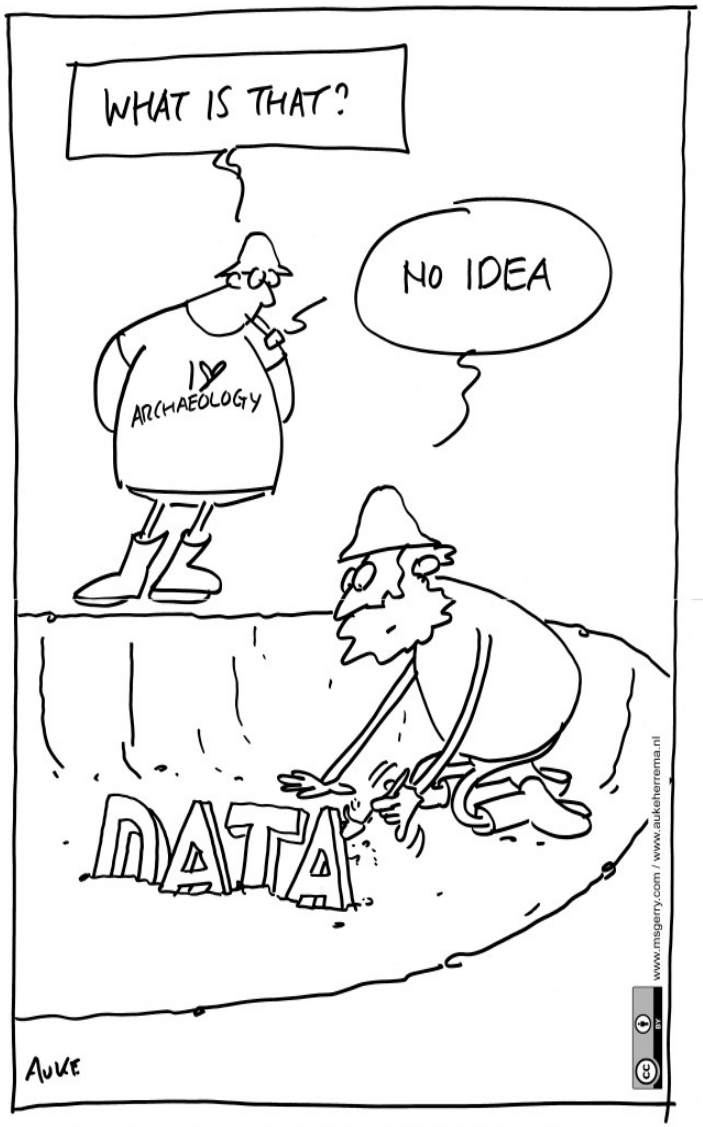
CELLULE DATA
GRENOBLE ALPES

Décrire et expliquer ses données



DATA FOR FUTURE GENERATIONS

Plan

Pourquoi décrire ?

Bonnes pratiques

Les standards

Outils

Pourquoi décrire ? (1)

1 - Pour soi
2 - Pour ses collègues (directeur, labo, partenaires etc.)
3- Pour tous

Intérêt :

Comprendre l'origine des données et leur contexte de création ou de collecte

Connaître les conditions de réutilisation et de partage des données

4- Pour les « machines » (moissonnage, interopérabilité)

A noter : quand les données sont non accessibles ou obsolètes, les informations descriptives peuvent rester disponibles

Pourquoi décrire ? (2)

Une histoire pas si rare

Data Sharing and Management Snafu in 3 Short Acts

New York University Health Sciences Library

https://www.youtube.com/watch?v=N2zK3sAtr-4

(voir vers 3mn05)

Bonnes pratiques (1)

Métadonnées

Définition : une donnée servant à définir ou décrire une autre donnée

2 types :

- les métadonnées embarquées (création automatique par les équipements) : données GPS, date, calibrage, etc.

