



**HAL**  
open science

## Anticiper et déposer les données de recherche : introduction aux principes FAIR

Emmanuelle Morlock

### ► To cite this version:

Emmanuelle Morlock. Anticiper et déposer les données de recherche : introduction aux principes FAIR. École thématique. Ecole d'été du Consortium DISTAM 2024, L'arbresle, Couvent de la Tourette, France. 2024. hal-04774928

HAL Id: hal-04774928

<https://hal.science/hal-04774928v1>

Submitted on 9 Nov 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Anticiper et déposer les données de recherche : introduction aux principes FAIR

Emmanuelle Morlock  
IR CNRS – HiSoMA, Biblissima+

École d'été du consortium DISTAM - 9 juillet 2024 - Couvent de la Tourette

# Qui parle ? IR au laboratoire HiSoMA - dir. adj. Biblissima+



PRÉSENTATION

Le laboratoire HISOMA mène des recherches sur les mondes anciens abordés à travers les disciplines spécialisées des Sciences de l'Antiquité que sont l'archéologie, l'histoire, les études littéraires, la philologie, l'épigraphie, la numismatique, l'iconographie, sur une très longue période qui s'étend de l'Ancien Empire pharaonique à la fin de l'Antiquité tardive et dans un espace géographique cohérent qui correspond principalement aux civilisations de la Méditerranée antique.

Les travaux du laboratoire sont largement reconnus à l'étranger et sur certains programmes, HISOMA est même tête de réseau d'une recherche internationale.

Le laboratoire HISOMA est l'un des laboratoires constitutifs de la Fédération de recherche « Maison de l'Orient et de la Méditerranée Jean-Pouilloux ».

Il regroupe environ 80 chercheurs, enseignants-chercheurs, post-doctorants, ingénieurs et administratifs, rattachés à l'une de ses cinq tutelles (CNRS, Université Lyon 2, Université Lyon 3, Université de Saint-Etienne, ENS de Lyon), 70 doctorants et une centaine de chercheurs associés.

Une lettre d'information trimestrielle présente les actualités, les manifestations scientifiques, les missions sur le terrain, les projets, les publications, les stages de formation et la vie du laboratoire...



≡ Biblissima<sup>+</sup>

PORTAIL BIBLISSIMA    ↓ Explorer    ↓ Rechercher    ↓ Visualiser    FR | EN

Tout Biblissima    Rechercher une œuvre, une personne, un lieu, une cote, une enluminure...    🔍

Bibliothèque virtuelle des bibliothèques, ce portail vous invite à découvrir l'histoire de'une partie des textes et livres qui ont été écrits, traduits, enluminés, collectionnés ou inventoriés depuis l'Antiquité jusqu'au XVIIIe siècle.

[A propos de ce portail](#)

<https://portail.biblissima.fr> / <https://projet.biblissima.fr>

**Biblissima<sup>+</sup>** 

Observatoire des cultures écrites  
de l'argile à l'imprimé

# Exemple de réalisation : le PGD de Biblissima+

**Biblissima!**  
Observatoire des cultures écrites  
de l'argile à l'imprimé

---

## Plan de gestion des données

Version Initiale (V1.07)

---

Tableau de suivi du document	
Titre	Plan de gestion des données du projet Biblissima+ V1 (version initiale à 6 mois)
Auteurs	Emmanuelle Mörlock, Régis Robineau, Edouard Frouzeau, Kevin Bois
Contributeurs	Anne-Marie Turcan-Verkerk, Marie-Agnès Avenel, François Bougard
Rellecteurs	Responsables scientifiques et techniques de livrables
Validé par	Anne-Marie Turcan-Verkerk (responsable scientifique et technique de Biblissima+)
Date de création	24 mars 2022
Type	Texte
Langage	fr-FR
Confidentialité	Publique
Statut	En cours

Biblissima+ Plan de Gestion des données - version initiale - avril 2022 1

**Biblissima!**  
Observatoire des cultures écrites  
de l'argile à l'imprimé

---

Plan de gestion de données de l'infrastructure  
Biblissima+, observatoire des cultures écrites  
anciennes de l'argile à l'imprimé (ANR-21-  
ESRE-0005)

---

Version 2.0  
Octobre 2023

[zenodo.org/records/10131101](https://zenodo.org/records/10131101)

**PGD de Biblissima+**

Gitlab

- Introduction >
- Le PGD de Biblissima+ (cadre de référence) >
- Vue d'ensemble des jeux de données >
- PGD détaillé (Périmètre 1) >
- Annexes >

estion des données (PGD) de l'observatoire des cultures  
édérant 16 établissements et une entreprise privée, réunis  
rée une infrastructure numérique multipolaire de  
service consacrée à l'histoire de la transmission des textes  
d'argile mésopotamiennes aux premiers livres imprimés,  
tes les langues. Biblissima+ concerne donc l'ensemble des  
nettant des textes anciens, y compris les sources  
onnaies, mais aussi les archives d'érudits modernes et de  
d elles apportent des informations originales sur les textes

<http://dmp.biblissima.fr>

# Plan

1. Pourquoi anticiper le dépôt des données de recherche ?
2. Définitions : (Méta)données, jeux de données, dépôts...
3. Introduction aux principes FAIR
4. Application sur un projet fictif

**Pourquoi anticiper ?**

# 1. Les données sont aussi importantes que les publications

“Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications”.

=> il faut être en mesure de les conserver en état d'être comprises et ré-utilisées

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles.  
Martone M. (ed.) San Diego CA: FORCE11; 2014



## 2. Les données numériques ont besoin d'une gestion continue et active

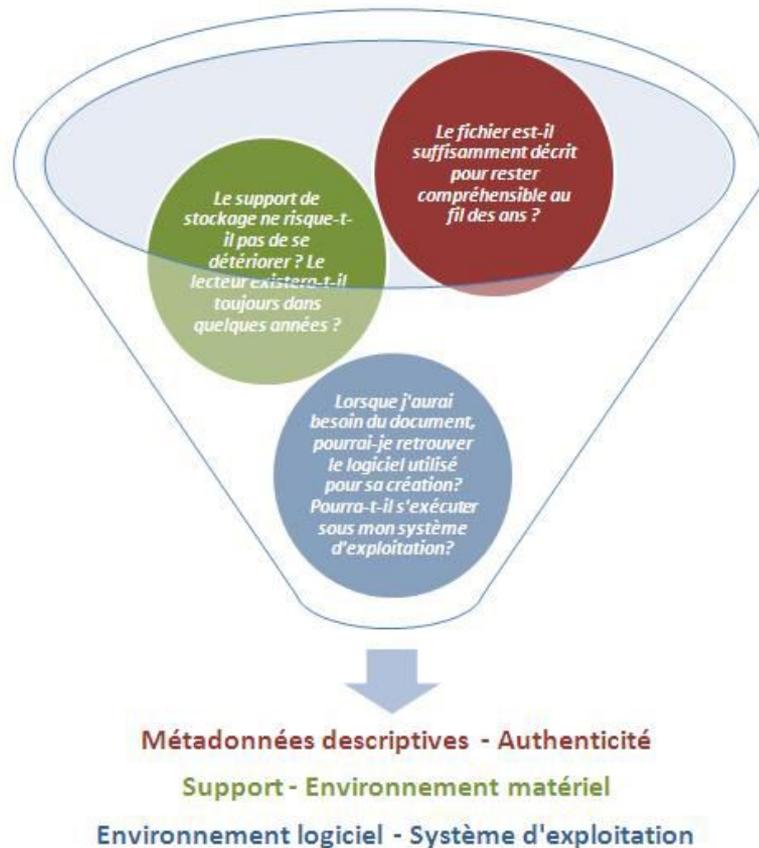
### Les 5 risques principaux

- Obsolescence (matériel, logiciels, formats)
- Vieillesse des supports
- Mauvaises manipulations
- Manque d'organisation
- Perte de la signification du contenu (documentation)

### Les solutions

- Copies et technologies de stockage multiples
- Veille technologique
- Privilégier les formats ouverts, les standards, les normes
- Conserver les métadonnées avec le document

Source : CINES



### 3. Pour s'inscrire dans la politique nationale de Science ouverte - A

2ème Plan national Science ouverte (2021-24)

#### Axe 2 : Structurer, partager et ouvrir les données de la recherche

Mettre en œuvre  
l'**obligation de diffusion  
des données de  
recherche** financées sur  
fonds publics.

Créer **Recherche Data  
Gouv**, la plateforme  
nationale fédérée des  
données de la recherche.

Promouvoir l'adoption  
d'une **politique de  
données sur l'ensemble  
du cycle des données de  
la recherche**, pour les  
rendre faciles à trouver,  
accessibles,  
interopérables et  
réutilisables (**FAIR**).

