CELLULE DATA
GRENOBLE ALPES

Décrire et expliquer ses données

DATA FOR FUTURE GENERATIONS

Plan

- **Pourquoi décrire ?**
- **Bonnes pratiques**
- **Les standards**
- **Outils**

Pourquoi décrire ?

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

Intérêt :

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

Comprendre ou retrouver le contenu des données

Les présenter, les réutiliser

Permettre leur réutilisation et leur reproduction

(pour vos lecteurs) connaître les conditions de réutilisation et de partage de vos données

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

Pourquoi décrire ?

- Une histoire pas si rare :

https://uquebec.libguides.com/gdr/videos#s-lib-ctab-16359068-0

ou

Data Sharing and Management Snafu in 3 Short Acts,
New York University Health Sciences Library,
https://www.youtube.com/watch?v=N2zK3sAtr-4

(à partir de 3mn05)

Bonnes pratiques

Votre avis ?

D’après vous,

- quels sont les éléments dont vous avez besoin pour utiliser des données ?
- quels sont les éléments importants à décrire pour vos données ?

Bonnes pratiques

Métadonnées

Une donnée servant à définir ou décrire une autre donnée

2 types :

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

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

Bonnes pratiques

Quelques conseils :

Quand décrire ?

Décrire au fur et à mesure !!! (votre thèse dure plusieurs années)

Votre contexte …

Prendre en compte votre contexte de travail

- Les **pratiques de votre discipline**
  Par ex, que constatez-vous dans les publications que vous lisez ?
- La **stratégie** de diffusion envisagée
  En discuter avec votre direction de thèse
- Les éventuelles **obligations** réglementaires ou financières

Bonnes pratiques

Comment ?

Décrire 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**

>>> facilitera la réutilisation (par vous, par d'autres) et la diffusion des données

Bonnes pratiques

Les **métadonnées importantes**

- Auteur (responsable du jeu)
  >>>>> identifiant (ORCID/affiliation)
- Titre
- Description/résumé
- Mots-clés
- Date
- Type
- Format
- Licence d’usage

- Identifiants uniques (de type doi)

- **Fichier Read-me**

Bonnes pratiques

Les **métadonnées importantes (suite)**

Autres métadonnées utiles :

- Contexte : projet de recherche, méthodologie

- **Lien avec les publications ou autres jeux de données ou codes**

- Le cas échéant :
  - Version
  - Source
  - Couverture temporelle ou spatiale
  - Taille/volume
  - Langue
  - Contributeur
  - Financement
  - etc.

Bonnes pratiques

Les métadonnées de fichiers

Nom du fichier

Chemin d'accès au fichier

Fichier de la provenance

Et Readme avec arborescence des fichiers

Bonnes pratiques

Un exemple : .

Yi-Kai Hsieh, Yoshiharu Omura, & Yuko Kubota. (2021). Energetic electron precipitation induced by oblique whistler mode chorus emissions [Data set]. Zenodo.

https://doi.org/10.5281/zenodo.5545963

Exercice

- Est-ce une bonne description ?
- Qu'est ce qui est bien décrit ?
- Qu'est-ce qui manque ?

Bonnes pratiques

- Yi-Kai Hsieh, Yoshiharu Omura, & Yuko Kubota. (2021). Energetic electron precipitation induced by oblique whistler mode chorus emissions [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5545963
-
- **Ce qui manque** :
- Résumé succinct
- Pas de description
- Pas de fichier « read me »
- Pas d'explication des paramètre
- Pas de mots clés

-
- **Ce qui est OK**
- Doi
- Date
- Auteurs avec identifiant
- Licence d'usage

Bonnes pratiques

• **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

Bonnes pratiques

Comment décrire une date ?

Que choisissez vous ?

1 : 2020, le 5 mars

2 : le 5 mars 2020

3 : 5/03/2020

4 : 05/03/2020

5 : 2020-03-05

Bonnes pratiques

Quelques conseils :

>>> Utiliser les **formulaires** dans les entrepôts

>>>> Des guides (par ex, sur Recherche Data Gouv)
- Guide de saisie des métadonnées générales
- Guide de saisie des métadonnées de fichiers
- Modèle de README

Les standards

Pour plus de cohérence, utiliser :

Les schémas de métadonnées (liste structurée d'éléments descriptifs – liste de champs)
Par ex :

Auteur : Nom, Prénom, identifiant
Date : AAAA-MM-JJ
Type : liste déroulante
Description : texte libre

Les standards/normes :
Schéma adopté comme modèle par une communauté

Reconnu, normalisé et utilisé à grande échelle.

