Session 4 International and open access strategy for the future

A trip to the "data management" wonderland

Cécile Pignol - Engineer
Sedimentory Plateform & Archives collections
Data management group (LTER-Fr -BED) *Bancariser Ensemble les Données (Give access to Data together)*

13:30-16:00

Fondamental in Data management - F.A.I.R. Principles

Data management (Samples & data registering strategies)

Summary :

PART I- INTODUCTION

***

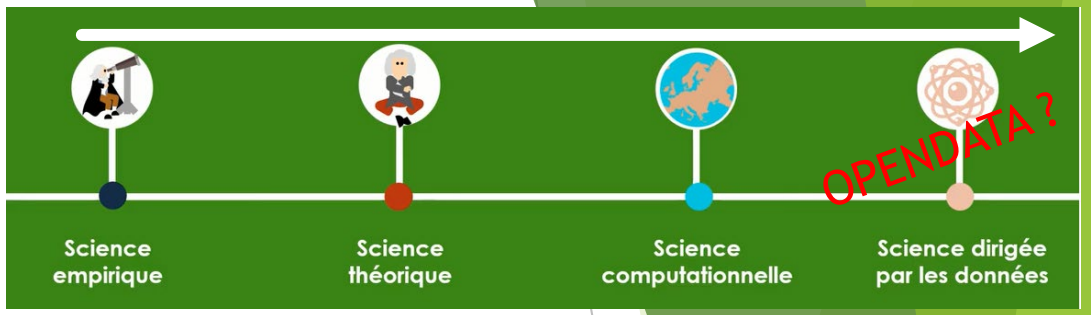
1- Issues & beneficts of data sharing

2- Process for collecting scientific data and publications

3- What are the brakes in 2021 ?

4- Learning path / Self-training courses

BENEFICITS OF DATA SHARING



FOR RESEARCHERS

TRANSCRIPTION OF NEW REQUIREMENT & LAWS

Research data from publicly funded must be openly disseminated

➔ Européen H2020 Projets (DataMmangementPlan),

➔ In France, the Plan national pour la science ouverte (some of which aim to structure and open research data) Research Agency ANR (DMP)

OPPORTUNITIES

- increase the visibility & re-use of the work of researchers

FOR SCIENCE

NEW PERSPECTIVES

Data sharing allows researchers to build on a larger data foundation and paves the way for ***new methods of investigation***.

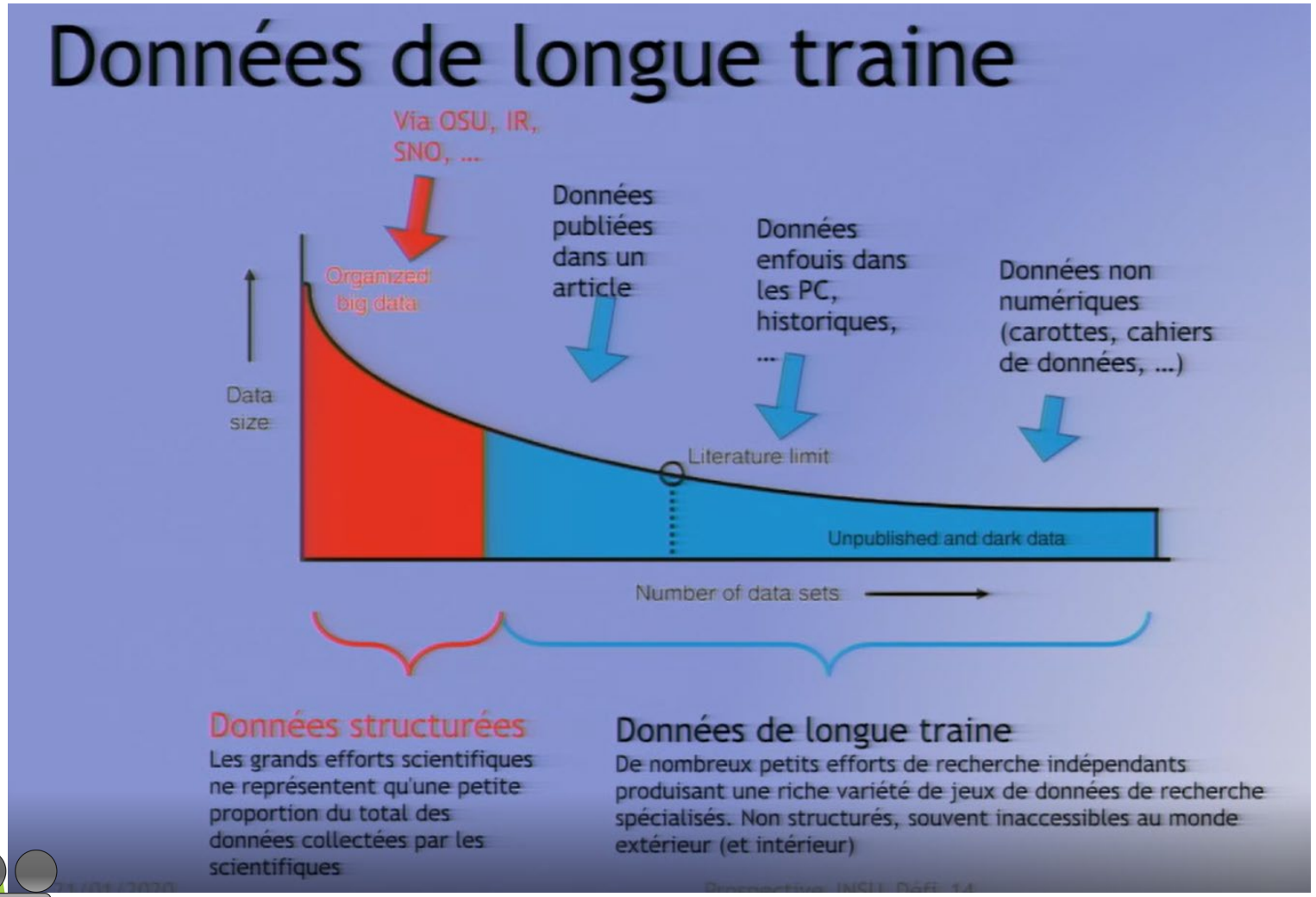
EXAMPLE: Use reference databases, approach the same dataset in different ways…

STONGER VALIDATION - QUALITY - REPRODUCTIBILITY

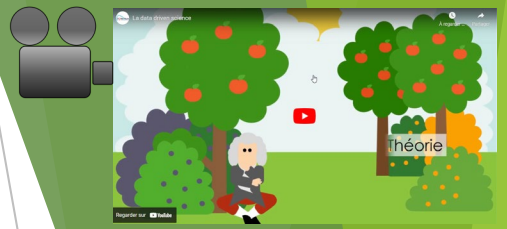
- conditions for validating articles by access to data
- Filing of data is increasingly requested by the reading committees.
- Re-use of the data by other researchers may lead to reconsideration of the initial results.

https://doranum.fr/enjeux-benefices/fiche-synthetique/

« LONG TAIL OF DATA »



Vidéo 2:30 :



Accessible structured data (Big data, Environmental observatories, Ecology, etc.)

Vs

Unstructured data, **+/- available** for some time, to researchers around a publication or, available on request

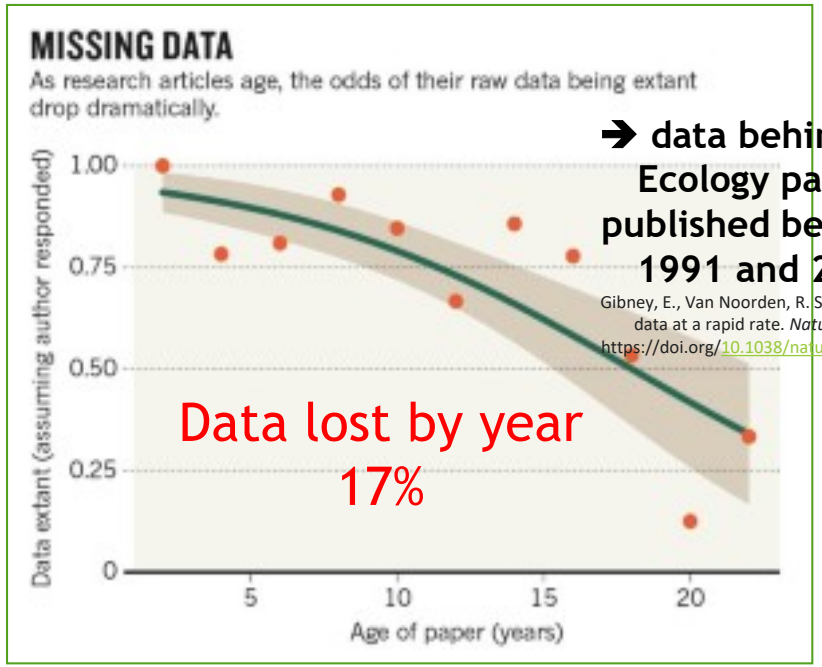
Non-Digital Data (Samples, Paper Lab Book,…)

➔ Talk about Smalls & Darks Data !

Video (in french) à 2:30 : S Galle, V Stoll, Données de longue traine, accès aux données attachées aux publications, données hors SNO, données non numériques, Atelier défi 14 de la prospective transverse INSU, Du 20 janv. 2020 au 21 janvier 2020

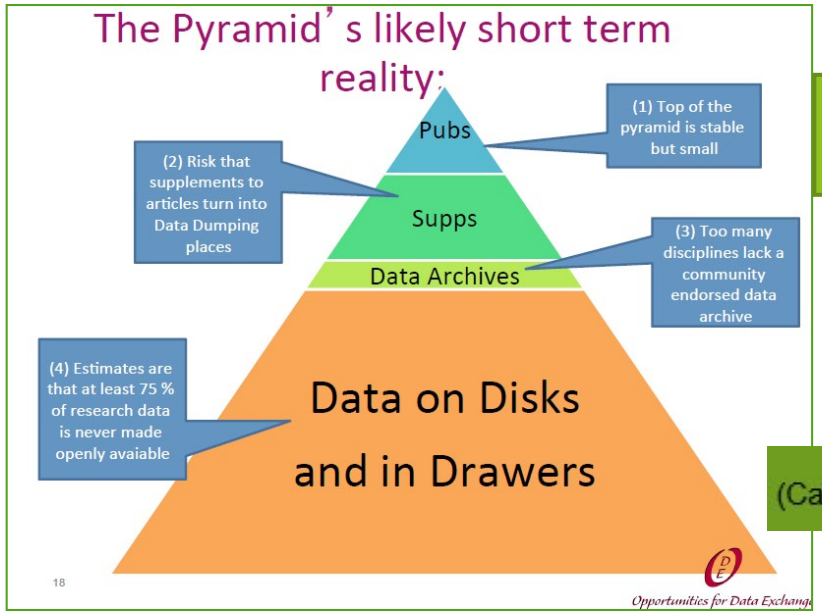
Ferguson, A., Nielson, J., Cragin, M. *et al.* Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nat Neurosci* **17,** 1442–1447 (2014). https://doi.org/10.1038/nn.3838

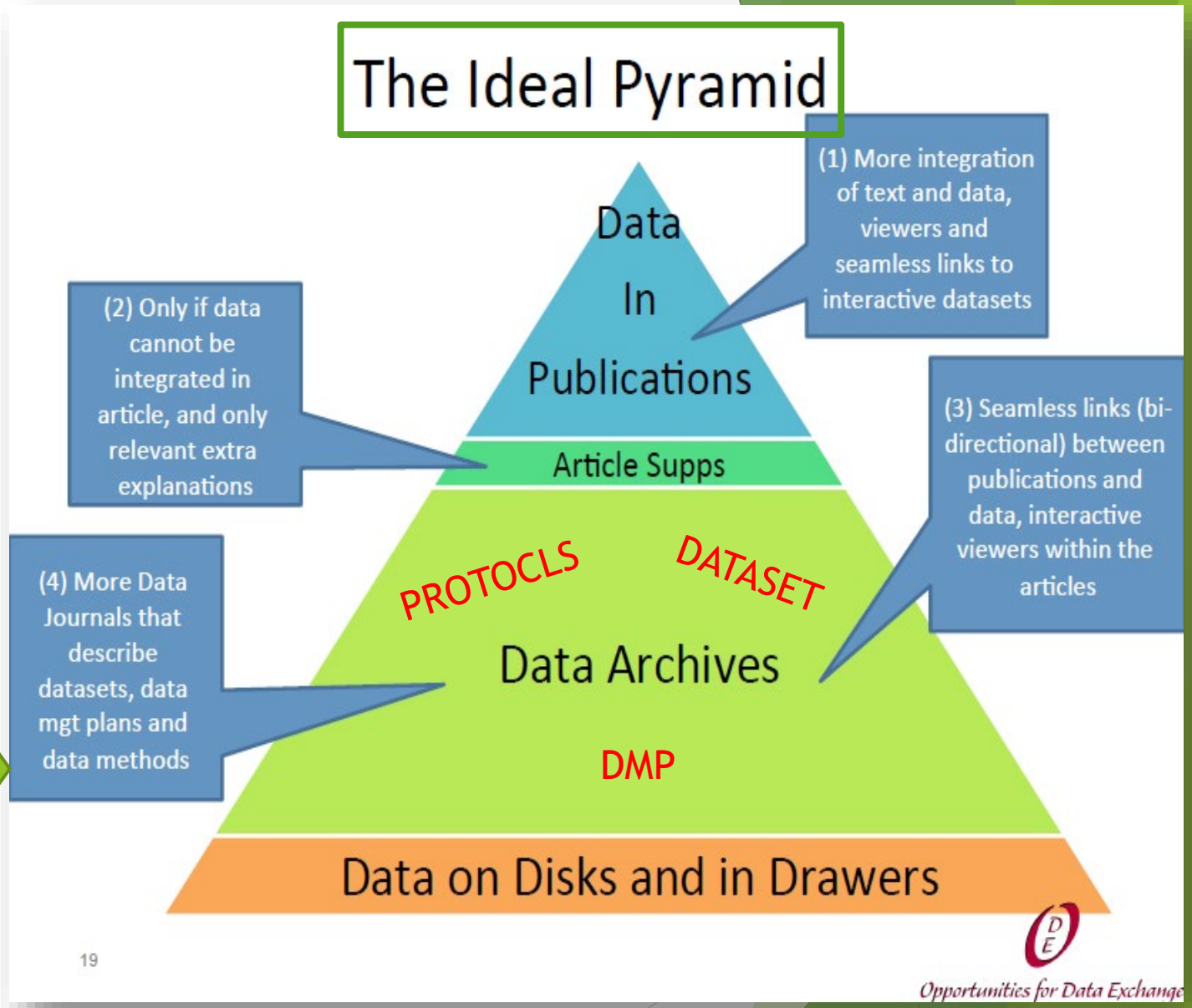
The reality !



➔ **data behind 516 Ecology papers published between 1991 and 2011**

Gibney, E., Van Noorden, R. Scientists losing data at a rapid rate. *Nature* (2013). https://doi.org/10.1038/nature.2013.14416





(Callaghan, 2013) d'après : (Reilly, Schallier, Schrimpf, Smit, & Wilkinson, 2011)

Process for collecting scientific data and publications (1)

< 200?

Collecte de données
- Données instrumentales brutes (in-situ)
- Données instrumentales brutes (ex-situ)

Processus Traitements
- Données traitées
- Données agrégées

Article scientifique

Publication d'un article
+ « Supplementary material »
(*PDF ou image, très rarement un fichier tabulaire de données)*

200?-2020

Collecte de données
- Echantillons de terrain
- Données instrumentales brutes (in-situ)
- Données instrumentales brutes (ex-situ)

Processus Traitements
- Données traitée
- Données agrégées

Parfois en bases de données spécifiques

Article scientifique

* **Publication d'un article**

* **Dépôt des données partielles** de l'article (entrepôts internationaux)

Bases de données internationale
NOAA, PANGAEA, NEOTOMA, …

IS THEIR ANY OBSTACLES IN 2021 ?

- **Less incentive from local institutions ?**

  ➔ **No, country** build **nationals' roadmap**

- **Lack of appropriate IT infrastructure** ?