### 3. Pour s'inscrire dans la politique nationale de Science ouverte - B

2ème Plan national Science ouverte (2021-24)

#### **Axe 3 – Ouvrir et promouvoir les codes sources produits par la recherche**

Valoriser et soutenir la **diffusion sous licence libre** des codes sources issus de recherches financées sur fonds publics.

**Mettre en valeur la production des codes sources** de l'enseignement supérieur, de la recherche et de l'innovation.

Définir et promouvoir une **politique en matière de logiciels libres.**

# Deuxième Plan national pour **la science ouverte**

“Généraliser les pratiques de science ouverte en France”

“Faire de la science ouverte la pratique habituelle et quotidienne”

“Étendre le mouvement de partage des données, déjà généralisé en astronomie, en sismologie ou en génétique, aux autres disciplines”

Remarque : Il n’y a pas d’obligation au dépôt (mais une forte incitation...)

“Aussi ouvert que possible,  
pas plus fermé que nécessaire”

En pratique :

- Il faut se renseigner sur les politiques de Science ouverte de ses établissements de tutelle (cf. FAQ PGD de l’ANR)
- Il est demandé (implicitement) d’évaluer les possibilités d’ouverture de chaque jeu de données collecté ou produit.
  - En cas de non ouverture (contenus sous droits d’auteur, données sensibles ou données personnelles), il y a une justification à donner dans le dans le PGD.
  - Pas de “droit des données de recherche” : statut juridique parfois complexe

## 4. Il y a des coûts de gestion à financer / maîtriser

### Exemples

- Frais de personnel ou prestation de service
- Frais de stockage
- Frais de dépôt
- ...

### Dépense financière ou temps passé à “budgeter”

- le coût facturé par une plateforme ou un prestataire
- le temps de travail nécessaire pour préparer les données pour le partage ou la préservation

“On n’ouvre pas  
les données de la  
recherche  
comme on ouvrirait  
un robinet d’eau”

*Anne-Laure Sterin*

On ne rend pas  
les données  
compréhensibles  
par autrui  
ou réutilisables  
et en un clic...

# Quelques définitions

# Définitions - 1

## Données

L'ensemble des données générées, collectées, modifiées, ou enrichies au cours d'un processus de recherche.

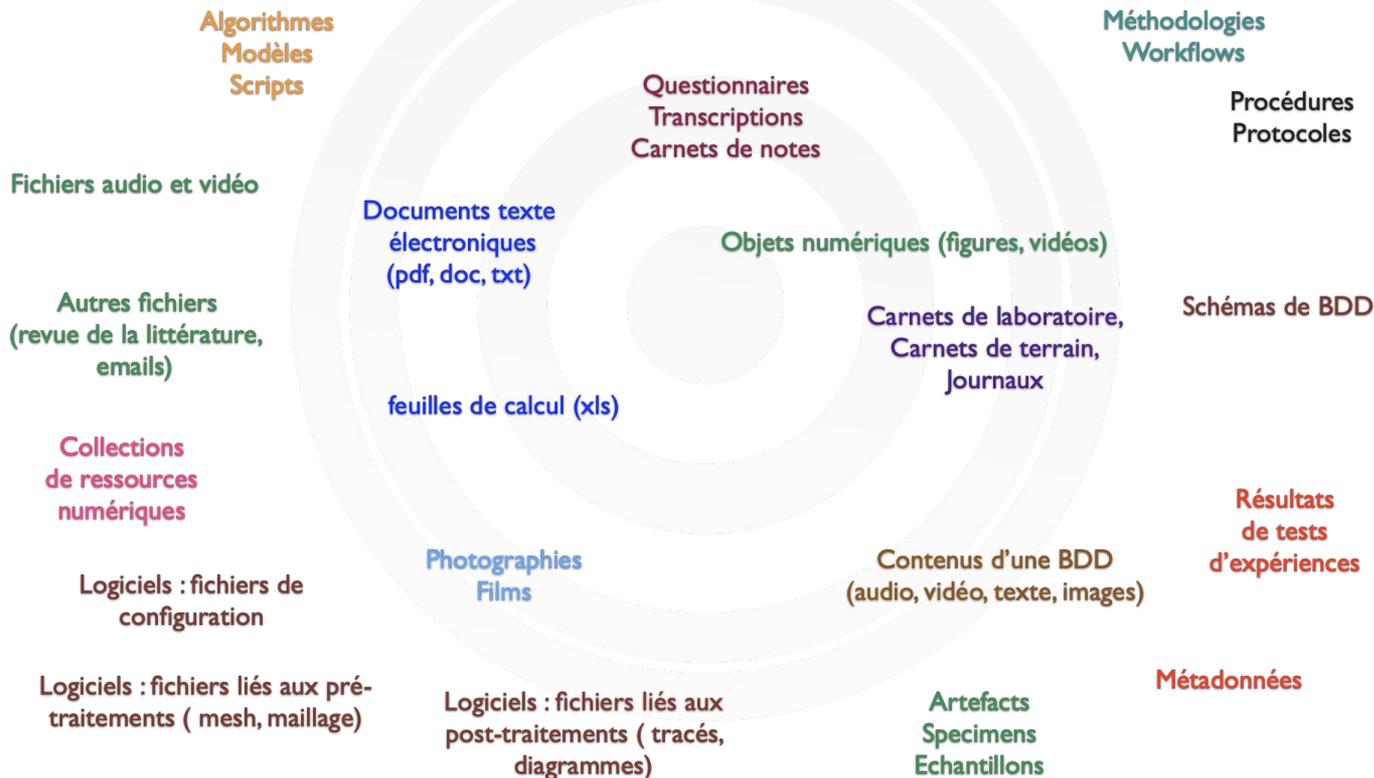
## Jeu de données

Une agrégation de données (brutes ou dérivées) regroupées pour former un ensemble cohérent. Concrètement : une collection de fichiers électroniques rassemblés selon des critères ou des niveaux de granularité laissés à l'appréciation des scientifiques.

## Métadonnées

Données servant à définir ou décrire d'autres données. Peuvent être administratives, descriptives, techniques etc.

L'ensemble des données  
générées, collectées, modifiées, ou enrichies au cours  
d'un processus de recherche



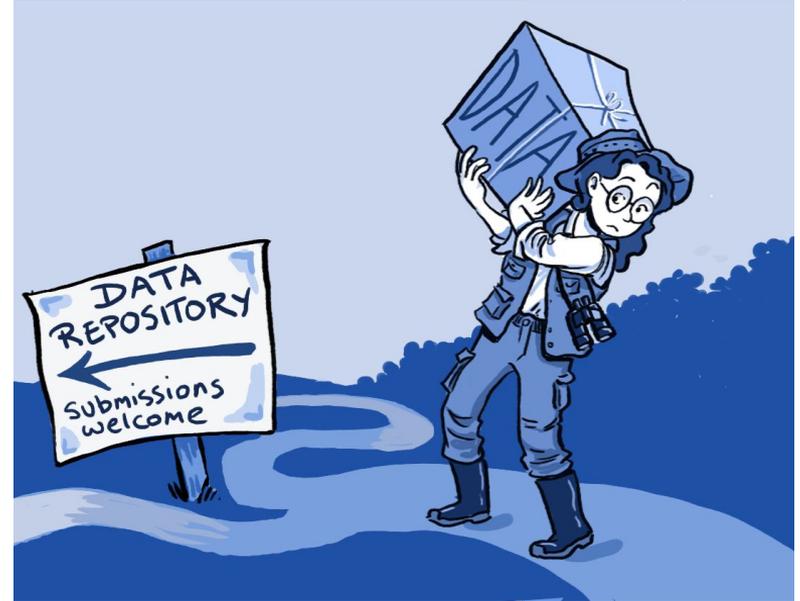
# Définitions - 2

## Dépôt

Démarche de diffusion et de préservation de jeux de données, dans une logique de “partage” ouvert facilitant la réutilisation.

Concrètement :

- Transfert d’un jeu de données dans un entrepôt (plateforme en ligne) avec renseignement d’un formulaire plus ou moins détaillé, constituant les métadonnées du dépôt.
- L’attribution d’identifiants uniques et pérenne de type **DOI** par l’entrepôt permet de citer le jeu de données (ou établir un lien solide avec des publications qui présentent les résultats obtenus avec ces données).



