



Décrire et expliquer ses données



DATA FOR FUTURE GENERATIONS



- **Bonnes pratiques**
- **Les standards**
- **Outils**

# Pourquoi décrire ? (rappel)



- 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)**



Vous avez regardé un ou deux jeux de données :

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



## 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.



## Quelques conseils :

### Quand décrire ?

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

Vérifier/compléter au moment du partage des données

### 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



## 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



## 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



## 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**



## 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.



## Les métadonnées de fichiers

Nom du fichier

Chemin d'accès au fichier

Et Readme avec arborescence des fichiers



## Point de vigilance : lever les ambiguïtés

•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



## Autre point de vigilance : normaliser

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



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



## 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



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

DDI (Data Documentation Initiative) : sciences sociales, comportementales et économiques.

EML (Ecological Metadata Language) : sciences de l'environnement

PDBx/mmCIF (Protein Data Bank Exchange Dictionary and the Macromolecular Crystallographic Information Framework) : biologie

EAD (Encoded Archival Description) : description des archives.

## **Trouver le standard de sa discipline**

RDA metadata standard catalogue (Research Data Alliance)

>>> Index par sujets

Digital Curation Center :

« Disciplinary Metadata » Guidance

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

# Les standards

## 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).

## 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)



## **Vocabulaires spécifiques :**

Facilite la réutilisation des données

Mots-clés, classifications, nomenclature des formules chimiques

N'existe pas dans toutes les disciplines

## **Intérêt :**

Enrichissement de la description

Ex : Aspirin

Désambigüiser

## **Exemples :**

### **Environnement :**

Thésaurus GEMET (GEneral Multilingual Environmental Thesaurus)

Référentiel taxonomique TAXREF

Ex : *Leontopodium nivalesubsp.alpinum*

### **Médecine :** Thésaurus MeSH

Ex : Asthme à l'effort



## Quelques conseils :

>>> Utiliser les **formulaire**s 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



**Chercher l'inspiration ?**

**Allons voir le formulaire de dépôt sur RechercheData Gouv**



- Outils pour créer ses métadonnées
- DataCiteMetadataGenerator (xml)
- ModèleOTELo (csv)

Pour aller plus loin

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



## 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