(where to put your data, safe place and easily accessible)

➔ **No !** but *Yes* we still **need easly & ergonomic IT & Curation procedures** to describe & deposit data

- **Lack of time ?** managing your data is a time-consuming activity ?

  ➔ **True !** especially if it taken into account too late

  ➔ Need to **well organize / to anticipate / to create or optimize procedures ?**

- **Lack of skills ?**

  ➔ **True !** data management and sharing require various skills in IT, data curation, legal knowledge, etc.)

https://data.gov.be/fr/info-faq

https://www.snf.ch/fr/FAiWVH4WvpKvohw9/dossier/points-de-vue-politique-de-recherche

YES

YES

➔ 1- Increase skills and use easly tools to save time

➔ 2- Create teams : Researcher - Data-librarian - IT Profils

Self-training courses

➔ <u>Parcours interactif</u> adapté aux domaines de l'environnement

In french : Big effort for training

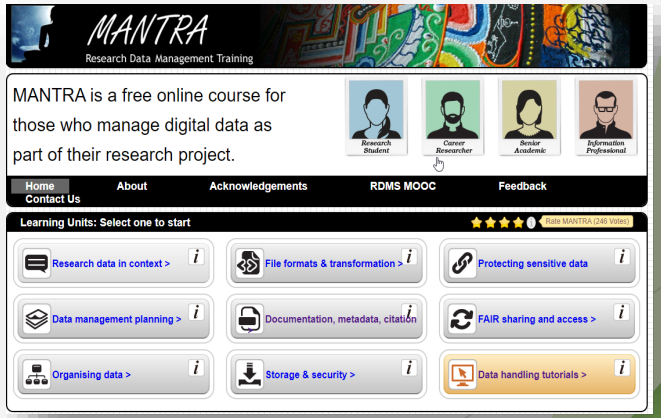
DoraNUM (INST CNRS), MITI (CNRS), CoopIST (CIRAD),

➔ Guide, Fiches, Vidéos…

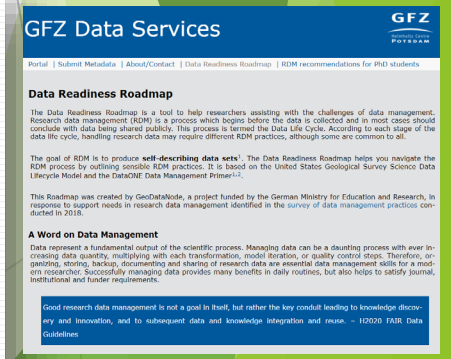
https://doranum.fr/enjeux-benefices/parcours-interactif-sur-la-gestion-des-donnees-de-la-recherche/

https://coop-ist.cirad.fr/gerer-ses-donnees

In english :



https://mantra.ed.ac.uk/#profiles



Before continuing the rest of the presentation

Opendata concept (*1 janv. 2020)*

[… ***data should be "as open as possible and as closed as necessary"***
"open" in order to foster the reusability and to accelerate research, but at the same time they should be closed" to safeguard the privacy of the subjects…]"* according to the H2020 Program Guidelines on FAIR Data

https://direct.mit.edu/dint/article/2/1-2/47/9998/The-A-of-FAIR-As-Open-as-Possible-as-Closed-as

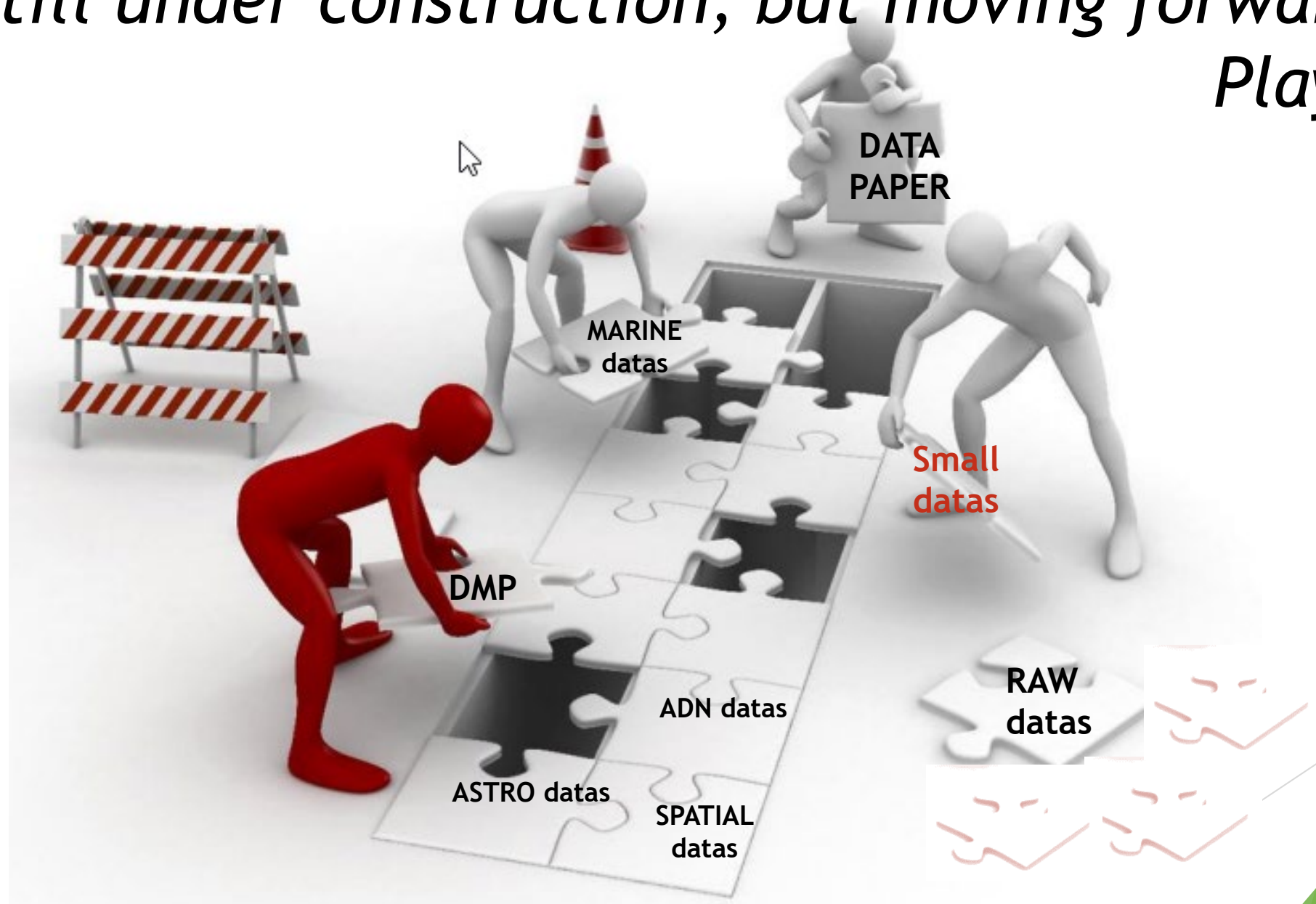
** persons, industriel contact, military information, …*

As Opendata is a way of work for the future decades , think that you do what you want, within the limit of :

- your **funder injunction & laboratory or doctoral institution**
- (your **data manager need's** because the institution tell him to do)
- the **time you dedicated to this job :** this (new) job isn't even an option now !

Still under construction, but moving forward !

Play together !



-PART II-

Fondamentals in Data management

***

1-Data Life cycle

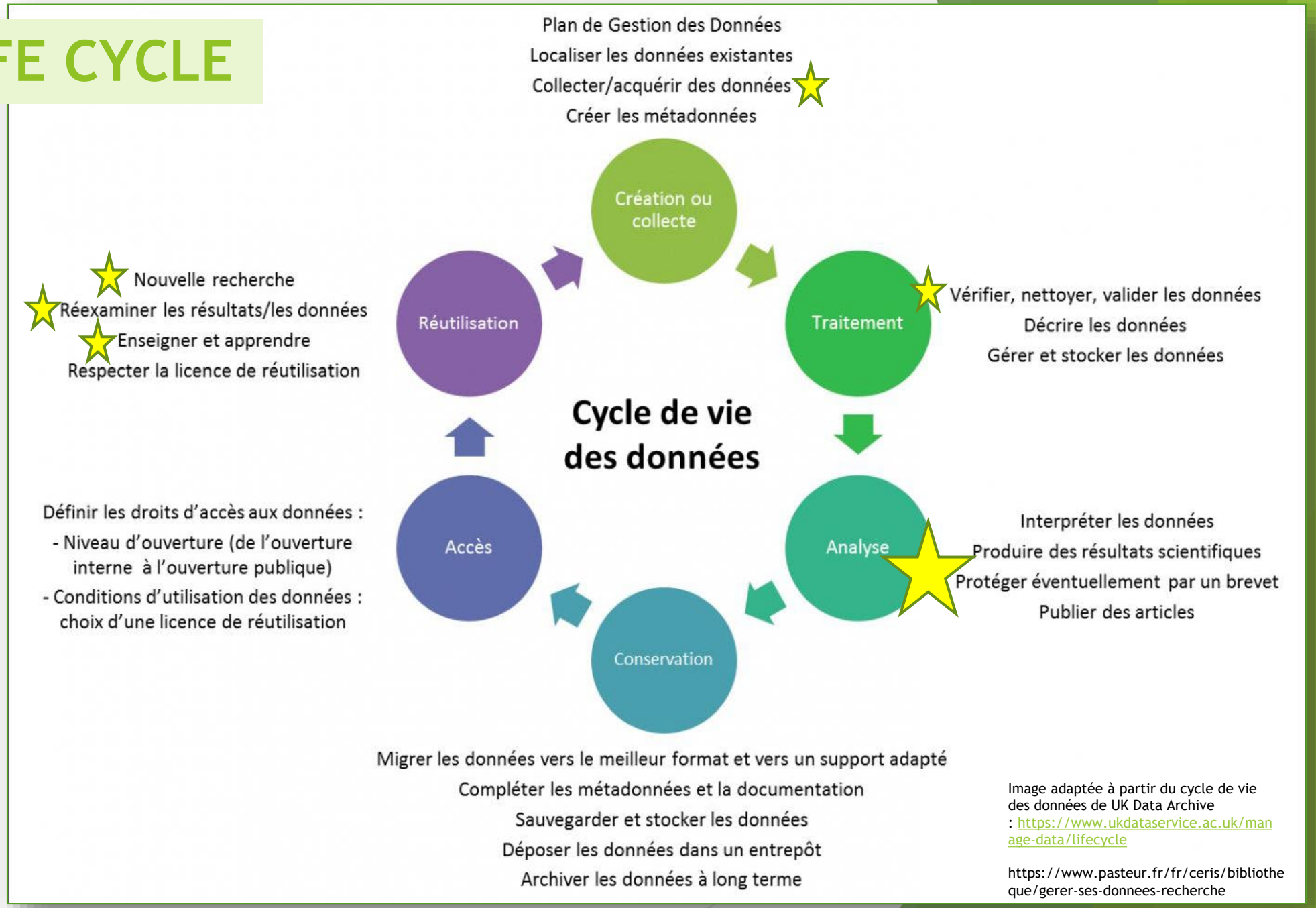
2-What kind of data ?

3-F.A.I.R. Principles

THE DATA LIFE CYCLE

6 STEPS

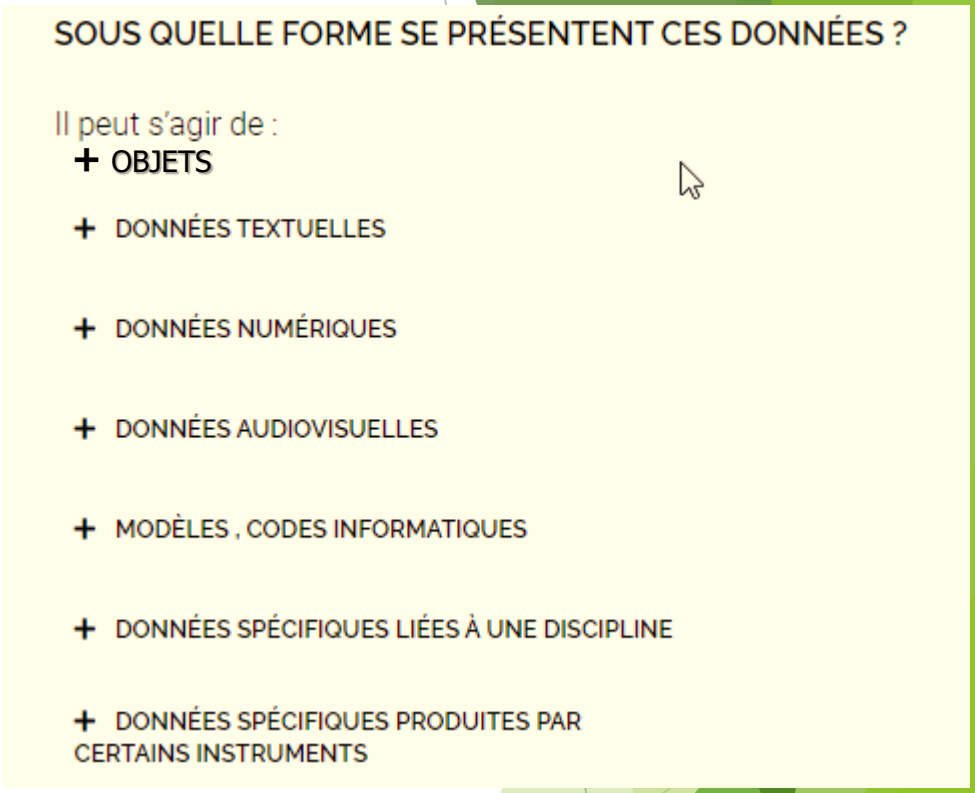
But **not necessarily** in the research timeline



What kind of data is involved ? (in the environment)

- identify and list the **types of data sets** that will be collected or generated (in the project/thesis, internship) :

Ex : Samples, experimental data, observation data, data from social surveys, simulation, software, images, space data *etc.)*



What Kind of informations do we need ?

- specify the **study object (feature of interest)**, **geographical origin**, **time period**
- identify **the purpose for collecting**
- specify the **generation** of these different types of data (method, protocols, instruments, ...)

https://doranum.fr/plan-gestion-donnees-dmp/origine-description-donnees-recherche/
https://coop-ist.cirad.fr/gerer-des-donnees/rediger-un-pgd/6-garantir-la-comprehension-et-l-accessibilite-des-donnees

F.A.I.R. Principles

"it evokes a proactive and altruistic behaviour of the data producer, which seeks to make them easier to find and use by all, while facilitating downstream sourcing (possibly automatic) by the data user"

comportement proactif et altruiste du producteur de données, qui cherche à les rendre plus facilement trouvables et utilisables par tous, tout en facilitant en aval le sourçage (éventuellement automatique) par l'utilisateur des données

➔Technics & tools enabling FAIRisation of data

https://www.go-fair.org/fair-principles/

FAIR Aware

https://fairaware.dans.knaw.nl/

Australian FAIR Tools diagnostic

https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/

Européen FAIR Tools diagnostic

https://www.f-uji.net/index.php

F.A.I.R. Principles : Who can applic & help for this ?

IT skills are everywhere in FAIR !