Crédit: Ainsley Seago. [doi:10.1371/journal.pbio.1001779.g001](https://doi.org/10.1371/journal.pbio.1001779.g001)

# Quoi déposer ?

- **Jeux de données**
  - validant des résultats
  - garantissant l'intégrité scientifique
  - valeur patrimoniale intrinsèque
- **Codes sources et logiciels**
- *Aussi : méthodes (si applicable)*

# Où déposer ?

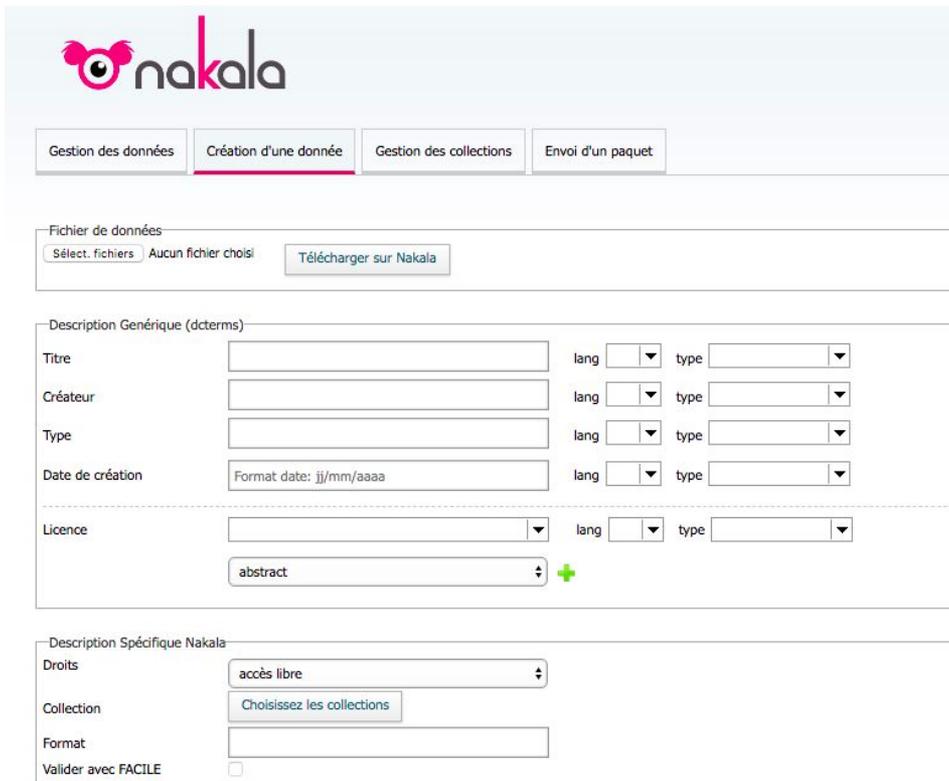
Entrepôt de référence  
dans votre discipline



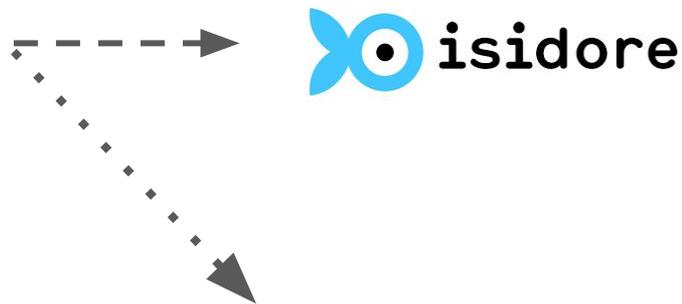
zenodo

nakala

# Nakala : identifiant pérenne, accès persistant, sécurisation



The screenshot shows the Nakala web interface. At the top left is the Nakala logo, which consists of a stylized pink eye icon followed by the word 'nakala' in lowercase. Below the logo is a navigation bar with four buttons: 'Gestion des données', 'Création d'une donnée' (which is highlighted with a pink underline), 'Gestion des collections', and 'Envoi d'un paquet'. Below the navigation bar is a section titled 'Fichier de données' containing a 'Sélect. fichiers' button, the text 'Aucun fichier choisi', and a 'Télécharger sur Nakala' button. The main form area is divided into two sections: 'Description Générique (dcterms)' and 'Description Spécifique Nakala'. The 'Description Générique' section contains four rows of input fields: 'Titre', 'Créateur', 'Type', and 'Date de création'. Each row has a text input field, a 'lang' dropdown menu, and a 'type' dropdown menu. The 'Date de création' field has a placeholder 'Format date: jj/mm/aaaa'. The 'Description Spécifique Nakala' section contains three rows: 'Droits' with a dropdown menu showing 'accès libre', 'Collection' with a button 'Choisissez les collections', and 'Format' with a text input field. At the bottom left of this section is a checkbox labeled 'Valider avec FACILE' which is currently unchecked.



Niveau avancé de préservation, en partenariat avec le CINES (sur demande).

zenodo   [Upload](#) [Communities](#) emmanuelle.morlock@mom.fr

## New upload

**instructions:** (i) Upload minimum one file or fill-in required fields (marked with a red star). (ii) Press "Save" to save your upload for editing later. (iii) When ready, press "Publish" to finalize and make your upload public.

Files

Drag and drop files here

— or —

(minimum 1 file required, max 50 GB per dataset - contact us for larger datasets)

Communities

Upload type

- Publication
- Poster
- Presentation
- Dataset
- Image
- Video/Audio
- Software
- Lesson
- Other

 Digital Object Identifier

Optional. Did your publisher already assign a DOI to your upload? If not, leave the field empty and we will register a new DOI for you. A DOI allows others to easily and unambiguously cite your upload. Please note that it is NOT possible to edit a Zenodo DOI once it has been registered by us, while it is always possible to edit a custom DOI.

 Publication date \*

Required. Format: YYYY-MM-DD. In case your upload was already published elsewhere, please use the date of first publication.

 Title \*

Required.

 Authors \*

Optional.

[+ Add another author](#)

 Description \*



Required.

 Version

Optional. Mostly relevant for software and dataset uploads. Any string will be accepted, but semantically-versioned tag is recommended. [See FAQ](#) for more information on semantic versioning.

# New upload

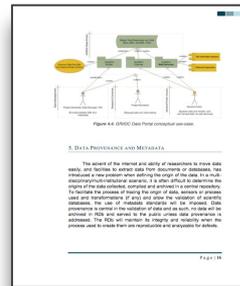
**Instructions:** (i) Upload minimum one file or fill-in required fields (marked with a red star ). (ii) Press "Save" to save your upload for editing later. (iii) When ready, press "Publish" to finalize and make your upload public.

Files >	<a href="#">Choose files</a>	<a href="#">Start upload</a>
Communities ⓘ	recommended	>
Upload type	required	>
Basic information	required	>
License	required	>
Funding	recommended	>
Related/alternate identifiers	recommended	>
Contributors	optional	>
References	optional	>
Journal	optional	>
Conference	optional	>
Book/Report/Chapter	optional	>
Thesis	optional	>
Subjects	optional	>
<a href="#">Delete</a>	<a href="#">Save</a>	<a href="#">Publish</a>

**Comment procéder ?**

# Anticiper via le PGD / DMP

Plan de gestion de données / Data Management Plan



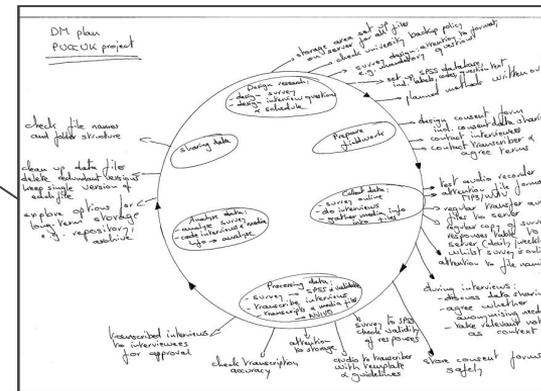
1

Inventaires questions audits

Politiques

Consultations

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
<b>Primary Data</b>				
Raw				
Processed				
Analyzed				
Finalized				
<b>Ancillary Data</b>				
Ancillary Data #1				
Ancillary Data #2				



2

Choix d'un modèle de PGD/DMP

Modèle "Science europe" recommandé par l'ANR

3

Saisie dans un outil collaboratif



Validation organisationnelle, technique, archivistique..

Révisions

4



5

PGD/DMP de fin de projet

# Etapes du dépôt d'un jeu de données

## 1. Préparation et documentation des données

- Sélection
- Structuration
- Vérifications (accord des co-auteurs, RGPD, vérifications scientifiques...)
- Préparation des données (nettoyage, renommage, exports vers des formats de diffusion, amélioration du niveau **FAIR**...)
- Préparation de la documentation jointe (fichiers README, dictionnaires de données, grilles d'analyses, etc.)

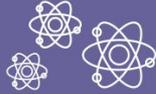
## 2. Choix de l'entrepôt

- Rassemblement des informations pour la fiche de métadonnées

## 3. Saisie des métadonnées / transfert des fichiers

## Comment lier des données à un article avant publication dans une revue scientifique ?

Vos données ne sont pas encore finalisées ?



Vous pouvez **pré-réserver** un PID !