Différents types de standards :

- généralistes
- par type de données
- par discipline

Les standards

Standard **généraliste** : le DublinCore

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor : par ex, photographe
7. Date : AAA-MM-JJ
8. Type
9. Format
10. Identifier : par ex, doi
11. Source
12. Language
13. Relation : par exemple, avec une publication
14. Coverage : géographique et temporelle
15. Rights : licences

Les standards

- Standard **généraliste** : Datacite

ID	*Property*	*Obligation*
1	Identifier	x
2	Creator	x
3	Title	x
4	Publisher	x
5	PublicationYear	x
10	ResourceType	x

Les standards

ID	Property
6	Subject
7	Contributor
8	Date
9	Language
11	AlternateIdentifier
12	RelatedIdentifier (pour relier les publications et les données : IsReferencedBy)
13	Size
14	Format
15	Version
16	Rights
17	Description

Les outils

- Outils pour créer ses métadonnées

- DataCiteMetadataGenerator (xml)
- ModèleOTELo (csv)

- Guide OTELo/INIST pour la gestion des données (voir métadonnées)

Les standards

Chercher l’inspiration ?

Allons voir le formulaire de dépôt sur Recheche Data Gouv

Les standards

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

DDI (Data Documentation Initiative) : sciences sociales, comportementales et économiques.
EML (EcologicalMetadataLanguage): sciences de l'environnement
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.
IPTC (International Press Telecommunications Council) : description d'une image par l'auteur.

Les standards

- **Intérêt** :
    - Enrichissement de la description
        - Ex : Aspirin
    - Désambigüiser

Trouver le standard de sa discipline

- RDA metadata standard catalogue (Research Data Alliance)
- >>> Index par sujets
-
- Digital Curation Center :
- « DisciplinaryMetadata » Guidance

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

Les standards

1AER

Download File | View File

DOMAIN III OF PSEUDOMONAS AERUGINOSA EXOTOXIN COMPLEXED WITH BETA-TAD

Li, M., Dyda, F., Benhar, I., Pastan, I., Davies, D.R.

(1996) Proc Natl Acad Sci U S A **93**: 6902-6906

Released	1996-06-10
Method	X-RAY DIFFRACTION 2.3 Å
Organisms	Pseudomonas aeruginosa
Macromolecule	EXOTOXIN A (protein)

1AG0

Download File | View File

STRUCTURE OF CYS 112 ASP AZURIN FROM PSEUDOMONAS AERUGINOSA

Faham, S., Rees, D.C.

To be published

Released	1997-10-29
Method	X-RAY DIFFRACTION 2.4 Å
Organisms	Pseudomonas aeruginosa
Macromolecule	AZURIN (protein)

1AKL

Download File | View File

ALKALINE PROTEASE FROM PSEUDOMONAS AERUGINOSA IFO3080

Miyatake, H., Hata, Y., Fujii, T., Hamada, K., Morihara, K., Katsube, Y.

(1995) J Biochem **118**: 474-479

Released	1996-03-08
Method	X-RAY DIFFRACTION 2 Å
Organisms	Pseudomonas aeruginosa
Macromolecule	ALKALINE PROTEASE (protein)

Collection Mode(s)

The method(s) used to collect the data.

- *audio computer-assisted self interview (ACASI)* -- Interview administered by the respondent, without an interviewer, assisted by a computer with audio prompts.
- *audiovisual touch-screen computer-assisted self-interview (AVT-CASI)* -- Interview administered by the respondent, without an interviewer, assisted by a touchscreen computer.
- *coded on-site observation* -- Observation that is conducted in a natural environment.
- *coded video observation* -- Observation that is conducted by video.
- *cognitive assessment test* -- Assessment of knowledge, skills, aptitude, or educational achievement by means of specialized measures or tests.
- *computer-assisted personal interview (CAPI)* -- Data collection method in which the interviewer reads questions to the respondents from the screen of a computer, laptop, or a mobile device like tablet or smartphone, and enters the answers in the same device. The administration of the interview is managed by a specifically designed program/application.
- *computer-assisted self-interview (CASI)* -- Respondents enter the responses into a computer (desktop, laptop, Palm/PDA, tablet, etc.) by themselves. The administration of the questionnaire is managed by a specifically designed program/application but there is no real-time data transfer as in CAWI, the answers are stored on the device used for the interview. The questionnaire may be fixed form or interactive. Includes VCASI (Video computer-assisted self-interviewing), ACASI (Audio computer-assisted self-interviewing) and TACASI (Telephone audio computer-assisted self-interviewing).

Les standards

Vocabulaires spécifiques :

Facilite la réutilisation des données

Mots-clés, classifications, nomenclature des formules chimiques

N'existe pas dans toutes les disciplines

Exemples :

Environnement :

Thésaurus GEMET (GEneral Multilingual Environmental Thesaurus)

Référentiel taxonomique TAXREF

Ex : Leontopodium nivalesubsp.alpinum

Médecine : Thésaurus MeSH

Ex : Asthmeàl'effort

Les standards

Comment choisir ?

- Consulter ses collègues : directeurs de thèse, documentaliste, informaticiens
- Vérifier ce qui est utilisé dans les entrepôts de votre discipline
- Consulter les **répertoires de standards**

Se former : DoraNum

Défi

A vous de jouer !

Essayez de décrire votre jeu de donnée en renseignant les éléments suivants :

- Auteur
- Titre
- Date
- Résumé/description
- Mots clés
- Type
- Format
- Licence d’usage
- **Read me**

Défi

- Bilan ...
- Des difficultés ?