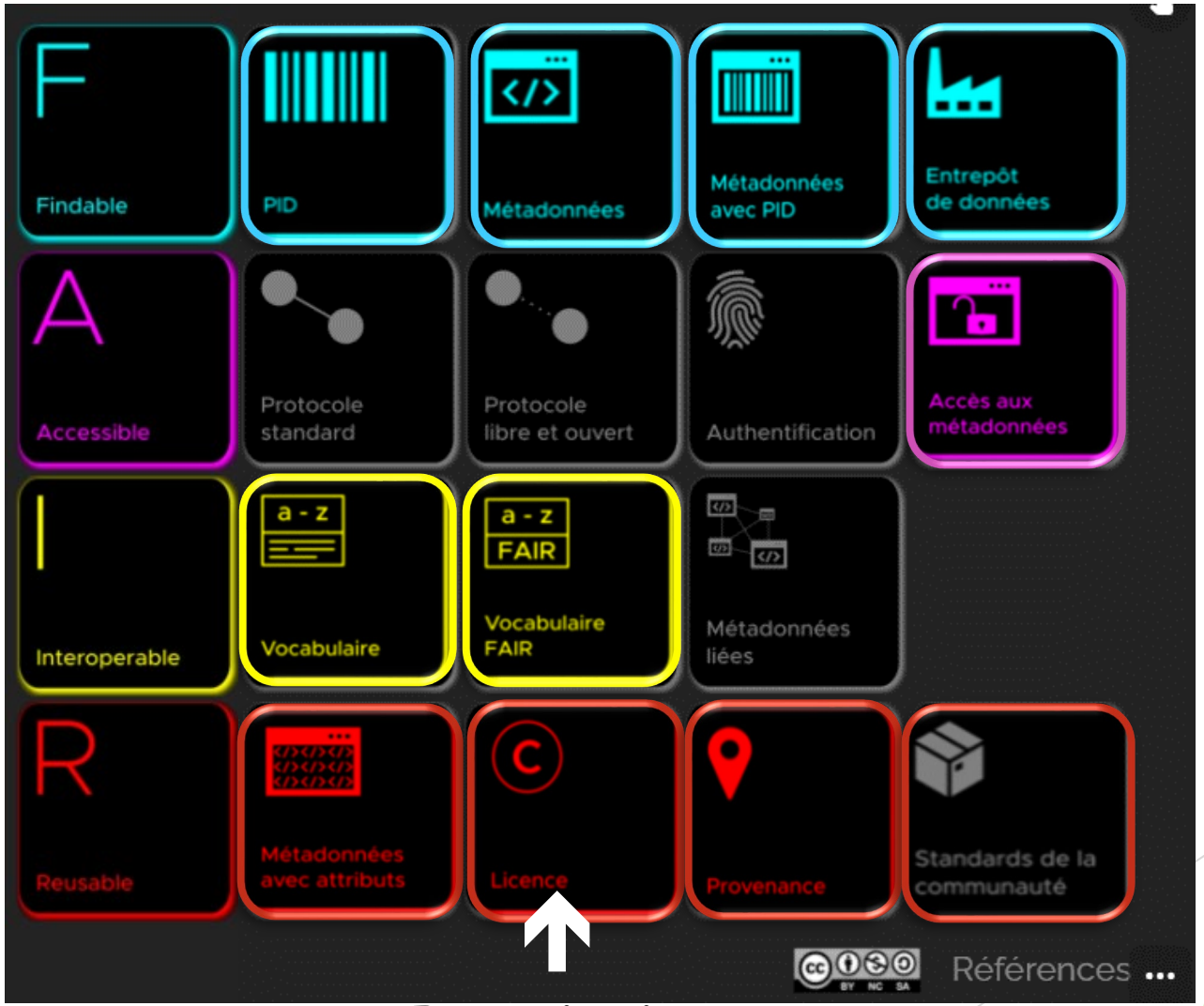
But one IT guy don't have all the time and skills

➔ Make a team !

Data librarian ➔

R=Researcher ? ➔

From you, for you !



➔ Data manager

➔ Data librarian

Data jurist

https://doranum.fr/depot-entrepots/criteres-choix-entrepot/

F.A.I.R Principles : 2 levels of management

level « Discovery »

hight level of quality

Nb : you can't do a Data-paper only with this level !

- **Level 1 = Findable + Accessible**

➔ Don't except a hight quality of description

LEVEL 1

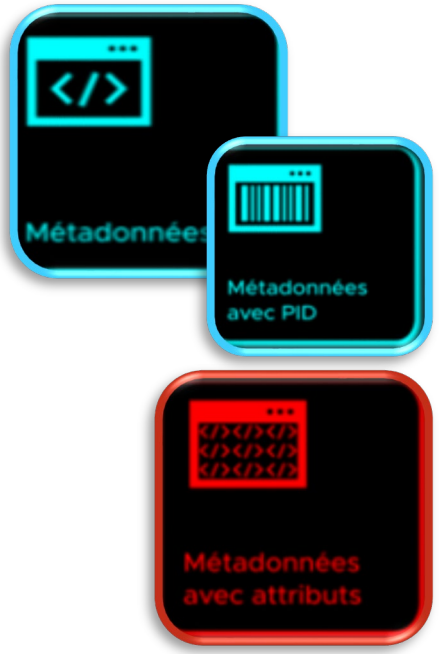
- **Level 2 = Level 1 + Interoperable & Reusable**

beyond « Discovery »

➔ Except a hight quality of description, precision for protocols, sensor, mesurements, …

LEVEL 2

Somes basic F.A.I.R. principles on practices



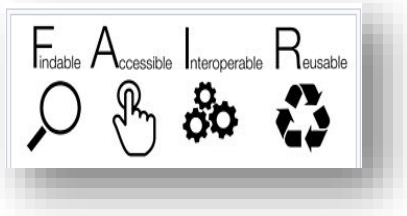
To be FAIR, the data must be finely described using metadata.

- **Metadata automates** sorting and prioritization tasks when **searching** for data.
- They also allow the **reuser to better understand** the context of the data, the conditions under which they were created or collected, their characteristics, etc.

The more information you give about your data, the easier they will be to find and understand

➔ ***Discovery Metadata (generalist or interdisciplinary) (for level 1)***

➔ ***Thematic Metadata (for level 2)***



The basic principles on practices

To be FAIR, the data must be described using a controlled vocabulary allowing interoperability.

By describing a data set according to a model of representation and a controlled vocabulary, you will provide resources understandable by both humans and machines

To be FAIR, metadata must remain accessible in catalogues <u>even if the data is no longer accessible</u>.

- Acces by (harvested) **catalogues**
- Over time, data may disappear. **Metadata** can be very useful in this case, as it will provide **valuable information about the missing dataset** and allow other researchers to take over and continue the associated research.

Focus on METADATA STANDARDS : WHY AND WHICH ?

Within a scientific framework, metadata helps to describe the resources, research data and productions made (article, repository, photo, measurement, software, web page, etc.)

➔ Facilitating the **search** for scientific research data and publications in catalogues (Level 1).

➔ Help to **understand the datasets** (collection, structure and context), for their validation (peers) and/or reuse (***Level 2***)

https://doranum.fr/metadonnees-standards-formats/standard-metadonnees/

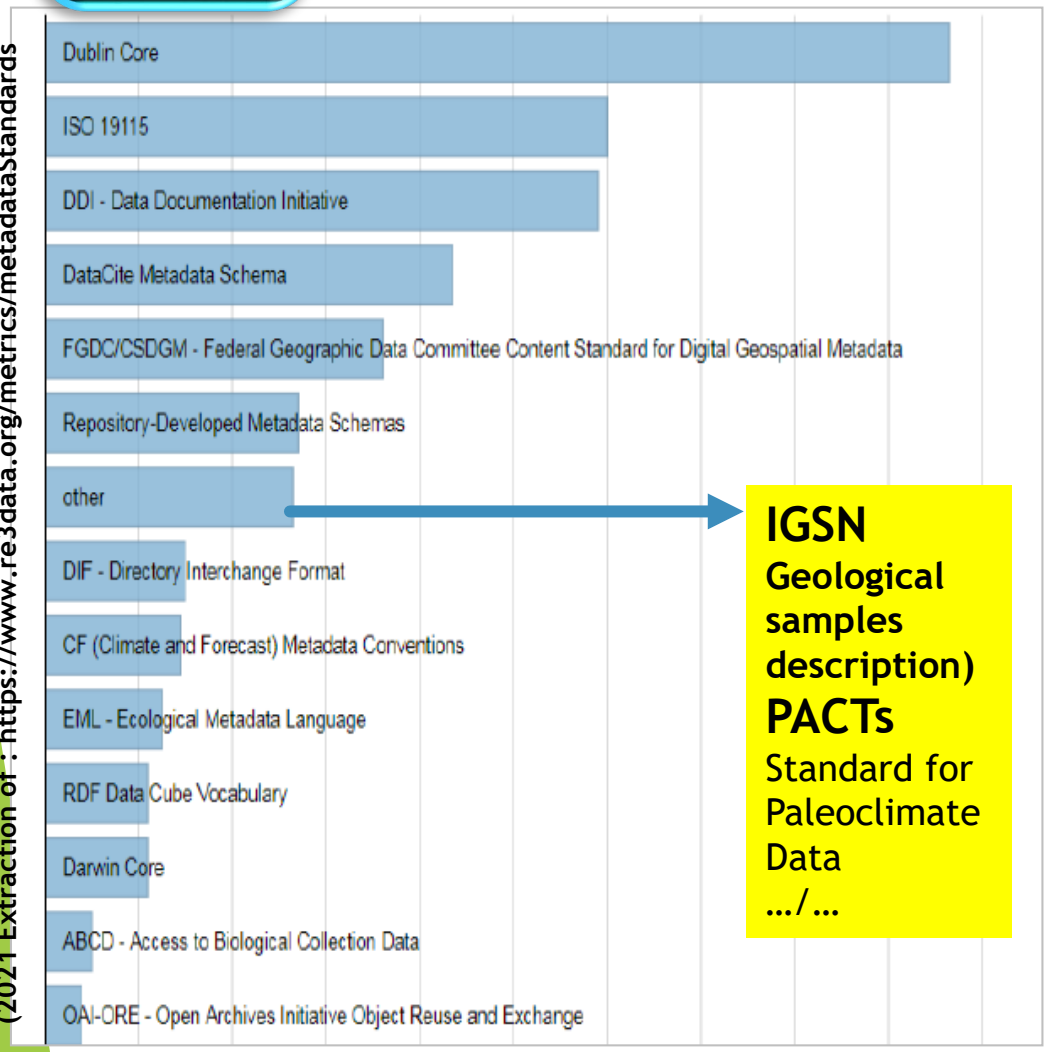
METADATA STANDARDS : WHICH ?

FAIRsharing.org standards, databases, policies

DCC Because good research needs good data

RDA Metadata Standards

➔ Catalogue of standards metadata & vocabulairy



(2021 Extraction of : https://www.re3data.org/metrics/metadataStandards

IGSN
Geological samples description)
PACTs
Standard for Paleoclimate Data
…/…

Level 1 : exemples interdisciplinary or generalist

- Dublin Core : *interdisciplinary, description* of digital resources.
- DataCite Metadata Schema : fairly generalist and can be useful to you regardless of your context *(Use for DOI PID)*

Level 2 : exemples thematics

- Iso19.115/139 : Used for geo-localized environmental data framework or geographic data
- EML (Ecological Metadata Language) : Used in Ecology
- Data Documentation Initiative (site officiel) : Used in Social science for social surveys

➔ **Some of them are « cross mapp » : ex Author = Creator**

Controlled vocabularies, thesauri, ontologies = interoperability & qualitity increase

- **Vocabularies (for « Human reading ») :** controlled vocabularies hierachic therms + defintitions

Level 1 : generalist // Level 2 : Specific & Thematic

- **Vocabularies *F.A.I.R* (for « marchine reading ») :** vocabularies hierachic therms with defintitions + having resolvable globally unique and persistent identifiers (URI)

Géographie thématique: ***Pollution***
Pollution (en), Contaminación (es)
http://data.loterre.fr/vocabs/BGI/GT/2440

Géographie thématique: ***Pollution chimique***
Chemical pollution (en), Contaminación química (es)
http://data.loterre.fr/vocabs/BGI/GT/2442

Géographie thématique: ***Pollution radioactive***
Radioactive pollution *(en),* ***Contaminación radioactiva*** *(es)*
http://data.loterre.fr/vocabs/BGI/GT/2447

Term's catalogues (multi-thesaurii)

1-Catalog's vocabularies :

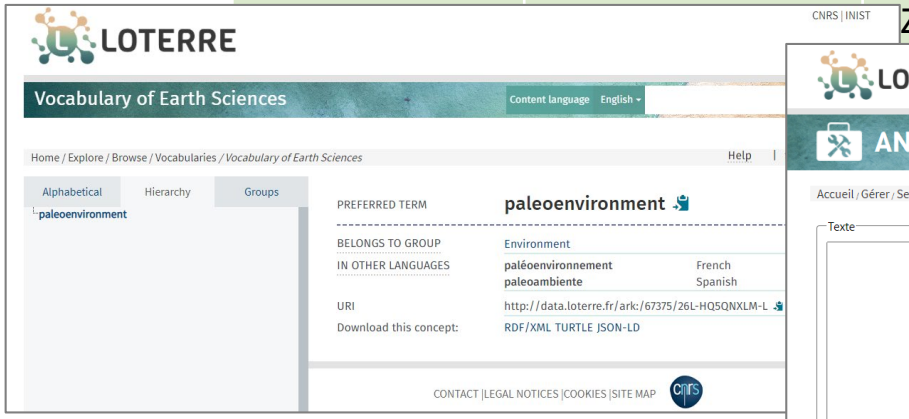
FAIRsharing.org
standards, databases, policies

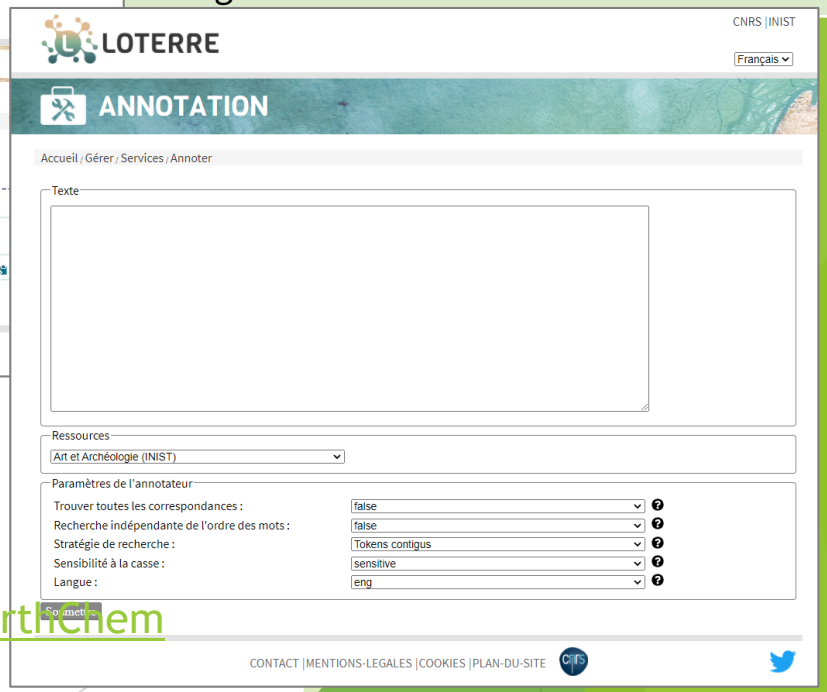
LOTERRE

Sciences de la terre et de l'univers	*Chimie*	*Sciences du vivant et environnement* *(non-exhaustif)*
Aéronomie, Météorologie Climatologie Astronomie Géologie Géophysique Glaciologie Océanographie	Chimie analytique Chimie générale Chimie théorique Chimie physique Chimie minérale (inorganique) Chimie organique	***Agronomie, écologie, environnement*** Biologie du développement végétal, botanique Évolution, écologie, biologie des populations Biotechnologies, sciences environnementales, agronomie, foresterie ***Biologie, santé*** Biologie moléculaire et structurale, biochimie Génétique, génomique, bio-informatique Zoologie

2- Search in multi-catalogs :

- **LOTERRE : https://www.loterre.fr**
  FR *(Fr/Eng/Sp/It, Ger)* F.A.I.R.