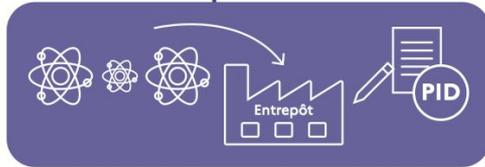
PID

1. Pré-réservez un PID pour vos données auprès de l'entrepôt de votre choix

2. Indiquez ce PID dans l'article à publier



3. Quand les données seront finalisées, n'oubliez pas de les déposer dans l'entrepôt de données !

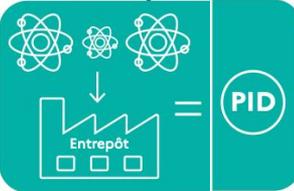


4. Et quand l'article sera publié, indiquez le PID article dans l'entrepôt où sont déposées les données si celui-ci le permet

Vos données sont finalisées ?



Vous pouvez les déposer dans un **entrepôt** !



1. Déposez vos données dans l'entrepôt de votre choix : un PID est attribué à vos données

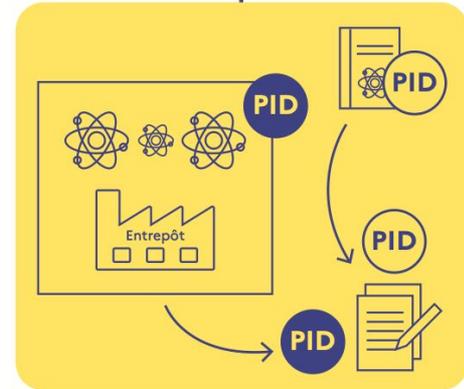
2. Indiquez ce PID dans l'article à publier



Vos données ont été valorisées par la publication d'un data paper ?

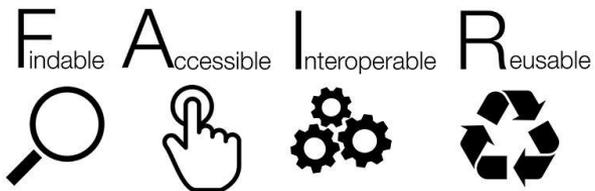


Vos données ont donc **déjà** été déposées dans un entrepôt !



Indiquez le PID attribué à vos données par l'entrepôt + le PID du Data paper dans l'article à publier

# Introduction aux principes FAIR



**Faciles à trouver**

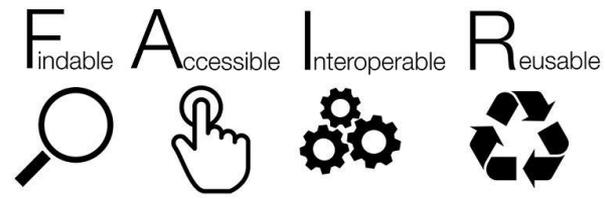
**Accessible**

**Interopérable**

**Réutilisable**

Principes génériques développés en **2014-2016** par un groupe de travail au sein de FORCE 11, une communauté internationale composée de chercheurs, d'éditeurs, de sociétés savantes, d'universités, de bibliothécaires, d'archivistes.

Rédigés sous forme de **recommandations**



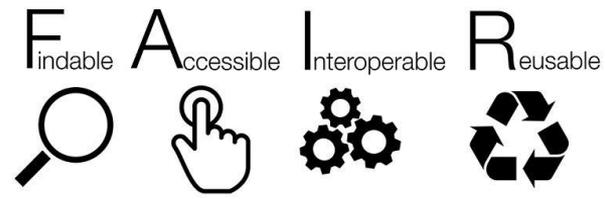
## Faciles à trouver

- Identifiants pérennes
- (Méta)données riches, recherchables et trouvables en ligne

## Accessible

## Interopérable

## Réutilisable



## Faciles à trouver

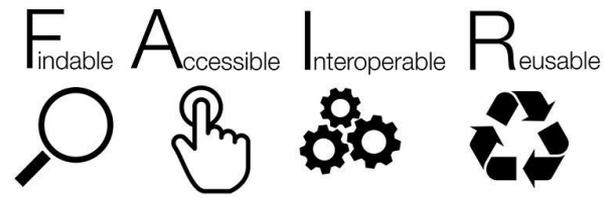
- Identifiants pérennes
- (Méta)données riches, recherchables et trouvables en ligne

## Accessible

- (Méta)données récupérables via un protocole de communication standardisé ouvert
- Authentification et restrictions possibles

## Interopérable

## Réutilisable



## Faciles à trouver

- Identifiants pérennes
- (Méta)données riches, recherchables et trouvables en ligne

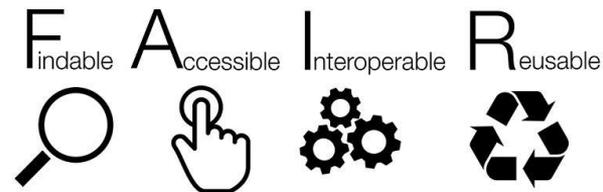
## Accessible

- (Méta)données récupérables via un protocole de communication standardisé ouvert
- Authentification et restrictions possibles

## Interopérable

- Formats communs et standards
- Vocabulaires contrôlés

## Réutilisable



## Faciles à trouver

- Identifiants pérennes
- (Méta)données riches, recherchables et trouvables en ligne

## Accessible

- (Méta)données récupérables via un protocole de communication standardisé ouvert
- Authentification et restrictions possibles

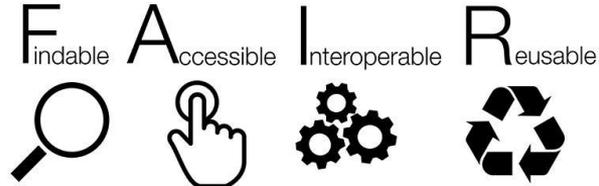
## Interopérable

- Formats communs et standards
- Vocabulaires contrôlés

## Réutilisable

- Données documentées (fichier readme)
- Licence d'utilisation explicite et lisible par machine
- Informations de provenance
- (Méta)données spécifiques à la discipline donnant le contexte nécessaire à la réutilisation scientifique

Les machines peuvent découvrir et utiliser les données automatiquement



## Faciles à trouver

- Identifiants pérennes
- (Méta)données riches, recherchables et trouvables en ligne

## Accessible

- (Méta)données récupérables via un protocole de communication standardisé ouvert
- Authentification et restrictions possibles