- les métadonnées ajoutées par l'auteur : auteur, titre, description, mots-clés, laboratoire ou organisme, licence, etc.

Format d'échange : représentation numérique du standard et des métadonnées associées (convention ou encodage).

Deux **formats** courants :

- le XML
- le CSV.

Bonnes pratiques (2)

Quelques conseils

Quand décrire ?

Décrire au fur et à mesure !!!

>>> facilitera la réutilisation, la diffusion des données

Comment ? Prendre en compte :

- Les **caractéristiques** des données à décrire
  - leur nature (molécule, corpus, matériau, gène, enquête)
  - leur méthode d'acquisition (observation, expérimentation)
  - leur organisation
  - leur caractéristiques techniques (format, volume)
  - leur **potentiel de réutilisation**
- Les **pratiques de la discipline**,
- La **stratégie** de diffusion envisagée
- Les éventuelles **obligations** réglementaires ou financières (ex : programme européens, la directive INSPIRE)

>>> Une aide : utiliser les **formulaires de dépôts** dans les entrepôts

Bonnes pratiques (3)

Les **métadonnées importantes**
- Auteur (responsable du jeu)
  >>>> Identifiant (ORCID) / Affiliation (laboratoire, etc.)
- Titre
- Description/résumé
- Mots-clés
- Date
- Type
- Format
- Licence d'usage

Autres métadonnées utiles
- Contexte : projet de recherche (ex ANR)
- Lien avec les publications ou autres jeux de données (via des doi)
- le cas échéant : version, source, couverture, taille/volume, langue, contributeur, financement, etc.

Eléments recommandés :
- Identifiants uniques (de type doi)
- Fichier Read-me

Bonnes pratiques (4)

Un exemple : Diana Francis, Narendra Nelli, Ricardo Fonseca, Michael Weston, Cyrille Flamant, & Charfeddine Cherif. (2021). The radiative impact of the June 2020 historical Saharan dust storm [Data set]. Zenodo.http://doi.org/10.5281/zenodo.4572733

Ce qui manque :

- Pas de résumé/description
- Pas d'explication des protocoles
- Pas de fichier « read me »
- Pas d'unités pour les valeurs
- Utilisation de notations non définies

Ce qui est OK

- Doi
- Mots-clés
- Date
- Auteurs avec identifiant
- Licence d'utilisation

Bonnes pratiques (5)

Type	OK	Ambigu
Auteur	Claire Martin Orcid : 0000-0001-5340-8323 Affiliation : CEA, etc	Claire Martin
Format	Pdf 1.7	pdf
Lieu	Vienne (France ; cours d'eau)	Vienne
Vitesse	10 Mètres / seconde	rapide

Comment décrire une date ?

1 : 2020, le 5 mars

2 : le 5 mars 2020

3 : 5/03/2020

4 : 05/03/2020

5 : 2020-03-05

Les standards (1)

Pour vous aider :

Les schémas de métadonnées : liste structurée composée d'éléments descriptifs (champs) reliés entre eux.
- Nom du champ (Ex : titre, mots-clés)
- Type de contenu attendu : texte libre, nombre (par ex date au format AAAA-MM-JJ)
- Valeurs possibles (par ex, liste déroulante)

Les standards/normes : caractéristiques ou attributs communs pour décrire des contenus afin de permettre l'interopérabilité et une compréhension commune des éléments décrits.
Un standard est un schéma qui a été adopté comme modèle par un ensemble d'utilisateurs : il est reconnu, normalisé et utilisé à grande échelle.
Différents types de standards :
- généralistes
- par type de données
- par discipline

Les standards (2)

Standard **généraliste** : le Dublin Core
15 champs

- 1. Title (Titre)
- 2. Creator (Créateur)
- 3. Subject (Sujet)
- 4. Description (Description)
- 5. Publisher (Éditeur)
- 6. Contributor (Contributeur) : par exemple, photographe,
- 7. Date (Date) : format AAA-MM-JJ
- 8. Type (Type)
- 9. Format (Format)
- 10. Identifier (Identifiant de la ressource) : par ex, doi
- 11. Source (Source)
- 12. Language (Langue)
- 13. Relation (Relation) : par exemple, avec une publication
- 14. Coverage (Couverture) : géographique et temporelle
- 15. Rights (Gestion des droits) : licences

  >>>> Dublin Core étendu (par ex, audience, provenance, etc.)