- **Quantity & Units :** *Units of Measure, Quantity Kinds, Dimensions*

  (Eng) ➔ *the more complete*

- **NERC / NOC BODC Vocabularies :**

  GB *(Eng)* ➔ *Marine's Community (EU-SeaDataNet)*

- **CZA Critical Zone :**

  US (Eng) ➔ *Thematics* Observation Data Model v2 (ODM2), CUAHSI CVs, IEDA EarthChem

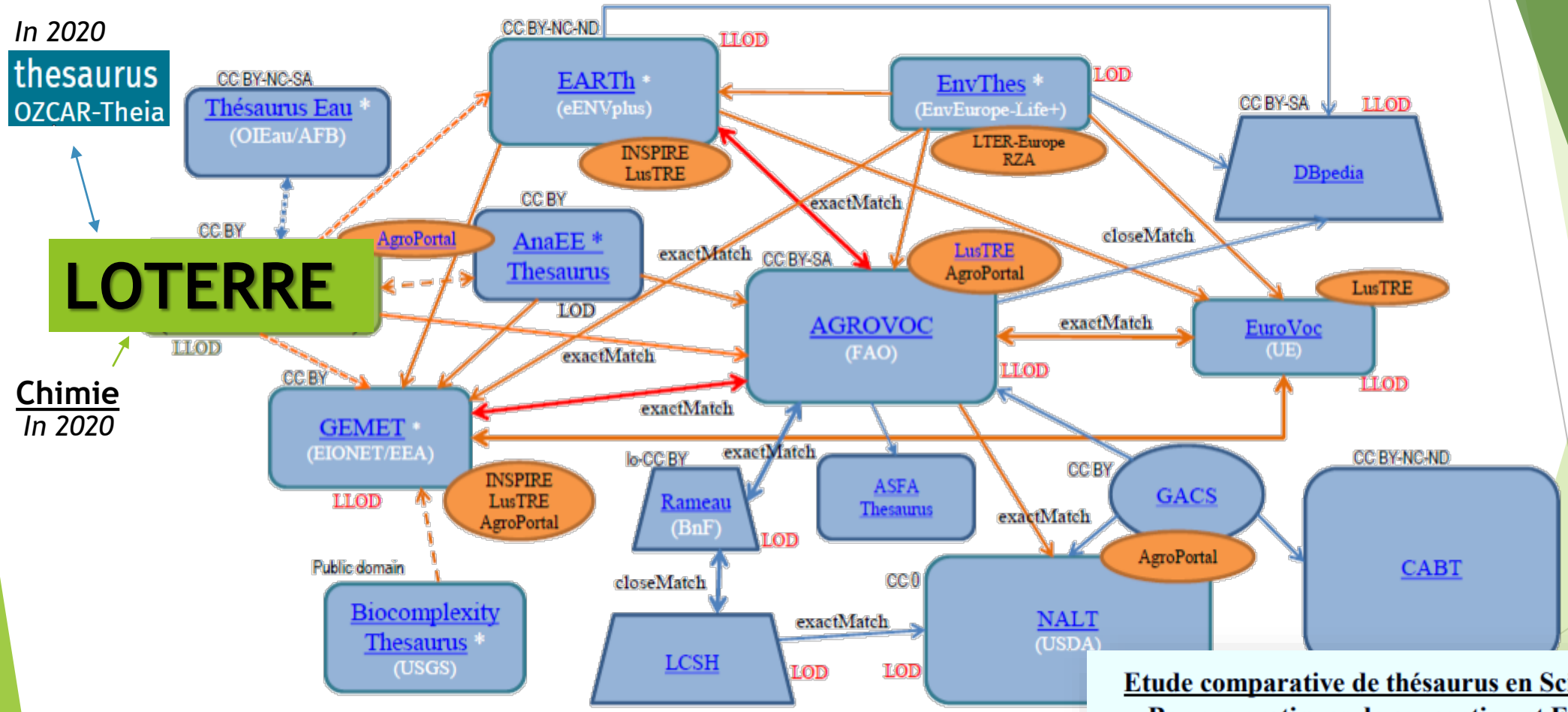
- **PANGAEA « List » Vocabularies :**

Parameter list

- Complete list of parameters used in PANGAEA

Term's catalogues (multi-thesaurii)

in 2018 exemple of thesaurii interopérability



Institut de l'information scientifique et technique (Inist-CNRS)

2018- Dominique Vachez

Etude comparative de thésaurus en Sciences de l'Environnement
Bonnes pratiques de conception et FAIRisation de thésaurus

Dominique Vachez
dominique.vachez@inist.fr
INIST-CNRS, 2 rue Jean Zay CS 10310 F-54519 Vandoeuvre-lès-Nancy, France

HAL - Inist
Publications scientifiques de l'Inist

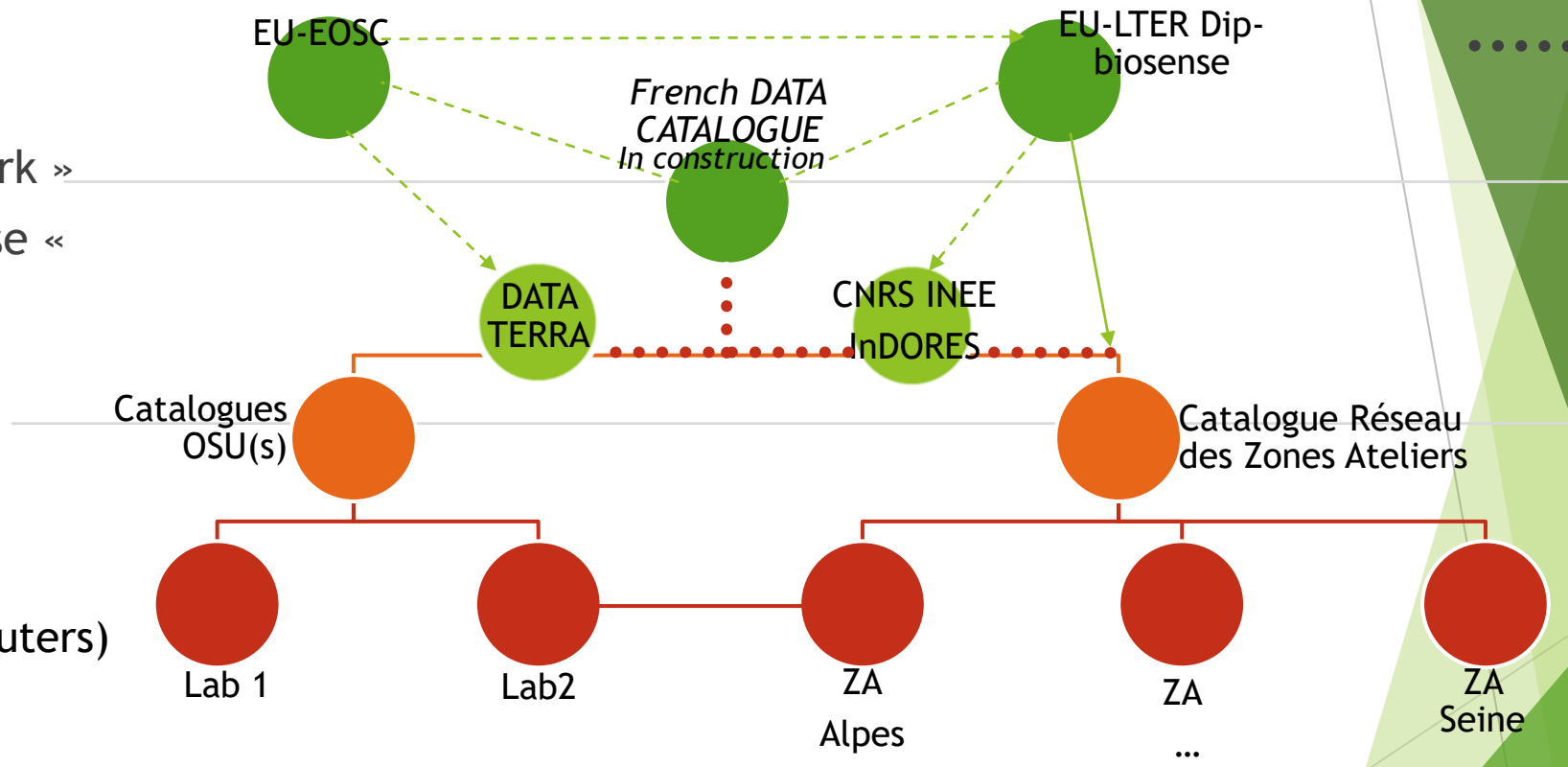
Exemple of interest of Metadata for catalogs Harvesting : *Interest of catalogues in data visibility* *(oct 2021)*

Catalogs & Repository

- Metadata « Geonetwork »
- Repository « Dataverse «
- Others …

Meta-search tools

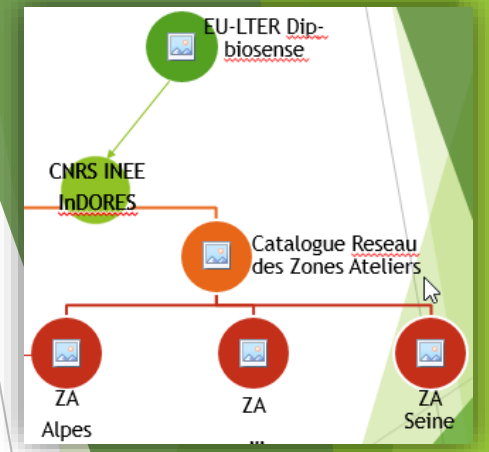
- Data Cite search
- Data Citation Index (Reuters)
- Google Dataset Search
- Data Search (Elsevier)



harvested by catalogues, integrators, European data infrastructures… more and more numerous (EOSC, OpenAIRE, DataONE, etc.)

Exemple of « metadata sheets trip »

(metadata standard - Iso19115 & Geonetwork catalogue)



1-metadata sheets with low metadata

LEVEL 1

Pollen analysis - Grozon (F-39), L'étang (Marais)

Téléchargements et liens

041165667
http://www.sudoc.fr/041165667

(suspicion of IT harvesting problems …)

2-metadata sheets with more metadata

LEVEL 1

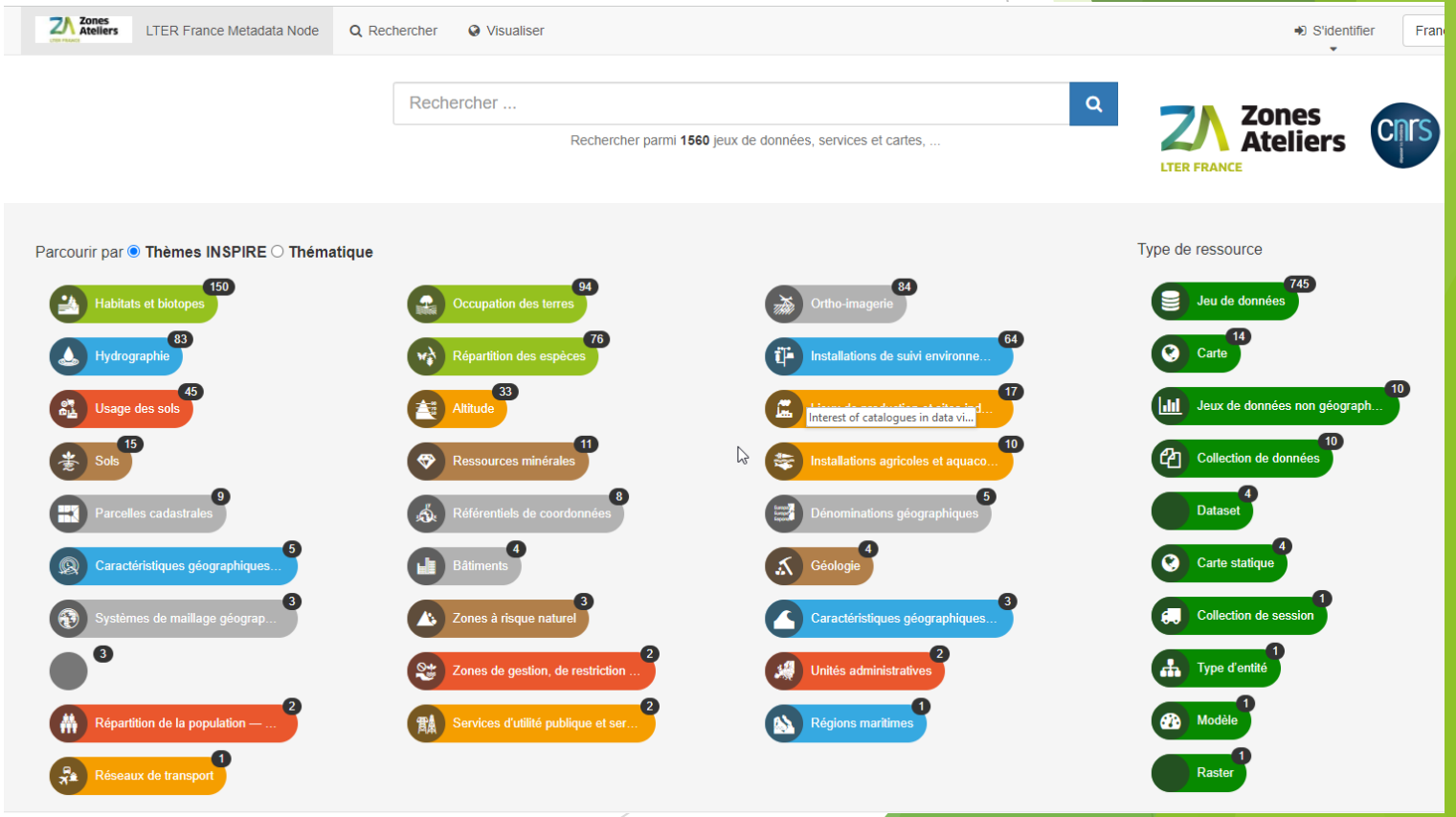
Section de carottage MUZ12-01A (IEFRA094Y) - Lac de la MUZELLE, Carottage (04/2012) Zone Atelier Alpes ZAA (LTER-FRANCE ROZA)

(do not have the link to the data)

3-metadata sheets with complete hight metadata quality

LEVEL 2

Lac de la Muzelle - Carottage Muz12-I – Plant cover and erosion dynamics : impacts of pastoral activities and climatic changes over the last 1600 years - 2017 – (IGSN:IEFRA00A4)



Exemple of « metadata sheets trip »

thanks to metadata standard (Iso19115 - Geonetwork catalogue)

ex of keyword : study area

➔ ***Interest of indexation catalogues in data visibility***

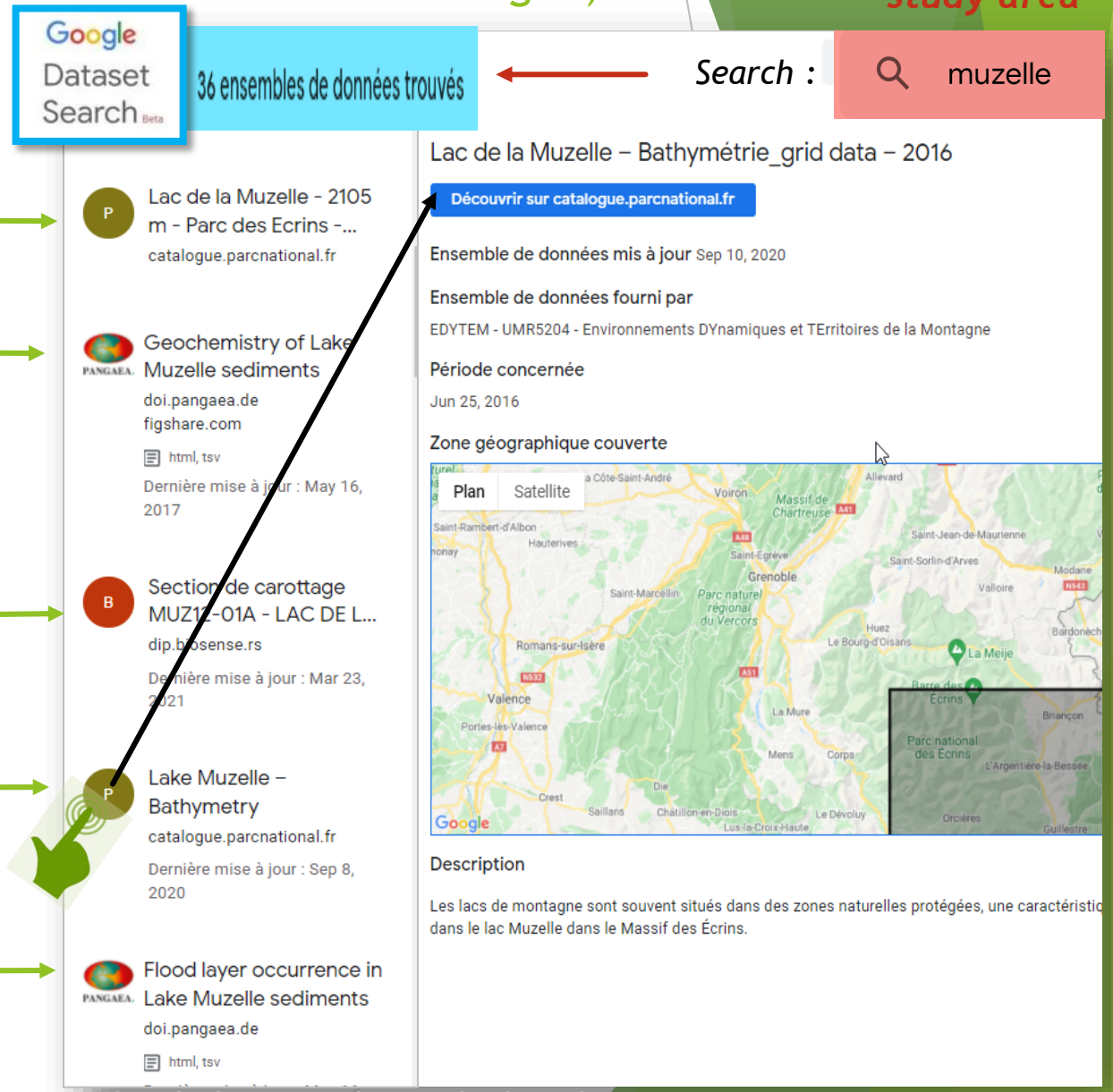
Local Catalogue

International thematic Repository

International Catalog (Sheet harvesting from à Local Catalog)