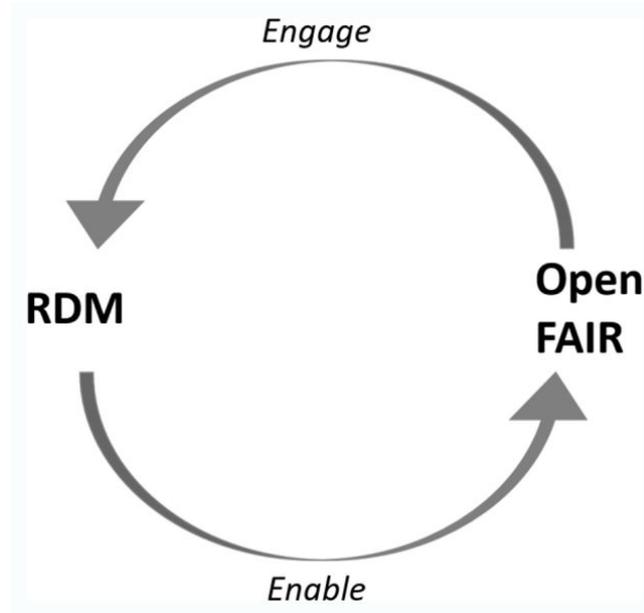
## Interopérable

- Formats communs et standards
- Vocabulaires contrôlés

## Réutilisable

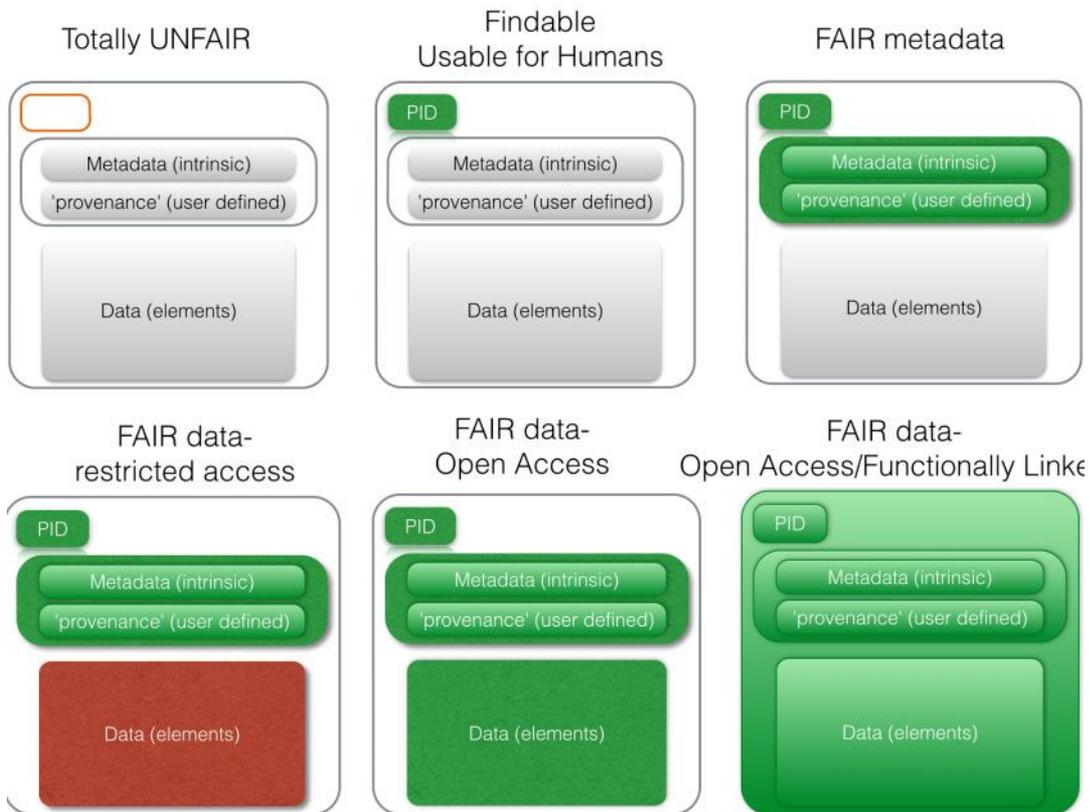
- Données documentées (fichier readme)
- Licence d'utilisation explicite et lisible par machine
- Informations de provenance
- (Méta)données spécifiques à la discipline donnant le contexte nécessaire à la réutilisation scientifique

# 3 concepts distincts : FAIR vs Open vs RDM



Source : Higman, R., Bangert, D., & Jones, S. (2019). Three camps, one destination: the intersections of research data management, FAIR and Open. *Insights*, 32(1), 18. DOI: <http://doi.org/10.1629/uksg.468>

# Progressivité du FAIR



5	<b>Managed Data Assets</b>	<b>Enterprise Level.</b> Data at this level is optimally managed at the most granular level in an environment offering <i>data governance</i> , <i>master data management</i> and <i>reference data management</i> capabilities.
4	<b>Semantically Typed Data</b>	<b>Cross-community Level.</b> This level focuses on cross-domain interoperability and is meant to be the level required for larger harmonization and integration projects.
3	<b>Standardised Data</b>	<b>Community Level.</b> Data at this level complies with community standard domain models, terminologies and formats, and is hosted in an environment offering searching and retrieval capabilities.
2	<b>Described Data</b>	<b>Project Level.</b> All datasets generated within a project are consistently described against a locally defined schema, controlled terminologies, and hosted in an environment offering data catalogue level searching capabilities.
1	<b>Identifiable Data</b>	<b>Data Object level.</b> Data at this level is identifiable as individual generic data objects and described by generic metadata elements. Hosting environment offers limited retrieval capabilities.
0	<b>Single Use Data</b>	No potential for re-use beyond lifetime of the research project



DoRANum. Données de la recherche : apprentissage numérique [En ligne]. France : DoRANum; 2019. Les principes FAIR [modifié le 04 décembre 2019 ; consulté le 07 juillet 2024].

Disponible : [https://doranum.fr/enjeux-benefices/principes-fair\\_10\\_13143\\_z7s6-ed26/](https://doranum.fr/enjeux-benefices/principes-fair_10_13143_z7s6-ed26/)

# How to comply with Horizon Europe mandate

## for Research Data Management



What are the requirements?

The FAIR principles

Developing a Data Management  
Plan (DMP)

Providing access to research data  
in trusted repositories

Validation (and re-use)  
requirements

Costs of Research Data  
Management

### What are the requirements?

Proper Research Data Management (RDM) is mandatory for any Horizon Europe project generating or reusing research data. It is a key part of Horizon Europe's open science requirements.

In Horizon Europe, *beneficiaries must manage the digital research data generated in the action ('data') responsibly, in line with the **FAIR principles***, and should at least do the following:

- Prepare a Data Management Plan (DMP) and keep it updated throughout the course of the project
- Deposit data in a trusted repository and provide open access to it ('as open as possible, as closed as necessary')
- Provide information (via the same repository) about any research output or any other tools and instruments needed to re-use or validate the data

Keep in mind that 'research data' is a very broad concept and certainly not limited to numerical/tabular data.

# How to make your data FAIR

Basic information with links to resources



Introduction

What is FAIR data?

Training materials

FAIR - in depth

How FAIR are your data?

More resources

## Introduction

Are you at the start of your project and planning to create research data? Read on to find out how to make it more findable, accessible, interoperable and reusable via the FAIR principles.

Why are the FAIR principles needed? The increasing availability of online resources means that data need to be created with longevity in mind. Providing other researchers with access to your data facilitates knowledge discovery and improves research transparency.

In this context, during the Lorentz Workshop "[Jointly Designing a Data FAIRport](#)" (2014), participants formulated the FAIR data vision to optimise data sharing and reuse by humans and machines, which resulted in the publication of [The FAIR Guiding Principles for scientific data management and stewardship](#), published in "[Scientific Data](#)".

The FAIR principles describe how research outputs should be organised so they can be more easily accessed, understood, exchanged and reused. Major funding bodies, including the European Commission, promote FAIR data to maximise the integrity and impact of their research investment.

# Etude de cas

# Projet fictif : une édition de sources anciennes sur le web

Données

Images d'objets  
ex. statues égyptiennes

Données

Transcriptions en XML /TEI  
(schéma EpiDoc)

Codes sources

Plateforme de gestion et  
d'indexation XML :  
eXist-db  
+  
Système de publication :  
TEI Publisher

# Inventaire

Nom	But de la collecte / création	Type	Format / standard	Provenance	Taille / volume
Images des statues	Vérification des transcriptions et publication en vis à vis de la transcription	images 2D couleur et N&B	Tiff / Jpeg	Institut XXX	100 images / 4 To
Transcriptions	Publication en ligne	Texte	XML/TEI (EpiDoc)	Créés dans le cadre du projet	100 Mo
Application web	Site web dédié pour l'édition en ligne	Codes	XQuery - CSS - javascript	Base générée par TEI Publisher développée et enrichie par l'ingénieure web	500 Mo

# Quels objectifs/stratégies de dépôts / valorisation ?

Images des statues

Stockage ?

Publication ?

Dépôt ?

fichiers TEI

Stockage ?

Publication ?

Dépôt ?

Codes TEI Publisher

Stockage ?

Publication ?

Dépôt ?

# Quels objectifs/stratégies de dépôts / valorisation ?

Images des statues

**Stockage** dans un cloud  
Sharedocs-Huma-Num

**Publication** dans l'application  
Web via un futur Serveur IIIF

**Pas de dépôt**

fichiers TEI

**Stockage/Partage** via un  
repository Gitlab Huma-Num

**Publication** dans l'application  
web

**Partage/préservation** via  
Nakala (collection dédiée) -  
Licence Etalab 2.0

Codes TEI Publisher

**Stockage/Partage** via un  
repository Github

**Publication** sur une instance  
exist-db via Huma-Num

**Partage/préservation** via  
Zenodo (automatisé pour via  
un versionnage) : application  
web seule - Licence GPLv3

# Repérage FAIR simplifié avec le mini-questionnaire des Fair Implementation Profiles (FIP)

## ~ FIP mini-questionnaire ~ Build your FAIR Implementation Profile

Community description	
Name of Community	e.g. ENVRI
Description of Community	
Supporting Links	
Research Domain	e.g. Environmental Sciences
Data Steward	e.g. ORCID #
Date of FIP creation	