Les standards (3)

- Standard **généraliste** : Datacite

ID	*Property*	*Obligation*
1	Identifier (with mandatory type sub-property)	M
2	Creator (with optional given name, family name, name identifier and affiliation sub-properties)	M
3	Title (with optional type sub-properties)	M
4	Publisher	M
5	PublicationYear	M
10	ResourceType (with mandatory general type description sub-property)	M

Table 2: DataCite Recommended and Optional Properties

ID	*Property*	*Obligation*
6	Subject (with scheme sub-property)	R
7	Contributor (with optional given name, family name, name identifier, and affiliation sub-properties)	R
8	Date (with type sub-property)	R
9	Language	O
11	AlternateIdentifier (with type sub-property)	O
12	RelatedIdentifier (with type and relation type sub-properties)	R
13	Size	O
14	Format	O
15	Version	O
16	Rights	O
17	Description (with type sub-property)	R
18	GeoLocation (with point, box, place, and polygon sub-properties)	R
19	FundingReference (with name, identifier, and award related sub-properties)	O
20	RelatedItem (with identifier, creator, title, publication year, volume, issue, number, page, publisher, edition, and contributor sub-properties)	O

Les standards (4)

Standards **disciplinaires** ou par type de données :

DDI (Data Documentation Initiative) : sciences sociales, comportementales et économiques.
MIDAS-Heritage : Architecture.
EML (Ecological Metadata Language): écologie
DwC (Darwin Core) : biodiversité.
PDBx/mmCIF (Protein Data Bank Exchange Dictionary and the Macromolecular Crystallographic Information Framework) : biologie

EAD (Encoded Archival Description) : description des archives.
EXIF (Exchangeable image file format) : description technique et automatique d'un cliché.
IPTC (International Press Telecommunications Council) : description d'une image par l'auteur.

Les standards (5)

Vocabulaires spécifiques :
Facilité la réutilisation des données
Mots-clés, classification taxonomiques, nomenclature des formules chimiques

Exemples :
Agriculture/agronomie : le vocabulaire contrôlé multilingue AGROVOC
Archéologie : le thésaurus PACTOLS
Environnement : le thésaurus GEMET (GEneral Multilingual Environmental Thesaurus) et le référentiel taxonomique TAXREF
Médecine : le thésaurus MeSH

Les standards (6)

Comment choisir ?

- Définir ses **objectifs**
  - Publication ?
  - Diffusion dans un entrepôt ?
  - Archivage ?

- Consulter ses collègues : directeurs de thèse, documentaliste, informaticiens
- Voir ce qui est utilisé dans les entrepôts de votre discipline
- Consulter les **répertoires de standards** :
  - Digial Curation Center
  - Research Data Alliance

Les outils (1)

Outils pour créer ses métadonnées

- DataCite Metadata Generator (xml)
- Modèle OTELo (csv)

Les outils (2)

Standards généralistes

Dublin Core

Guide de la Bibliothèque Nationale de France (BNF)

Datacite

Standards par disciplines

RDA metadata standard catalogue (Research Data Alliance)

>>> Index par sujets

Digital Curation Center :
« Disciplinary Metadata » Guidance

« Disciplinary Metadata » Guidance

>>>> Liste des outils d'édition de métadonnées

Web de données : les standards et le vocubulaire du W3C

Guide OTELo/INIST : guide pour la gestion des données (partie métadonnées)

Se former : DoraNum