Local Cataloge

International thematic Repository



TOOLS FOR METADATA CREATING

Many tools / Many ways to do

(template, formulars, package R, database,…)

➔ Find the one(s) that best suits you,
according to your profile, your(s) standard(s) needs, …

Exemples for our « Paleo-community » :

1) (eng) International metadata Geosample : **IGSN**
2) (eng) **interdisciplinary community Tool,** base on standards DublinCore, DataCite, Iso19.115 (et EML) : **Geoflow**
3) (fr) OSU (OTELo) recommended **template** for fields to be filled Observ. Terre Environnement Lorraine
4) (fr) **a webtool provided by** DataCite
5) etc …

Initiate by world Marine consortium

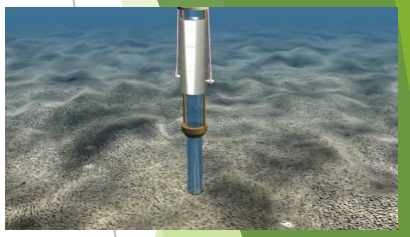
See Slide N°55

Funder FAO, IRD, INARe, CNRS LTER-fr …

https://doranum.fr/metadonnees-standards-formats/outils-creation-metadonnees/

-PART III- Practical from case studies & Basics FAIR Princips

- Usecase 1 for Sample collection : ex of cores
- Usecase 2 about Analytical data
- Which « Licences » ?
- Why How deposit your research data in a data-repository ?
- Use case of Workflow deposit data



	Total defects	A	B	C	D	E
A4636	131	37	21	28		45
A2524	86	20	24	21	1	20
A3713	75	17	13	18		27
A4452	73	5	33	17		18
A4088	72	14	16	12	2	28
A2103	68	14	13	14	1	26
A2156	68	16	13	19	2	18
A3681	66	12	16	9	1	28
A1366	50	11	15	12		12
A2610	39	5	7	12		15
Total	728	151	171	162	7	237

Remember F.A.I.R Principles : 2 levels of management

level « Discovery »

Nb : you can't do a Data-paper only with this level !

- **Level 1 = Findable + Accessible**

➔ Don't except a hight quality of description

hight level of quality

- **Level 2 = Level 1 + Interoperable & Reusable**

beyond « Discovery »

➔ Except a hight quality of description, precision for protocols, sensor, mesurements, …

Before usescases, let's continue to see some basics F.A.I.R. principles

LEVEL 2

To be FAIR, the data must indicate their origin (Provenance)

data are reusable = information to contextualize them

as the authors, their institution, the date of creation of the data, sources if aggregated data,

Link the relationship with other data already published, etc... (add "Read me" for additional information if neccessairy)

To be FAIR, the data are richly described with a plurality of specific and relevant attributes

*The more you know the **context in which data was created** (technical, setting, limit of detection,...), the more you can benefit from it. It can be technical metadata (contextual information on the data, even those which may seem useless to you)*

addition of a text file of the type "DataDictionary"

To be FAIR, the data must be uniquely identifiable and sustainable using a PID.

DOI - IGSN - ORCID ...

To be FAIR, the metadata must contain the PID of the dataset described

Case of SAMPLES : *in Environnemental Sciences*

UseCase 1

Most of the time, no science without samples !

Samples are DATA

and all « sample's data » become metadata of your future Analytical mesurement

➔ Important to well describe them a the begining of the project (+ fielwork report)

➔ They are part of your « **FAIR *Provenance*** » for your futur analytical datas

Only one standard for (geological) samples = IGSN

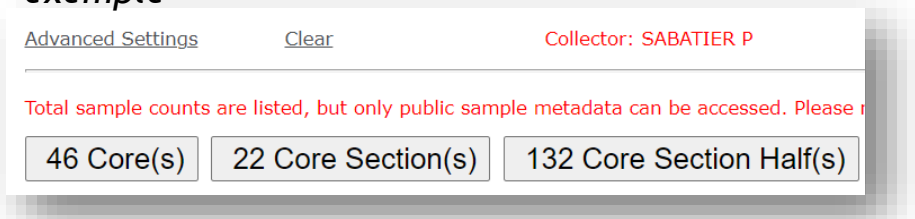
Nb : Find your own Allocation Agent in your country to declare your samples

https://www.igsn.org

SAMPLES : Sample collection *in Environnemental Sciences, Data become Metadata !*

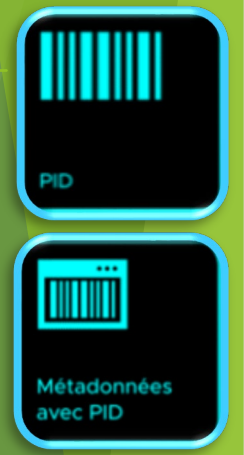
- **Common Metadata standart** to describe a general geological sample
- **A unique Identifier** (for each sample and children) link in your articles, datapaper or IT Data infrastructure

exemple



Editors (Recommends IGSN)
- Copernicus Publications
- AGU Publications, …

➔ **Assigne PID**
➔ **Good common metadata** but it is« Poor » technical metadata
➔ **Sufficient for discovery level**



Pit the PID in your général MD standard

UseCase 1

THU10-I

PANGAEA.
Data Publisher for Earth & Environmental Science

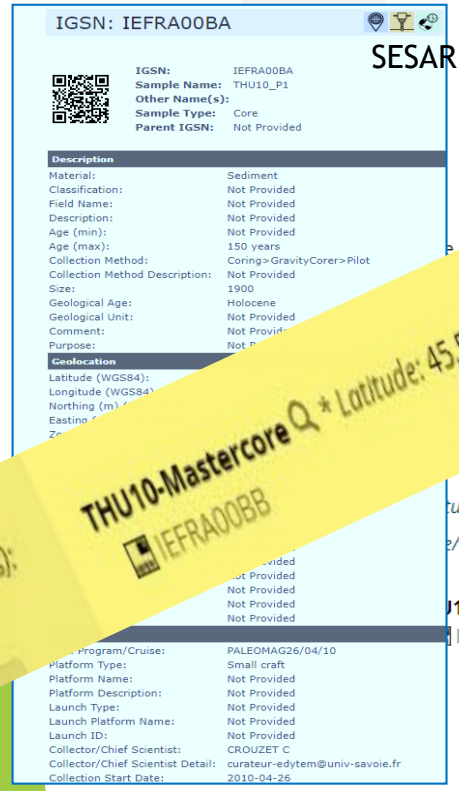
UseCase 1

Citation: **Bajard, Manon; Sabatier, Pierre; David, Fernand; Develle, Anne-Lise; Reyss, Jean-Louis; Fanget, Berna… Malet, Emmanuel; Arnaud, Daniel; Augustin, Laurent; Crouzet, Christian; Poulenard, Jérôme; … Fabien (2015):** Analyses of lake La Thuile sediment core. *PANGAEA*, https://doi.org/10.1594/…

Supplement to: Bajard, M et al. (2015): Erosion record in Lake La Thuile sediments (Pre… montane landscape dynamics throughout the Holocene. *The Holocene*, **26(3)**, 35… https://doi.org/10.1177/0959683615609750

Always quote citation above when using data! You can download the citation in several form…

RIS Citation | BibTeX Citation | Copy Citation | Facebook | Twitter | Show M…


IGSN: IEFRA00BA

SESAR

IGSN:	IEFRA00BA
Sample Name:	THU10_P1
Other Name(s):	
Sample Type:	Core
Parent IGSN:	Not Provided

Description

Material:	Sediment
Classification:	Not Provided
Field Name:	Not Provided
Description:	Not Provided
Age (min):	Not Provided
Age (max):	150 years
Collection Method:	Coring>GravityCorer>Pilot
Collection Method Description:	Not Provided
Size:	1900
Geological Age:	Holocene
Geological Unit:	Not Provided
Comment:	Not Provid…
Purpose:	Not P…

Geolocation

Latitude (WGS84):	
Longitude (WGS84):	
Northing (m):	
Easting:	

Program/Cruise:	PALEOMAG26/04/10
Platform Type:	Small craft
Platform Name:	Not Provided
Platform Description:	Not Provided
Launch Type:	Not Provided
Launch Platform Name:	Not Provided
Launch ID:	Not Provided
Collector/Chief Scientist:	CROUZET C
Collector/Chief Scientist Detail:	curateur-edytem@univ-savoie.fr
Collection Start Date:	2010-04-26

https://app.geosamples.org/sample/igsn/IEFRA00BA

…, provides an 18 m long sedimentary sequence spanning the entire Lateglacial/Holocene period. The high resolution multi-proxy (sedimentological, palynological, …eveals the Holocene dynamics of erosion in the catchment in response to landscape modifications. The mountain belt is at relevant altitude to study past human activities and the …rge valleys to capture a local sedimentary signal. From 12,000 to 10,000 cal. BP (10 to 8 ka cal. BC), the onset of hardwood species triggered a drop in erosion following the …0,000 to 4500 cal. BP (8 to 2.5 ka cal. BC), the forest became denser and favored slope stabilization while erosion processes were very weak. A first erosive phase was initiated at ca . 4500 cal. BP …sence in the catchment. Then, the forest declined at approximately 3000 cal. BP, suggesting the first human influence on the landscape. Two other erosive phases are related to anthropic activities: …P (550 cal. BC) during the Roman period and after 1600 cal. BP (350 cal. AD) with a substantial accentuation in the Middle Ages. In contrast, the lower erosion produced during the Little Ice Age, when climate …generally considered to result in an increased erosion signal in this region, suggests that anthropic activities dominated the erosive processes and completely masked the natural effects of climate on erosion in the …e.

…tude: 45.530000 * *Longitude:* 6.056700

…e/Time Start: 2010-04-25T00:00:00 * *Date/Time End:* 2013-05-22T12:25:30

…10-Mastercore * *Latitude:* 45.530000 * *Longitude:* 6.056700 * *Date/Time:* 2010-04-25T00:00:00 * *Elevation:* 874.0 m * *Method/Device:* Piston corer (PC) * *Comment:* IGSN of cores: THU10-P1: IEFRA00BA; THU10-I: IEFRA00BB

Chemical composition of sediment core THU10-Mastercore. https://doi.org/10.1594/PANG…
Declination of sediment core THU10-Mastercore. https://doi.org/10.1594/PANGAEA.855423
Grain size composition of sediment core THU10-Mastercore. https://doi.org/10.1594/PANG…
Loss on ignition in sediment core THU10-Mastercore. https://doi.org/10.1594/PANGAEA.855…
Radiocarbon age determination of sediment core THU10-Mastercore. https://doi.org/10.…
Short-lived radionuclides in sediment core THU10-Mastercore. https://doi.org/10.1594/PA…

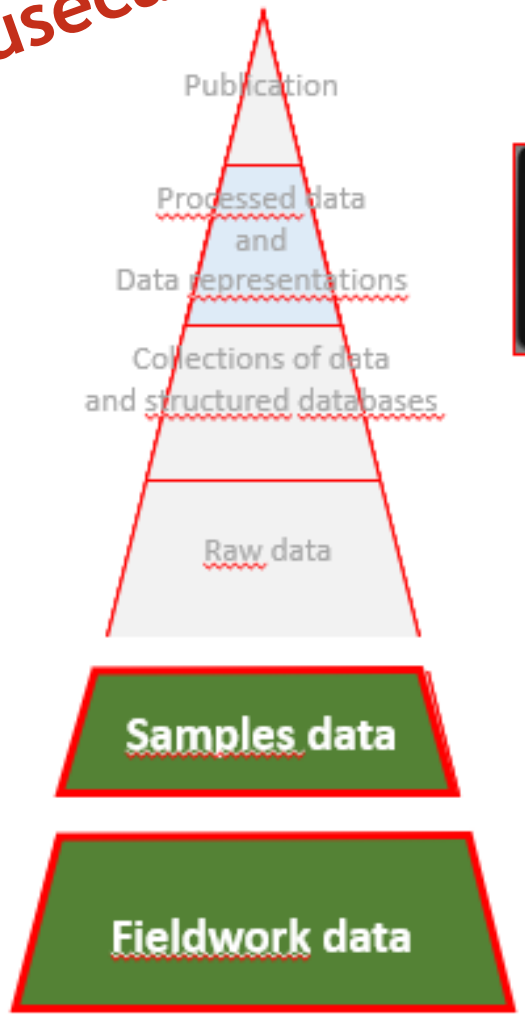
https://app.geosamples.org/sample/igsn/IEFRA00BB

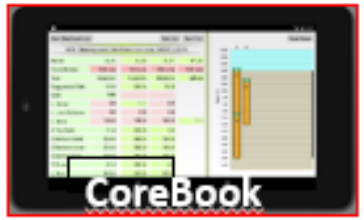
How to turn kilos of mud into megabytes of data?
10 years of efforts in curating lake sediment cores and their associated results

Exemple of usecase 1

➔ **Banking and displaying fieldwork data: the National Cyber-core repository (Cyber-carothèque nationale)**

Findable Accessible Interoperable Reusable





Flow WFS

OGC

Export geographical point & fielddata for GIS SYSTEM

Arnaud F., Pignol C. et al., OZCAR-TERENO-2021

Provenance : Collect your sample collection

Exemple of usecase 1

In France :

French Cyber-Core repository

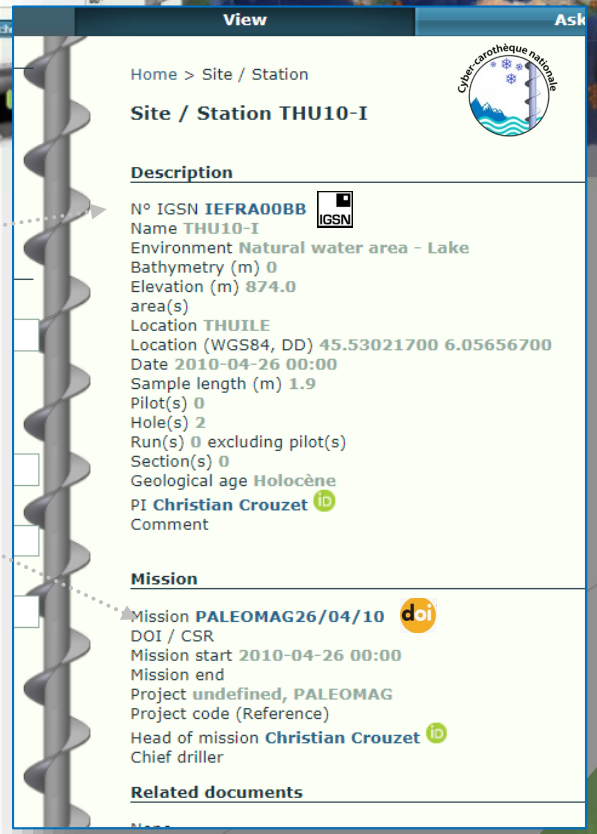
(marine, lacustrine, river, peatland, *ice cores)*

- *Directy on the field*, **collect data** by APK Android

OR

- after the field work, **Import CCN template**
  (or for your core legacy)

Download file here



In the end of 2021 = version v2 ➔ CYBER-Carothèque Core repository NATIONALE

+ IGSN automatic assignation

+ Campaign PDF report (DOI)

+ Pivot format for link to Core Storage Db (« Collec-Science » ?)