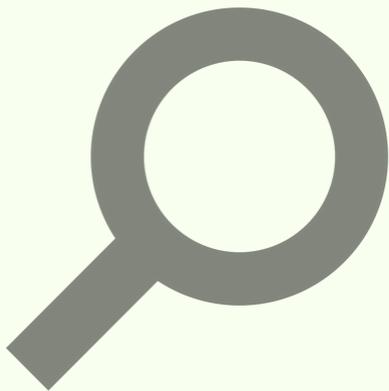
FAIR principle	Question	FAIR enabling resource types	Your answers
F1	What globally unique, persistent, resolvable identifiers do you use for metadata records?	Identifier type	e.g. PURL, DOI
F1	What globally unique, persistent, resolvable identifiers do you use for datasets?	Identifier type	
F2	Which metadata schemas do you use for findability?	Metadata schema	
F3	What is the technology that links the persistent identifiers of your data to the metadata description?	Metadata-Data linking mechanism	
F4	In which search engines are your metadata records indexed?	Search engines	
F4	In which search engines are your datasets indexed?	Search engines	
A1.1	Which standardized communication protocol do you use for metadata records?	Communication protocol	
A1.1	Which standardized communication protocol do you use for datasets?	Communication protocol	
A1.2	Which authentication & authorisation technique do you use for metadata records?	Authentication & authorisation technique	
A1.2	Which authentication & authorisation technique do you use for datasets?	Authentication & authorisation technique	
A2	Which metadata longevity plan do you use?	Metadata longevity	
I1	Which knowledge representation languages (allowing machine interoperation) do you use for metadata records?	Knowledge representation language	
I1	Which knowledge representation languages (allowing machine interoperation) do you use for datasets?	Knowledge representation language	
I2	Which structured vocabularies do you use to annotate your metadata records?	Structured vocabularies	
I2	Which structured vocabularies do you use to encode your datasets?	Structured vocabularies	
I3	Which models, schema(s) do you use for your metadata records?	Metadata schema	
I3	Which models, schema(s) do you use for your datasets?	Data schema	
R1.1	Which usage license do you use for your metadata records?	Data usage license	
R1.1	Which usage license do you use for your datasets?	Data usage license	
R1.2	Which metadata schemas do you use for describing the provenance of your metadata records?	Provenance model	
R1.2	Which metadata schemas do you use for describing the provenance of your datasets?	Provenance model	

<https://bit.ly/yourFIP>

# FINDABLE

RICH DESCRIPTIVE  
(META)DATA

---



- How will they be discovered?
- What is necessary to make explicit so they can be understood outside the team?

- F1. (Meta)data are assigned a globally unique and **persistent identifier**
- F2. Data are described with **rich metadata** (defined by R1 below)
- F3. Metadata clearly and explicitly **include the identifier** of the data they describe
- F4. (Meta)data are registered or **indexed in a searchable resource**

# Profil FIP simplifié pour le critère “Facile à trouver”

Nom	F1 - Identifiant pérenne (PID)	F2 - Quels schémas de métadonnées pour la découverte ?	F3 - Technologie pour lier les métadonnées aux données ?	F4 - Quels outils de découverte indexent vos données
Images des statues	L'institut n'a pas attribué d'identifiant pérenne. La diffusion via un serveur IIIF à l'étude pourrait en fournir	N/A	A discuter avec l'institut selon les modalités de gestion de leur catalogue	N/A
Transcriptions	Identifiant pour chaque texte dans l'application web (url) et identifiant pour le jeu de données dans Nakala	Schéma utilisé par l'entrepôt Nakala : Dublin Core	N/A ou géré par Nakala	Isidore (via la création d'une collection Nakala)
Application web	Identifiant générique dans Zenodo et identifiant pour chaque version donnant lieu à dépôt	Schéma utilisé par Zenodo : Datacite	N/A ou géré par Zenodo	Indexation automatique par Software Heritage et OpenAire

# ACCESSIBLE

EXPLICIT ACCESS PROTOCOLS  
FOR BOTH HUMANS AND  
MACHINES

---



- Where the data will be found?
- Are they openly and freely accessible or are the limitations clear?
- A1. (Meta)data are retrievable by their identifier using a **standardised communications protocol**
  - A1.1 The **protocol is open**, free, and universally implementable
  - A1.2 The protocol allows for an **authentication and authorisation procedure, where necessary**
- A2. **Metadata are accessible**, even when the data are no longer available

# Profil FIP simplifié pour le critère “Accessible”

Nom	A1.1 - Quel protocole de communication standard pour les données ?	A1.2 - Quelle méthode d'authentification et autorisation pour les données	A2 - Quel plan à long terme est prévu pour les métadonnées ?
Images des statues	https (application web)	Sous la responsabilité de l'institut - images publiées sans authentification sur le site web avec l'accord de l'institut	Sous la responsabilité de l'Institut
Transcriptions	https (application web et dépôt)	Données accessibles en Open data complet via l'application web et le dépôt Nakala	Aucun, tributaire de la politique de Nakala.
Application web	https (application web et dépôt)	Données accessibles en Open data complet via le dépôt Zenodo	Aucun, tributaire de la politique de Zenodo

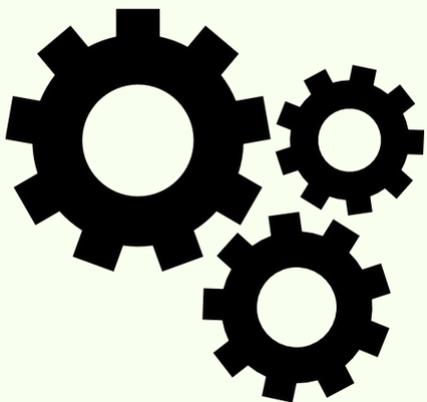
# Profil FIP simplifié pour le critère “Réutilisable”

Nom	R1.1 - Licence de réutilisation pour les données	R1.2 - Quels schémas de métadonnées de provenance utilisés pour les jeux de données
Images des statues	Sous la responsabilité de l'institut : Etalab 2.0	Aucun
Transcriptions	Etalab 2.0	Modèle EpiDoc (balise sourceDesc et descendants)
Application web	GPLv3	Métadonnées Zenodo de dépôt (champs Datacite)

# INTEROPERABLE

INTEGRATED WITH OTHER  
DATA & INTEROPERATE WITH  
APPLICATIONS

---



- What are the formal languages in use?
- Is the data intelinked to other (meta)data?
- Is the data machine-actionable?

- I1. (Meta)data use a **formal**, accessible, shared, and broadly applicable **language for knowledge representation**.
- I2. (Meta)data use **vocabularies** that follow FAIR principles
- I3. (Meta)data include qualified **references** to other (meta)data

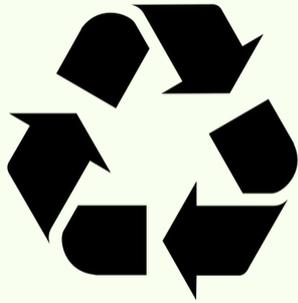
# Profil FIP simplifié pour le critère “Interopérable”

Nom	I1 - Quel langage de représentation de connaissance (exploitable par les machines) utilisez-vous pour les jeux de données	I2 - Quels vocabulaires contrôlés utilisez-vous pour encoder/indexer vos jeux de données	I3 - quels modèles de données pour les données ?
Images des statues	Aucun	Thésaurus iconographique Biblissima (SKOS)	N/A actuellement - à venir : manifestes IIIF
Transcriptions	Balisage XML	Pleiades, Typologie des statues spécifique au projet)	TEI / EpiDoc
Application web	Aucun	N/A	N/A

# REUSABLE

## OPTIMISE THE REUSE OF THE DATA

---



- How will they be discovered?
  - Are reuse conditions clear?
  - What is necessary to make explicit so they can be understood outside the team?
- 
- R1. Meta(data) are **richly described** with a plurality of accurate and relevant attributes
    - R1.1. (Meta)data are released with a clear and accessible **data usage license**
    - R1.2. (Meta)data are associated with detailed **provenance**
    - R1.3. (Meta)data meet **domain-relevant community standards**

# Récap et conclusion

# À Retenir

FAIR :

- des principes à très haut niveau de généralité orientés sur la réutilisation des données.
- Une interprétation est nécessaire
- L'implémentation dans les pratiques de recherche en est encore à ses débuts
- En SHS des questions spécifiques se posent (cf. Arnauld Ginglold <https://roadtofair.hypotheses.org/788> (sept 2023))

et non pas :

- un standard
- un état statique
- nécessairement Open
- FAIR requiert des procédures pour la machine : le fait que des humains puissent découvrir des données, y accéder, les reformater ou les réutiliser n'est pas suffisant
- FAIR n'est pas un passe-temps de documentalistes fanatiques mais un niveau de qualité que l'on s'efforce d'atteindre en fonction d'une stratégie de dissémination / diffusion

# Conclusion

- Dépôt dans un entrepôt “de confiance” ou certifié : 1er niveau de FAIR
- Anticiper les dépôts donne une vision claire d’objectifs concrets, qui facilite la rédaction du PGD et l’organisation quotidienne
- Le PGD est un outil d’aide à la décision
  - Enjeux des arbitrages à faire pour développer une “Sobrité numériques”
- La démarche et le workflow peuvent être adaptés à d’autres étapes
  - Par exemple :
    - Dépôts de fichiers liés à une publication ancienne,
    - Préparation d’un départ à la retraite,
    - Mise en ordre rétrospective et partage de données liées à un master, un doctorat...





<b>Data Stage</b>	<b>Output</b>	<b># of Files / Typical Size</b>	<b>Format</b>	<b>Other / Notes</b>
<b>Primary Data</b>				
Raw				
Processed				
Analyzed				
Finalized				
<b>Ancillary Data</b>				
Ancillary Data #1				
Ancillary Data #2				

Modèle d'inventaire de données – Data curation profiles - Purdue university Library

# La vision des archivistes...

« 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 preuve par les chercheurs et leurs équipes. À ce titre, elles constituent **une partie des archives de la recherche** ».