Accès aux métadonnées

IGSN
DOI
ORCID ...
PID

Métadonnées avec attributs

Vocabulaire

Other Databases for samples :

- Other IGSN AAgent IGSN like IFREMER, *SESAR*
- NOAA Index Marine Lacustrine IMLGS
- CSDCO Continental Scientific Drilling Holes / LacCore
- IODP Core Repository

Exemple of usecase 1

How to turn kilos of mud into megabytes of data?
10 years of efforts in curating lake sediment cores and their associated results

➔ **Banking and displaying fieldwork data: the National Cyber-core repository (Cyber-carothèque nationale)**

Mission information

Mission PALAS19-GUYNEMER

Description

Fichier source Corebook

Code
Début de mission 27-11-2019 06:00
Fin de mission 27-11-2019 18:00

Description

Projet PALAS
Contexte scientifique

Connaissances préliminaires

Commentaire

Confidentialité Public

Remerciements

DOI
URL
CSR

XML source file

Link to project: permit to display all surveys/cores within the same project

Automatic survey report with DOI (under development)

Participants

Chef foreur **Arnaud Fabien**
Chef de mission **Arnaud Fabien**

PI(s) des prélèvements (liste non exhaustive)
Arnaud Fabien

Financeur

Aucun(e)

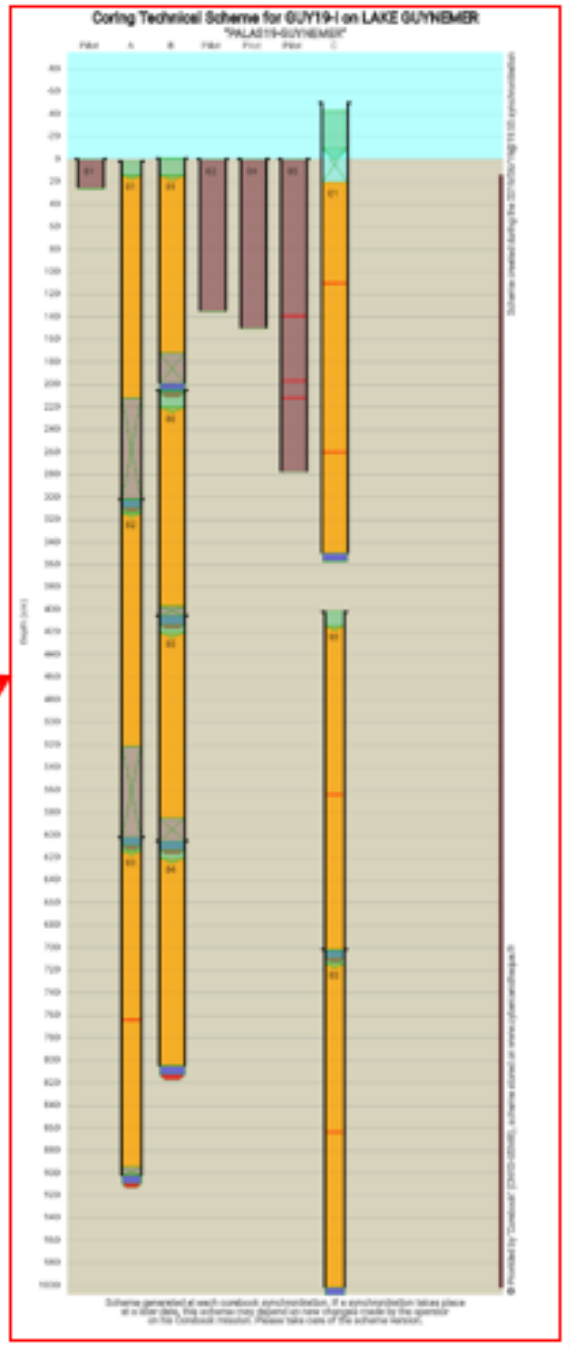
Documents liés

GUY19-I-Pil-05-20191223-141634.jpg
GUY19-I-Pil-05-20191223-141716.jpg
GUY19-I-Pil-05-20191223-141736.jpg
GUY19-I.20190619-195534.Coring.Schema.png
GUY19-I-B-01-20191201-064036.mp4
EPGUY19-01-20191204-113555.jpg
EPGUY19-01-20191204-113630.jpg

Acknowledgement of crew members with determined role and ORCID identification

Any additionnal document (pictures, sounds, videos) taken with CoreBook App.

When applicable: link to the sketches of parallel corings



Exemple of usecase 1

How to turn kilos of mud into megabytes of data?
10 years of efforts in curating lake sediment cores and their associated results

➔ **Banking and displaying fieldwork data: the National Cyber-core repository (Cyber-carothèque nationale)**

Sample (core) information

IGSN unique identifier

Codified usual name : Site-Year-Station-Hole-Section

Where is the core stored?

Link to coring tool information

Children subsamples

Description

N° IGSN **TOAE0000000078**
Site / Station **GUY19-I**
Nom GUY19-I-A-03
Matériel Sédiment
Type de prélèvement Carotte (recouvrement)
Environnement Étendue d'eau (douce) naturelle - Lac
Plateforme
Profondeur bathymétrique (m) 94
Altitude (m) 131.0
zone(s) French Southern and Antarctic Lands, France
Lieu LAKE GUYNEMER
Point GPS (WGS84, DD) -49.11344500 69.02049300
Date 28-11-2019 16:01
Nom du carottier **U-NIEDERREITER 63-3m (EDY)**
Nom de la configuration **2-NIED63-3m PO/TC (EDY)**
Diamètre intérieur (mm) 63
Longueur du prélèvement (m) 2.93
Longueur matière récupérée (m) 2.81
Époque géologique
PI **Fabien Arnaud**
Sondeur/Foreur en chef
Commentaire

Mission

Mission **PALAS19-GUYNEMER**
DOI / CSR
Début de mission 27-11-2019 06:00
Fin de mission 27-11-2019 18:00
Projet PALAS
Code projet (Reference)
Chef de mission **Fabien Arnaud**
Sondeur/Foreur en chef **Fabien Arnaud**

Repository

Repository Environnements, DYnamiques et TErritoires de la Montagne (EDYTEM)

Documents liés

Aucun(e)

Sous-prélèvements associés

Export csv

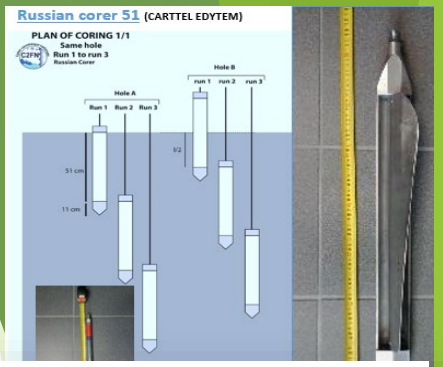
Nom	Parent	IGSN	Profondeur supé
GUY19-I-A-03-A Section	GUY19-I-A-03 Carotte (recouvrement)	TOAE0000000582	
GUY19-I-A-03-B Section	GUY19-I-A-03 Carotte (recouvrement)	TOAE0000000583	

Arnaud F., Pignol C. et al., OZCAR-TERENO-2021

Exemple of usecase 1

How to turn kilos of mud into megabytes of data?
10 years of efforts in curating lake sediment cores and their associated results

➔ **Banking and displaying fieldwork data: the National Cyber-core repository (Cyber-carothèque nationale)**



Sample (core) information

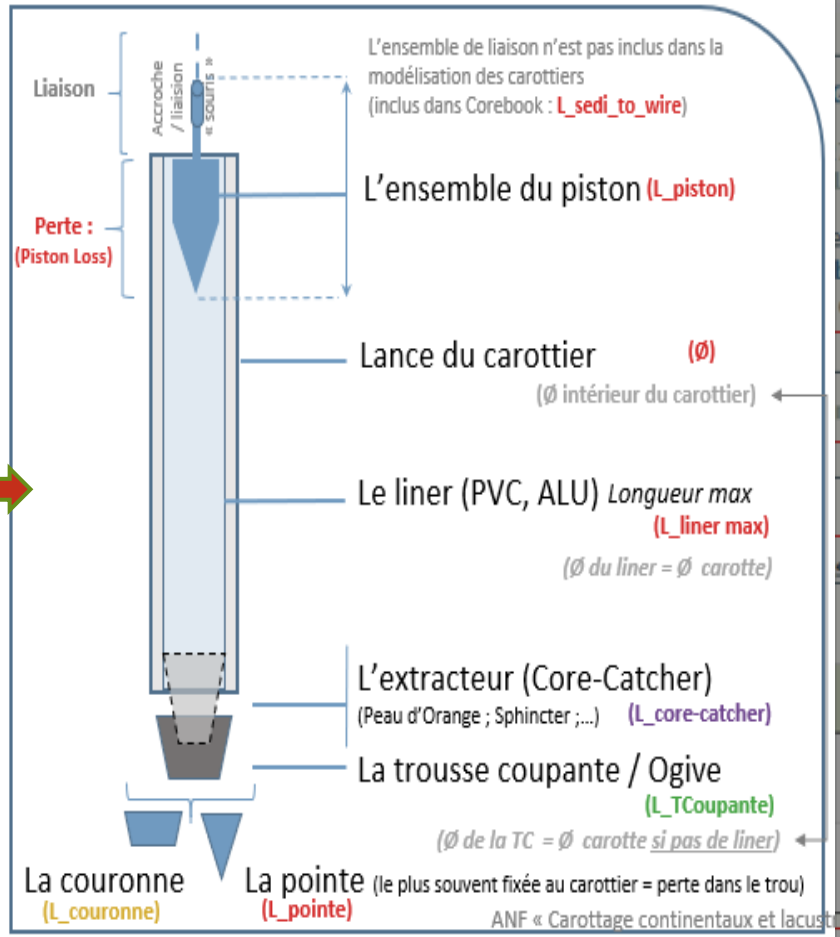
Description

N° IGSN **TOAE0000000078**
Site / Station **GUY19-I**
Nom GUY19-I-A-03
Matériel Sédiment
Type de prélèvement Carotte (recouvrement)
Environnement Étendue d'eau (douce) naturelle - Lac
Plateforme
Profondeur bathymétrique (m) 94
Altitude (m) 131.0
zone(s) French Southern and Antarctic Lands, France
Lieu LAKE GUYNEMER
Point GPS (WGS84, DD) -49.11344500 69.02049300
Date 28-11-2019 16:01
Nom du carottier **U-NIEDERREITER 63-3m (EDY)**
Nom de la configuration **2-NIED63-3m PO/TC (EDY)**
Diamètre intérieur (mm) 63
Longueur du prélèvement (m) 2.93
Longueur matière récupérée (m) 2.81
Époque géologique
PI **Fabien Arnaud**
Sondeur/Foreur en chef
Commentaire

Link to coring tool information

Arnaud F., Pignol C. et al., OZCAR-TERENO-2021

Variables d'un carottier, extracteur et trousse coupante



ANF « Carottage continentaux et lacustres », Aiguebelette, 17-20 octobre 2017

Qu'est ce qu'une « CONFIGURATION » de carottier ?

⁉ Selon les carottiers, multitudes de variation de 'Configuration' autour de :

Les outils

Carottier
(Piston + Lance/Liner + pointe)
+
Extracteur (core-catcher)
+
Trousse Coupante / Ogive
+
Couronne

CONFIGURATION DE CAROTTIER

⁉ **Les 'configs'** entre dans le **référentiel envoyé au Corebook**

UseCase 2
Provenance for Analytical Data

➔ Well structure data in tables

➔ MAKE REPETABLE "TEMPLATE"

What reference metadata standard for this type of data ?

- ***Sheet Tab*** « ***Read-me*** » : General & Context information contexte (from MD std) + thematic control vocabulary

(Geoflow template slide 55)

- ***Sheet Tab*** « **Data dictionary** » :

  -variable(s) names (+ definitions + control vocabulary)

  -unit of measure (+ control vocabulary),

  -quality : define unmeasured values (9999, -9999), n/a, value below the measurement …), uncertainty, standard calibration, suppressed value, ….

Keep PROTOCLES close to your data & share them !

➔ **well describe your lab or study protocols**

1- txt procols, sheet of laboratory notebook, Electronic labnotebook (ELN)

2-publication of a protocols

=> Link DOI to your article, Data deposit or Data paper

protocols.io

PLOS Published Peer-Reviewed Protocols

nature protocols

protocol exchange

"Five keys to writing a reproducible lab protocol", Nature, 2021

Provenance : Analytical Data
Well structure data in tables

UseCase 2

- ***Sheet Tab « Rawdata »*** (if necessary, otherwise link to the permanent repository wich store row data)
- ***Sheet Tab « Clean Data »*** *: only variables values, Good name for heider of colonnes (vocabularies?)*
- ***any information other than definine in the Read-me***

➔ **Do VERSIONNING** data as your work progresses
(with date consistent name : AAAA-MM-DD_xxxxx.xls)

- ***Sheet Tab « Work in progresse Interprétation session date 1 »***
- ***Sheet Tab « Work in progresse Interprétation session date 2 »***
- ***Sheet Tab « Work in progresse Interprétation session date 3 »***

Keep what you do at the end of time session

ROZA : standardised data & metadata template

Exemple of a 2014 « READ ME » of a sheet from a projet « ROZA »

UseCase 2

Usual name

IGSN

Usage license

NOMCAROTTE	PDU 14P 1w	SESAR IGSN	IGSNEDYFER0025	LICENCE UTILISATION (Creative Commu
METADONNEES / REFERENCES				
Version fichier Reference EDYTEM_XRF_data_12/05/2015				
WHAT KIND OF ANALYSE	XRF core scanning			
WHERE / LAKE	Paule			
WHO / NAME RESEARCHER	Pierre Sabatier			
WHEN	25/01/2016			
WHY	Application	PALAS		
DOI				
Tube lenght (mm)	890			
Top sediment (mm)	38			
Bottom sediment (mm)	889			
Sediment lenght (mm)	851			
Tube Diameter	63mm			
INTRUMENT & SETTINGS	AVAATECH Core Scanner			
Source	Rhodium			
Voltage (kV)	10	30	50	
Intensity (mA)	1,2	0,75		
LiveTime (s)	30	30		
Step (mm)	1	1		
Measured Area (CroscoreDownCore)	15*1	15*1		
Filter	No filter	Pd-Thin	Cu	
ELEMENT Soconcerned :	Al_Area	Cu_Area	Ag_Area	
	Si_Area	Zn_Area	Ba_Area	
	P_Area	Ga_Area		
	S_Area	Br_Area		
	Cl_Area	Rb_Area		
	K_Area	Sr_Area		
	Ca_Area	Y_Area		
	Ti_Area	Zr_Area		
	Cr_Area	Nb_Area		
	Mn_Area	Mo_Area		
	Fe_Area	Pb_Area		
	Rb_Area	Bi_Area		
Standards (Yes/No, Local Number Certified standards Names)	YES	134	MESS-3, SARM-4	
Operator	Anne-Lise Develle (orcid:0000-0002-4441-06...)			
Processing softwares	Winaxil, Winaxilbatch (Avaatech)			

Sample characteristics
(from Cyber-core repository)

Measurements settings

Frome Lab Note book

Measurement variables + definition
(here: list of elements)

Operator's acknowledgement (with ORCID)

Reusable

Ongoing effort

When possible, use thesaurus/controlled vocabularies (CHEBI; IUPAC; QUTD; ...) to provide unambiguous variables and units (interoperability)

➔ **Referenced in metadata catologs**

Arnaud F., Pignol C. et al., OZCAR-TERENO-2021

Provenance : Analytical Data
Well structure data in tables

UseCase 2

A Data Dictionnary

	A	B	C
1	**Data sheet names**	**Data description**	**Name of the file where data are located**
2	**Water silicone**	These data are giving the concentration in the water of the different PCB, concentrations calculated with two methods : Smedes and Yates. These data are named like this beceause silicone is used as a medium for the analys of PCB contained in water with a chromatograph.	PCB_Muzelle_2014 or PCB_Muzelle_2015
3	**Air**	These data are giving the deposit flow between air and water of PCB for the both phase of the atmosphere : dry air and wet air.	PCB_Muzelle_2014 or PCB_Muzelle_2015
4	**Sediment trap**	These data are giving the sedimentation flow of PCB at the surface and the bottom of the lake .	PCB_Muzelle_2014 or PCB_Muzelle_2015
5	**LOD**	These data are giving the limit of detection for each PCB in this study.	PCB_Muzelle_2014 or PCB_Muzelle_2015
6			
7	**Remark to users**		
8	Remark	All the given data in the files are first scale data. Some "second scale" data are available but th fluxes of transfers between compartments and mass balance in the lake. They are calculated b have acces to it by contacting Christine Piot, Edytem, Savoie-Mont Blanc University.	
9			
10			
11	**Signification of abreviation and specific values in both data (attribute) dictionnairy and datafiles**		
12	**Cw**	Concentration in water	
13	**<LOD**	Measures below the limit of detection. In chormatography the LOD is equal to three times the value of the background noise.	
14	**Profile**	Ratio between concentration of compound and the sum of contentrations of all compounds in percent	
15	**PCB**	Polychlorinated biphenyl compounds	
16	**NA**	Not Available	

A1 | Attribute / variable

	A	B	C	D
1	**Attribute / variable**	**Title**	**Définition**	**Comments**
2	**PCB 28**	Polychlorinated biphenyl 28	N°cas : 7012-37-5	7012-37-5@https://commonchemistry.cas.org/detail?cas_rn=7012-37-5&search=7012-37-5
3	**PCB 52**	Polychlorinated biphenyl 52	N°cas : 35693-99-3	35693-99-3@https://commonchemistry.cas.org/detail?cas_rn=35693-99-3&search=35693-99-3
4	**PCB 101**	Polychlorinated biphenyl 101	N°cas : 37680-73-2	37680-73-2@https://commonchemistry.cas.org/detail?cas_rn=37680-73-2&search=37680-73-2
5	**PCB 118**	Polychlorinated biphenyl 118	N°cas : 31508-00-6	31508-00-6@https://commonchemistry.cas.org/detail?cas_rn=31508-00-6&search=31508-00-6
6	**PCB 153**	Polychlorinated biphenyl 153	N°cas : 35065-27-1	35065-27-1@https://commonchemistry.cas.org/detail?cas_rn=35065-27-1&search=35065-27-1
7	**PCB 138**	Polychlorinated biphenyl 138	N°cas : 35065-28-2	35065-28-2@https://commonchemistry.cas.org/detail?cas_rn=35065-28-2&search=35065-28-2
8	**PCB 180**	Polychlorinated biphenyl 180	N°cas : 35065-29-3	35065-29-3@https://commonchemistry.cas.org/detail?cas_rn=35065-29-3&search=35065-29-3
9	**Sum PCB**	Sum of PCB		
10				

Read me PCB | key-words PCB | Water silicone | Air | Sediment trap | LOD

LICENCES : explicitly define the rights of reusers regarding your data

-Should they cite you?
-Can they modify the data?
-Make commercial use of it?

The choice depend on the **rights you can grant** depending on the legal **nature of the data**

*For example, they may already be protected by **copyright** (droit auteur) or be subject to **exceptions** to their open access (collaboration with private, military…).*

LICENCE OUVERTE
OPEN LICENCE

La Licence Etalab a été conçue par le Gouvernement français pour faciliter la mise en place de l'Open Data. Elle équivaut à la licence CC-BY.

ODC-BY, CC-BY 2.0

creative commons

Les licences CC permettent de définir plusieurs restrictions, comme l'interdiction d'usage commercial ou de modification.

CC0 / CC-By / CC-By-SA …

ODbL

L'Open Database Licence (ODbL) est une licence spécifique permettant d'exploiter publiquement des bases de données.

In France : Institutions get an « Déléguée à la protection des données (DPO)

Careful with **"Personal Data »**

The EU-General Data Protection Regulation – GDPR **(RGPD)**

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
data.gouv.fr
List of licences

➔ A **metadata catalogue must inform** the licenses of the dataset

➔ A **data repository must provide** the chosen license for its reuse

Ca commence a bien F.A.I.R !

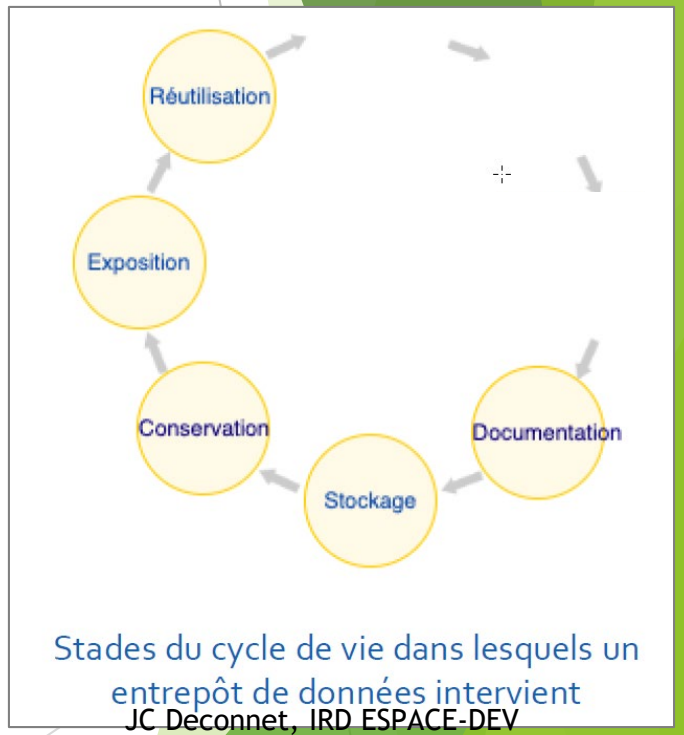
5 minutes ?

Why deposit your research data in a data-repository ?

Depositing your data in a warehouse (data and metadata) ensures their preservation, visibility and access, thus facilitating their sharing and reuse

Several benefits :

- ✓ compliance with the **Opendata recommendations of funders** and institutions
- ✓ **data visibility** and easy access for catalogues or search engines discovery
- ✓ **reuse by third parties** and **citation** of the dataset facilitated (PID identifier)
- ✓ management of data sharing arrangements through **licensing**
- ✓ **data interoperability** (use of metadata standards and vocabularies)
- ✓ data retention in a **secure environment**



Stades du cycle de vie dans lesquels un entrepôt de données intervient
JC Deconnet, IRD ESPACE-DEV

➔ **improved research reproducibility, integrity and scientific validation**

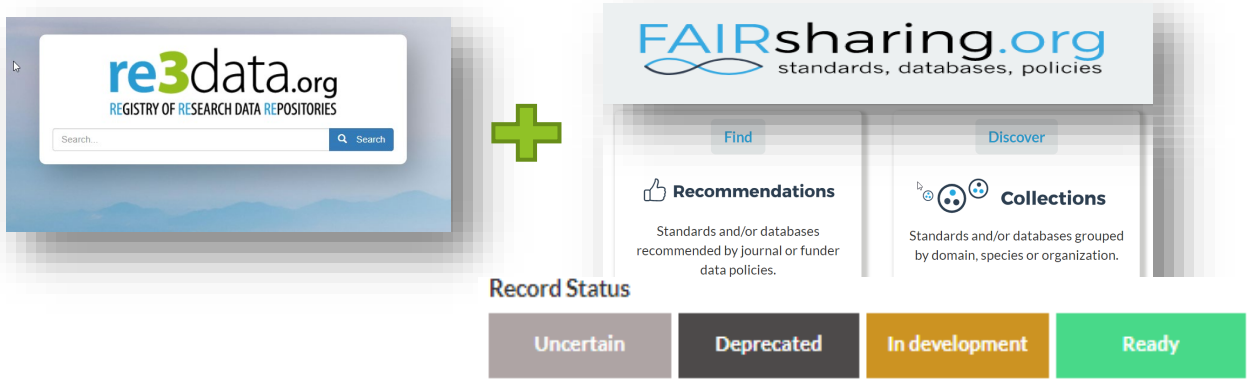
➔ **reuse in new studies and innovations**

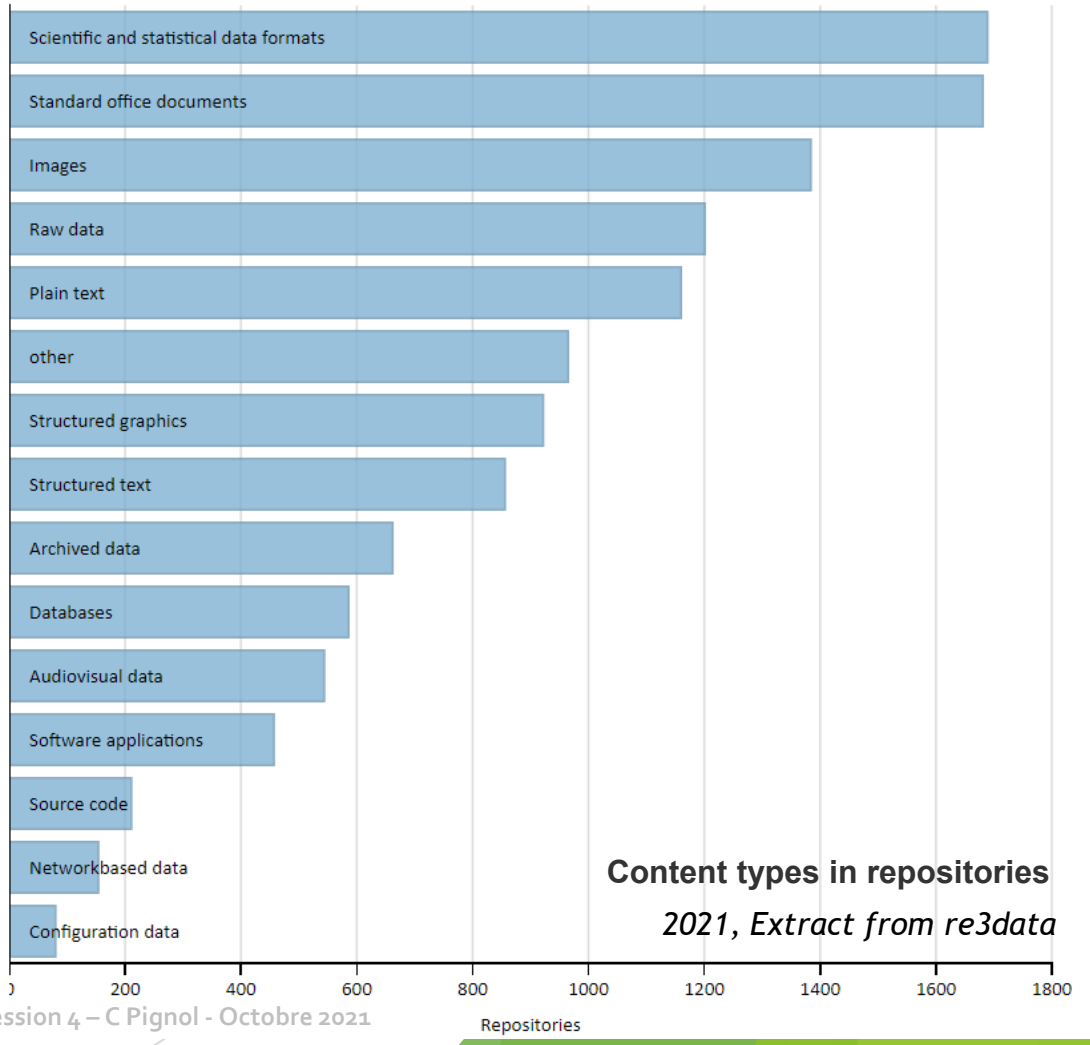
How to find a data-repository ?

https://doranum.fr/depot-entrepots/criteres-choix-entrepot/

https://sist20.sciencesconf.org/data/pages/SIST20_entrepot_de_donnees_Desconnets.pdf

1) Searching repository by « annuaire » (directories) or according editor recommandations





➔ **Search for storage** « Subject(s)" =
Oceanography Geology and Palaeontology Geochemistry

➔ ***Nature*** **Data Repository Guidance** (+90 entrepôts recommandés)
Scientific Data repository

Where : how to choose a data repository ?

Si vous ne savez pas lequel choisir, vous pouvez vous appuyer sur des critères présentés ci-dessous :

https://doranum.fr/depot-entrepots/criteres-choix-entrepot/

2) Aspects to be checked directly on the website :

- ✓ **Easy deposit : easy to use ?** user-friendly, ergonomic? Some warehouses also offer support, tutorials
- ✓ **The quality of the description (metadata) :** The quality of the description of the datasets is important for easy retrieval (**based on standards ?**)
- ✓ **Usage statistics :** Does the warehouse offer usage, consultation and data download statistics ?