# A finalité probante...

« Data are representations of observations, objects, or other entities **used as evidence** of phenomena for the purposes of research or scholarship. »

C.L. Borgman (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press

« Data is not in and of itself a kind of evidence but a **multifaced object** which can be **mobilized as evidence** in support of an argument ».

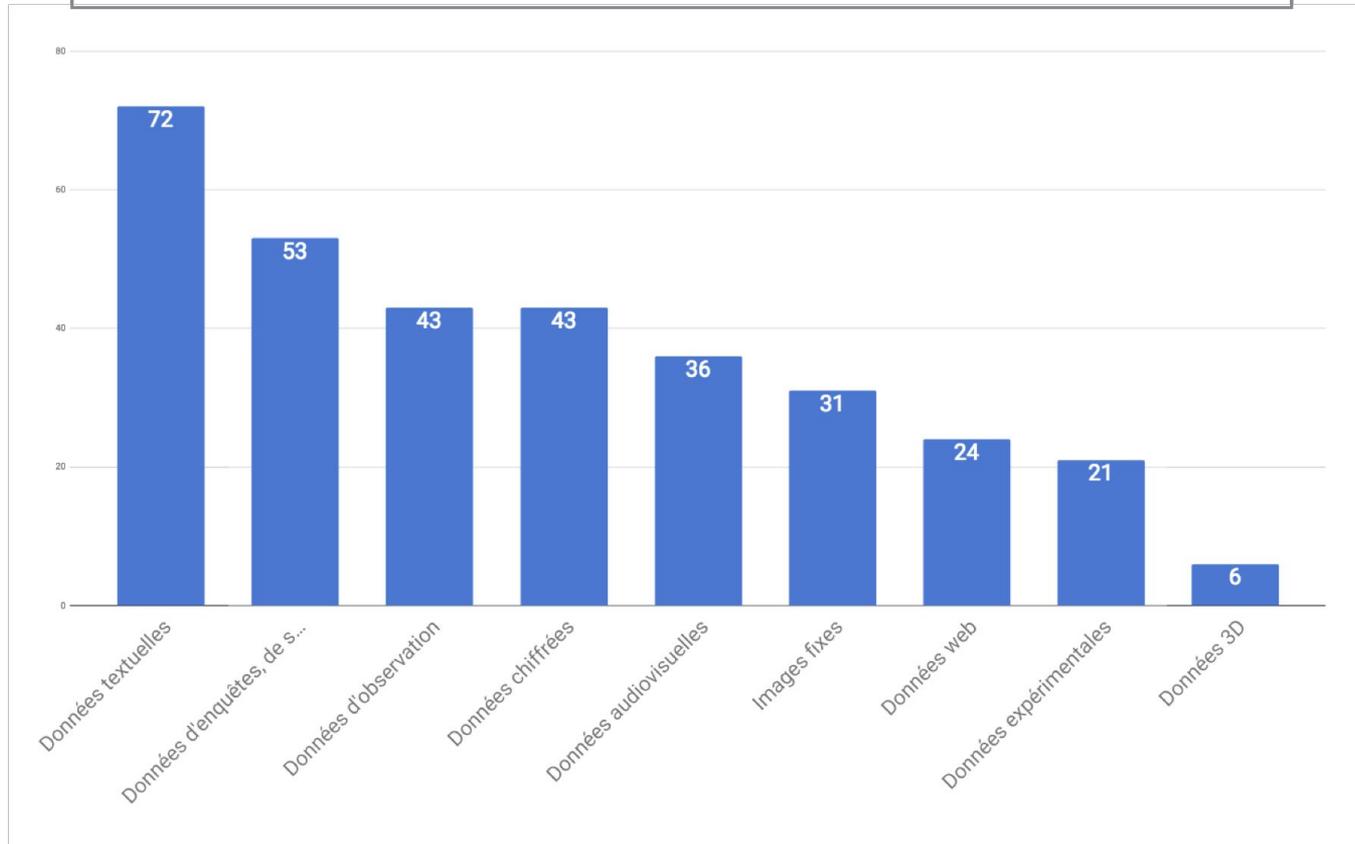
Trevor Owens, « Defining Data for Humanists: Text, Artifact, Information or Evidence? », *Journal of Digital Humanities*, Table of Contents for Vol. 1, No. 1 Winter 2011

## Données de la recherche

« machine actionnable »

“A digital, selectively constructed, machine-actionable abstraction representing some aspects of a given object of humanistic inquiry”

# Catégories de sources de données



Alexandre Serres, Marie-Laure Malingre, Morgane Mignon, Cécile Pierre, Didier Collet.  
Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs :  
une enquête à l'Université Rennes 2 : Rapport (hal-01635186)



Image credit : « **FAIR principles** » by Australian National Data Service – ANDS is licensed under **CC BY 4.0** and include icons made by **Freepik** from [www.flaticon.com](http://www.flaticon.com) are licensed by **CC 3.0 BY**

DoRANum. Données de la recherche : apprentissage numérique [En ligne]. France : DoRANum; 2019. Les principes FAIR [modifié le 04 décembre 2019 ; consulté le 07 juillet 2024]. Disponible : [https://doranum.fr/enjeux-benefices/principes-fair\\_10\\_13143\\_z7s6-ed26/](https://doranum.fr/enjeux-benefices/principes-fair_10_13143_z7s6-ed26/)

2007 - FP7

**OA Pilot Publications**  
L'ERC recommande l'OA  
pour les données dès  
début du projet



2014 - Horizon 2020

**OA obligatoire** pour les publications + **OA data pilot**



2017 - Horizon 2020

**OA par défaut** pour les publications et  
les données

2018 - EOSC

**Focus** sur la Science ouverte



sept 2018

initiative "Science Europe" pour  
**accélérer la transition**  
vers l'OA complet



2021 - Horizon Europe

La Science ouverte est le **modus operandi** (Open data par défaut)

**Stratégie européenne  
pour les données**

Faire en sorte que l'UE devienne  
un modèle et un acteur majeur  
d'une société dont les moyens  
d'action sont renforcés par les  
données

# Critères de choix d'un entrepôt

Si le choix de l'entrepôt ne vous est pas imposé par le financeur, l'institution ou la revue, quelques questions à se poser :

- disciplinaire ?
- certifié ?
- données hébergées en France ?
- archivage à long terme ?
- coûts ?
- types de données acceptées ?
- formats ?
- modifications possibles ?
- attribution d'un identifiant pérenne ?
- licence imposée ?
- restrictions d'accès ?
- embargos ?
- statistiques d'utilisation ?

**“A digital, selectively  
constructed, machine-actionable  
abstraction representing some  
aspects of a given object of  
humanistic inquiry.”**

Schöch, Christoph. “Big? Smart? Clean? Messy?” Data in the Humanities”.  
*Journal of Digital Humanities*. Vol. 2, n°3, 2013

# Classement selon un continuum



## Données brutes

Données recueillies mais non traitées, sans organisation ni mise en forme. *Ex. données d'observation de phénomènes particuliers.*



## Données traitées

Données produites après traitement de données brutes (calibration/étalonnage ou correction).



## Données dérivées

Présentent un résumé ou une re-présentation spécifique des données, par exemple sous une forme canonique (agrégation, compilation, calcul, réorganisation, reformulation).



## Données analysées

Données à l'étape de l'examen critique par les chercheurs pour en tirer des informations ou réponses à leur questions de recherche.



## Données finalisées

Dernière étape avant dépôt/archivage, où il n'y a plus de traitements / manipulations par les chercheurs.



## Données publiées

Données intégrées aux publications.

# Classement selon des usages

## Données primaires

Données recueillies mais non traitées, sans organisation ni mise en forme.  
*Ex. données d'observation de phénomènes particuliers.*

## Données secondaires

Données à l'étape de l'examen critique par les chercheurs pour en tirer des informations ou réponses à leur questions de recherche.

## Données auxiliaires

Données produites après traitement de données brutes (calibration/étalonnage ou correction).

## Données sources

Présentent un résumé ou une représentation spécifique des données, par exemple sous une forme canonique (agrégation, compilation, calcul, réorganisation, reformulation).

## Données résultats

Données intégrées aux publications.

# Classement selon le mode de production

## Données d'observation

- Capturées en temps réel.
- Souvent uniques et non reproductibles

*Relevés météo, clichés astronomiques, fouilles archéologiques.*

## Données expérimentales

- Obtenues à partir d'équipements de laboratoires
- Potentiellement reproductibles mais de manière coûteuse

*Chromatogrammes, poids biomasse, séquence peptide.*

## Données computationnelles

- Générées à partir de modèles informatiques
- Reproductibles si le modèle est documenté

*Modèle climatique, modèle économique.*

## Données dérivées

- Issues du traitement, de la compilation ou de la réorganisation de données brutes
- Souvent coûteuses à reproduire

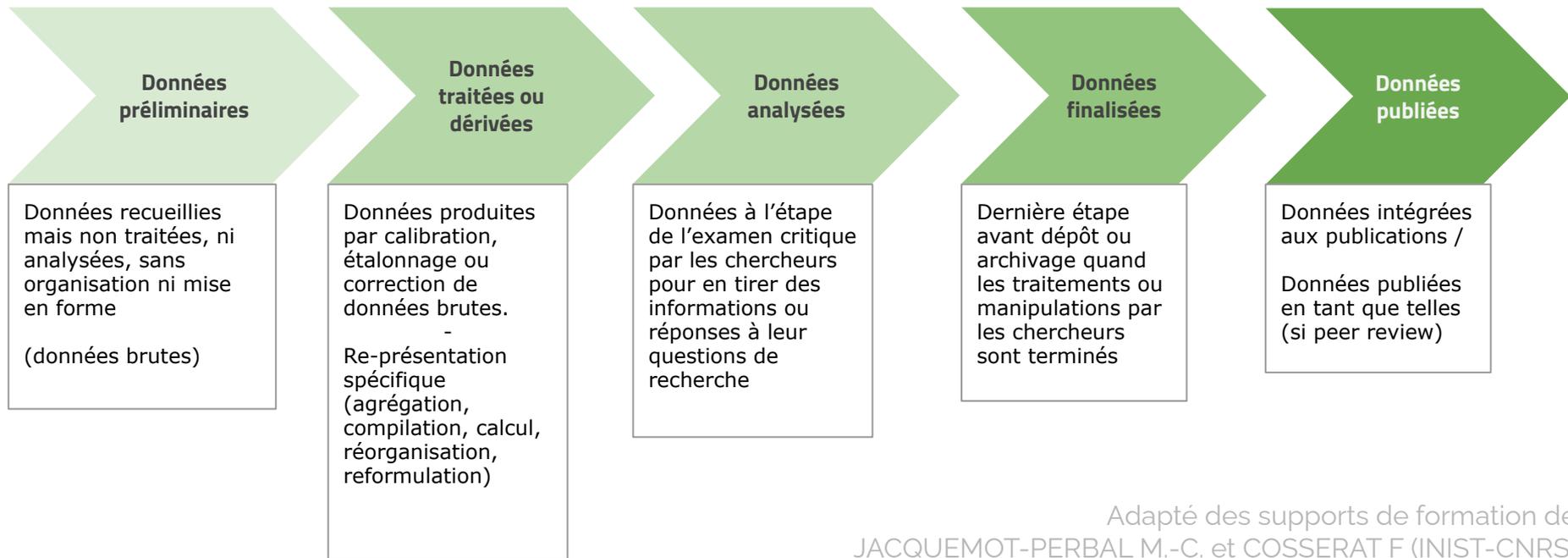
*Bases de données compilées, fouilles de textes.*

## Données de référence canoniques

- Agglomération statique ou dynamique de données, souvent éditorialistes et publiées.

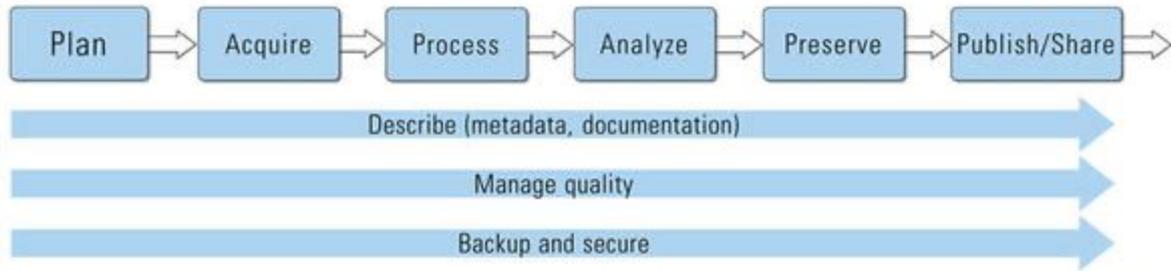
*Ex. : Séquences de gènes, correspondances, collections d'archives photographiques.*

# Typologie : un continuum d'états tout au long du cycle de vie



# **Le dépôt**

**répondre aux exigences des financeurs,  
augmenter la visibilité,  
permettre la validation**



The USGS Data Lifecycle produced by the U.S. Geological Survey

<https://www.usgs.gov/data-management/data-lifecycle>

# Findable

The data has sufficiently rich metadata and a unique and persistent identifier to be easily discovered by others. This includes assigning a [persistent identifier](#) (like a [DOI](#) or [Handle](#)), having rich [metadata](#) to describe the data and making sure it is findable through disciplinary local or international discovery portals.

To aid automatic discovery of relevant datasets

# Accessible

The data is retrievable by humans and machines through [a standardised communication protocol](#), with authentication and authorisation where necessary. The data does not necessarily have to be open. Data can be [sensitive](#) due to privacy concerns, national security or commercial interests. When it's not able to be open, there should be clarity and transparency around the conditions governing access and reuse.

Explicit limitations on the use of data, protocols for querying, downloading or copying data are made explicit for both humans and machines

# Interoperable

The associated data and metadata uses a 'formal, accessible, shared, and broadly applicable language for knowledge representation'. This involves using community accepted languages, [formats](#) and [vocabularies](#) in the data and metadata. Metadata should reference and describe relationships to other data, metadata and information through [identifiers](#).

(Meta)data should use standards, have references to other (meta)data and be machine actionable

# Reusable

The associated metadata provides rich and accurate information, and the data comes with a clear usage licence and detailed provenance information. Reusable data should maintain its initial richness. For example, it should not be diminished for the purpose of explaining the findings in one particular publication. It needs a clear machine readable [licence](#) and [provenance](#) information on how the data was formed. It should also use discipline-specific data and metadata standards to give it rich contextual information that will allow reuse.

(Meta)data are sufficiently well described for humans and computers to be able to understand them and have a clear and accessible data usage licence

# références

synthèse rédigée sur le site <https://ardc.edu.au/resource/fair-data/> d'après **Turning FAIR into reality**

Final report and action plan from the European Commission expert group on FAIR data

<https://op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en>

# ACCESSIBLE

EXPLICIT ACCESS PROTOCOLS  
FOR BOTH HUMANS AND  
MACHINES

---

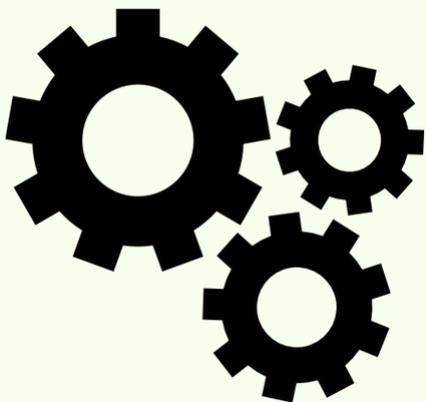


- Where the data will be found?
- Are they openly and freely accessible or are the limitations clear?
- A1. (Meta)data are retrievable by their identifier using a **standardised communications protocol**
  - A1.1 The **protocol is open**, free, and universally implementable
  - A1.2 The protocol allows for an **authentication and authorisation procedure, where necessary**
- A2. **Metadata are accessible**, even when the data are no longer available

# INTEROPERABLE

INTEGRATED WITH OTHER  
DATA & INTEROPERATE WITH  
APPLICATIONS

---



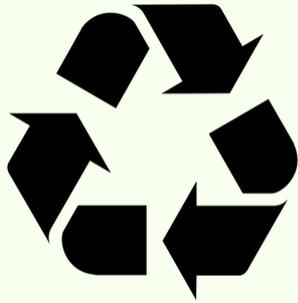
- What are the formal languages in use?
- Is the data intelinked to other (meta)data?
- Is the data machine-actionable?

- I1. (Meta)data use a **formal**, accessible, shared, and broadly applicable **language for knowledge representation**.
- I2. (Meta)data use **vocabularies** that follow FAIR principles
- I3. (Meta)data include qualified **references** to other (meta)data

# REUSABLE

## OPTIMISE THE REUSE OF THE DATA

---



- How will they be discovered?
  - Are reuse conditions clear?
  - What is necessary to make explicit so they can be understood outside the team?
- 
- R1. Meta(data) are **richly described** with a plurality of accurate and relevant attributes
    - R1.1. (Meta)data are released with a clear and accessible **data usage license**
    - R1.2. (Meta)data are associated with detailed **provenance**
    - R1.3. (Meta)data meet **domain-relevant community standards**