3) check the warehouse "policies":

Exemple, consulting the Policy (*Zenodo free & open*)

- ✓ **Server hosting location** : in Europe ? US ? Data sovereignty ? Lockdown ?
- ✓ **Long-term preservation of data : how many years ?**
- ✓ **Cost of deposit**
  Dryad, *Figshare* are paid/ Pangaea, Zenodoo are free.
  The cost can also be related to the volume of files
- ✓ **Statut public / academic or private ?**
- ✓ **Supported data types**
  Dryad, only accepts data related to a publication.
  Zenodo, accepts any type of data.
- ✓ **Accepted file formats**
  recommend specific file formats.
  (Non-proprietary formats, list recommended formats, preferred format ….)
- ✓ **Link(s) with articles ?**

A Comparative Review of Various generalist Data Repositories (1)

Extraction of 4 repository

From 2018

https://dataverse.org/blog/comparative-review-various-data-repositories

	Dryad	figshare	Mendeley Data	Zenodo
Data citation automatically generated	Yes	Yes	Yes	Yes
Ability to embargo files	Yes	Yes	Yes	Yes
Analyzing tabular data (aside from geospatial mapping)	No	No	Yes	No
Filetypes supported (list)	"All data types and formats within reason"	All file types	All file types	All file types
Handling large data	Yes - Limit of 20GB per file for "package", charge for more storage	Yes - Default limit of 5GB per file, can support up to 5TB	Planned Q2 18	Yes - limit of 50GB per file. Upload page mentions you can contact them about larger datasets.
Mapping of Geospatial files	No	Yes	Yes	No
Previewing and/or analyzing uploaded non-tabular data	No	Yes	Yes (preview images, tabular data)	Yes (previewing, not analyzing)
Previewing tabular files	No	Yes	Yes	Yes
Provenance	No	No	Yes	No
Users are able to control dataset file hierarchy + directory structure	No	No (can re-order files, but can't create hierarchy or directories)	Planned Jan 18	No
Robustness of metadata	Yes	Yes	Yes (depositors can describe files today;	Yes
Support for controlled vocabulary terms with URIs	?	No	No	Yes
Tracking citations with altmetrics	No	Yes	No	No
Ability to access older versions of files	Yes - on data package page	Yes	Yes	Yes
Dataset-level versioning	Yes	Yes	Yes	No
Faceted search	Yes	Yes	Planned Q1 18	Yes

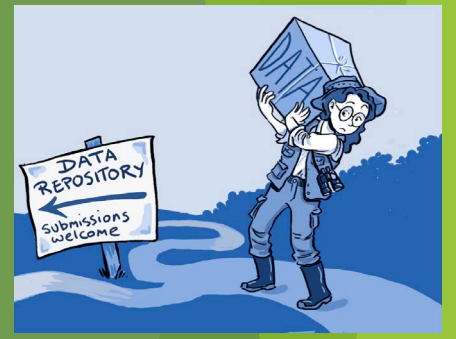
A Comparative Review of Various generalist Data Repositories (1)

Extraction of 4 repository

From 2018
https://dataverse.org/blog/comparative-review-various-data-repositories

	Dryad	figshare	Mendeley Data	Zenodo
Permissions	Yes (ability to apply embargo)	Yes (ability to apply embargo or make data "confidential")	Yes, with embargo	Yes (Includes embargoed, restricted, and closed access)
Terms of use, copyright	No - all uploaded data is CC0	Yes	Yes	Yes
Open Source	Yes	No	No	Yes
How is maintenance or development of repository funded?	Grants; Service charges	Unknown (Funded by Digital Science while retaining autonomy)	Private investments (Elsevier; Mendeley Ltd)	Institutional (OpenAIRE / European Commission); donations (via CERN &
Free To Use	No	Yes	Yes	Yes
Institutional Fees?	No	Yes - "Figshare for Institutions" is priced based on "research intensity of the institution"	No	No
Journal Fees?	No	Yes - "Figshare for Publishers"	No	No
Paid services?	Yes	Yes - dedicated support team	No	No
Can depositors pay for extra storage?	Yes - "For data packages in excess of the 20GB size limit, submitters w ill be charged $50 for each additional 10GB, or part thereof. (Packages betw een 20 and 30GB = $50, betw een 30 and 40GB = $100, and so on)."	No	No	No
Preservation technology?	Yes - DataONE	Yes - Chronopolis	Yes - Amazon S3 servers in Ireland and dark archive storage with DANS	Yes - All data files are stored in CERN Data Centres, primarily Geneva, with replicas in Budapest.
Most popular subject tags	Adaptation / Population Genetics - Empirical / Speciation	Unknow n	Unknow n	Taxonomy / Biodiversity / Animalia
Subject vocabulary	Folksonomy	Australian and New Zealand Standard Research Classification (ANZSRC), 2008	Elsevier's OmniScience taxonomy	Folksonomy

Exemple of data-repository non-exhaustive



- **Institutional (in France)** : Dataverse Cirad, Datapartage (INRAe), DataSuds (IRD),

DataTERRA (4 Pôles de données ODATIS, AERIS, THEIA, Form@Terre),

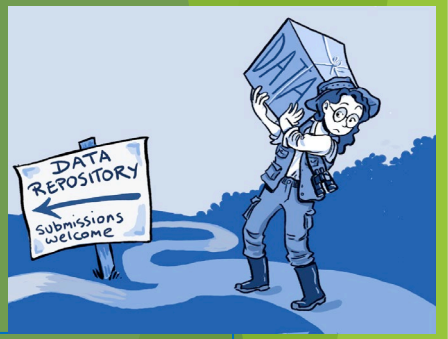
In nov 2021 Dataverse InDORES (INEE-CNRS) , in mars 2022 RecherheDataGouv

- **Institutional (in Europe)** : NERC Data Centres (GB), …
- **International multidisciplinary** : Zenodo, Dryad, Figshare
- **Own by editors** : GigaDB (Oxford Univ. Press), Dataverse Ubiquity Press,
- **Thematic and disciplinary : Repository or Database !**

GenBank (séquences génétiques), TRY (caractères botaniques), GBIF (biodiversité), NOAA , HydroShare (CUAHSI), EarthChem, Pangaea (sc de la terre&environnement)

- **For « Protocls » :** Protocols.io (protocols), Nature ProtocolsExchange, Plos Protocols , etc. (*txt, video*)
- **For «Code and software » :** Software Heritage (or via HAL)

Repository Data-repository in connection with other devices/ infrastructure



Stockage

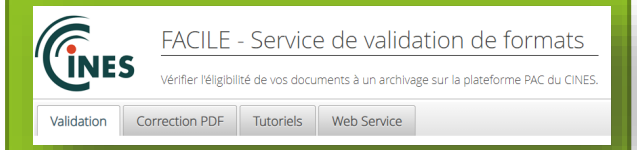
≠

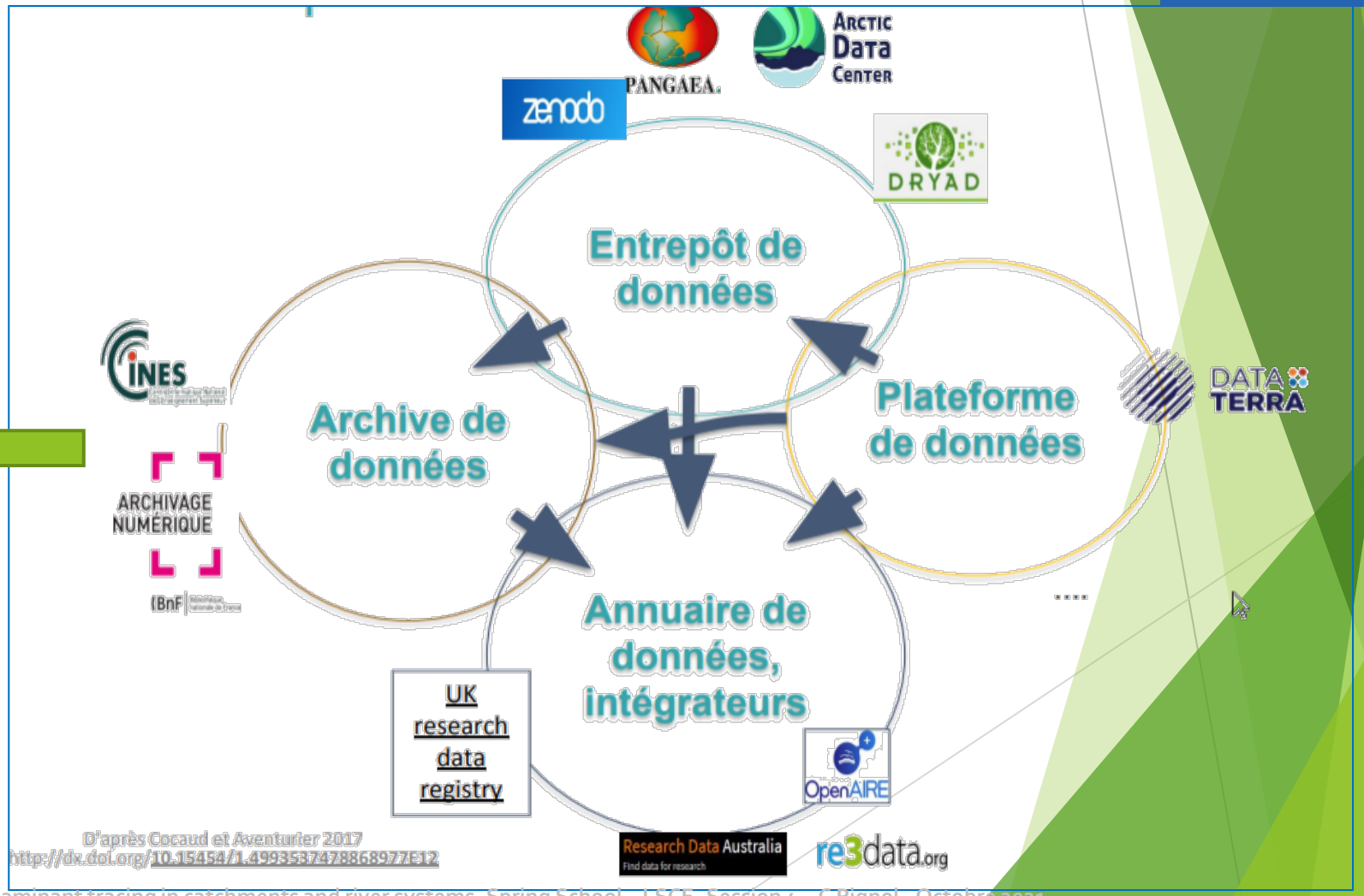
Archive

Focus

data archive
longevity of format file

https://www.cines.fr/archivage/des-expertises/les-formats-de-fichier/

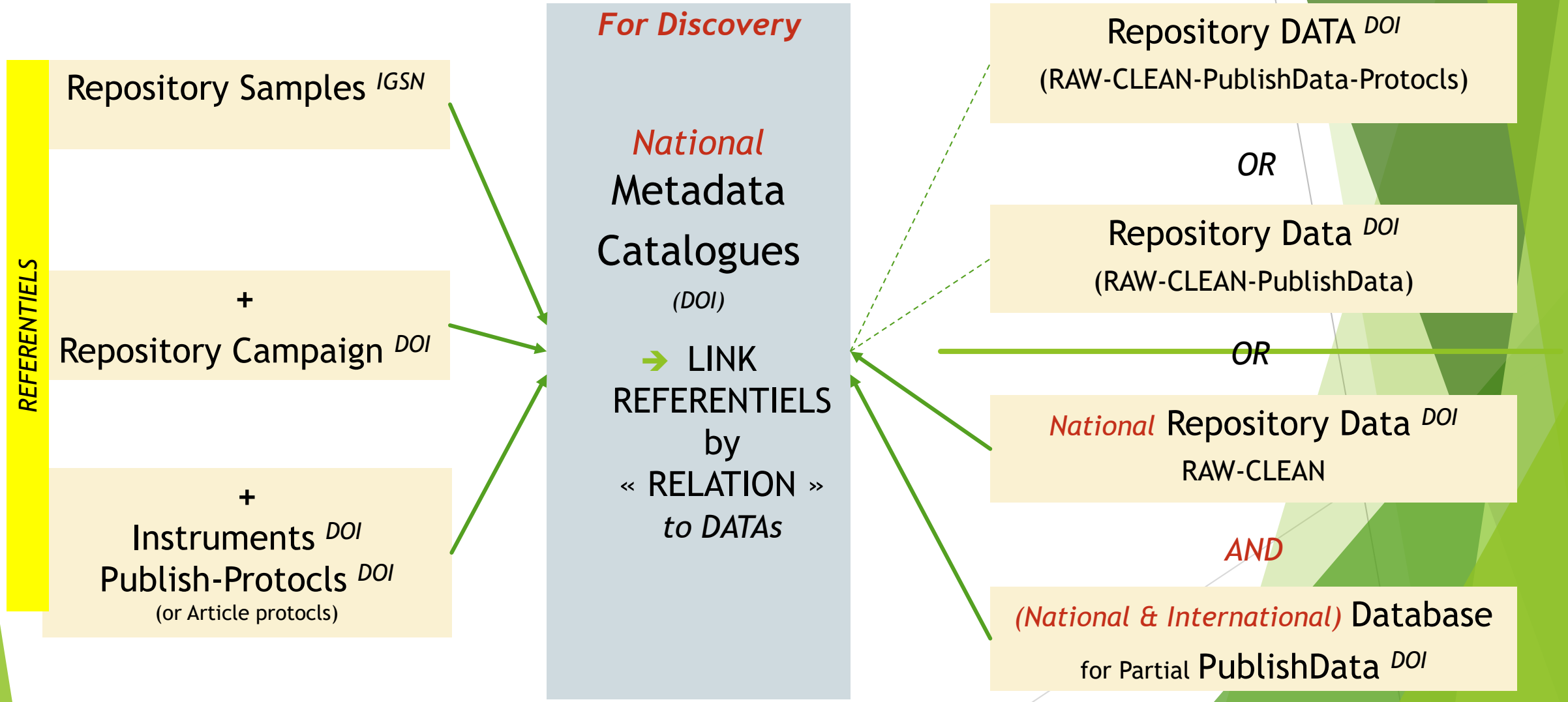




D'après Cocaud et Aventurier 2017
http://dx.doi.org/10.15454/1.4993537478868977E12

How repository-data network could be structured next ?

It's depend when you do it in your Datacycle !



An example of tools that can help

Geoflow a metadata management workflow for **everyone** !

➔ ***1a,b,c*** - **You prepare your data** (at differents moment of the collect

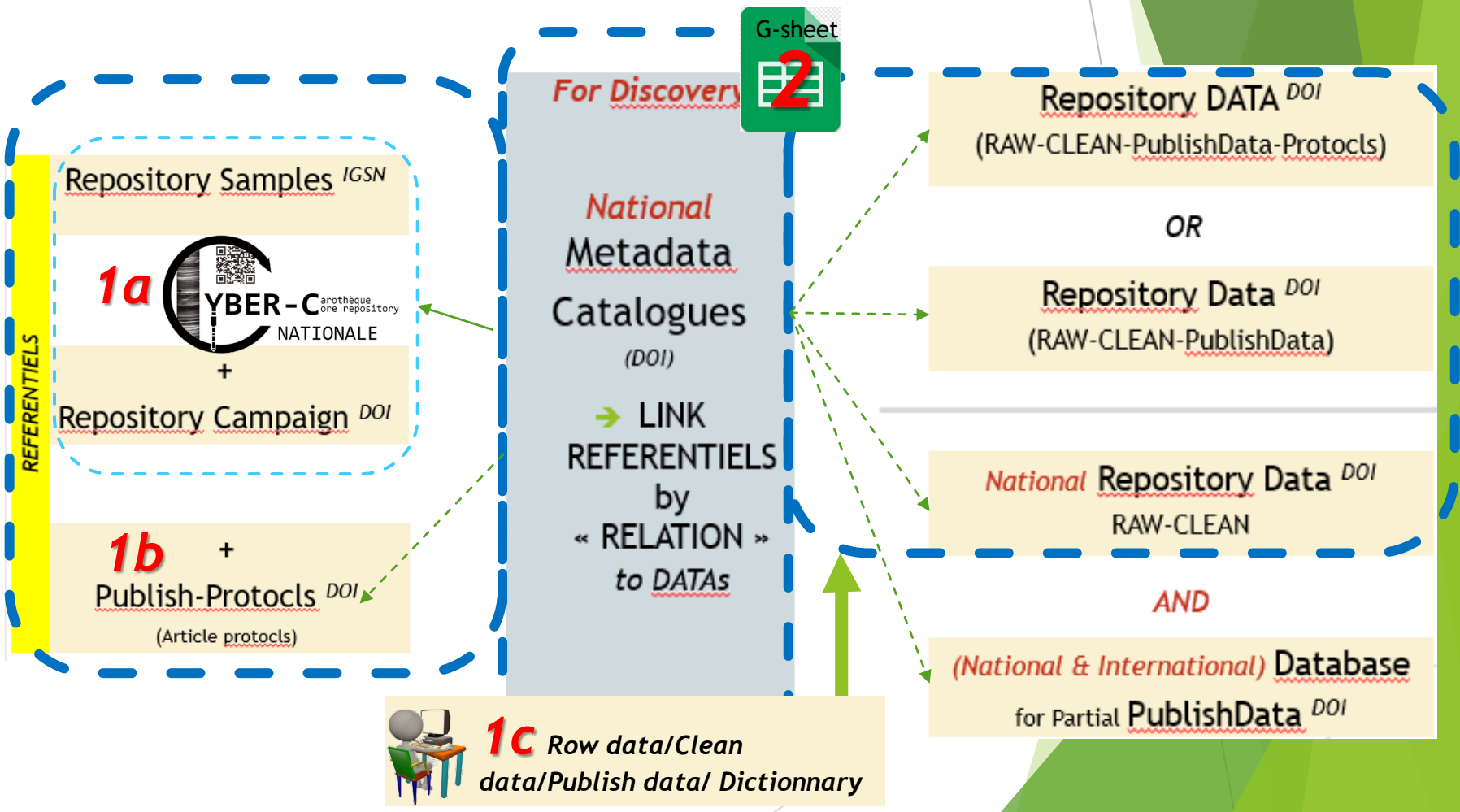
➔ ***2*** – You use **Geofow** to expose Metadata & Data

➔GEOFLOW

Multi metadata pivot for publishing ***in one time***

to catalogMD and Data Repository

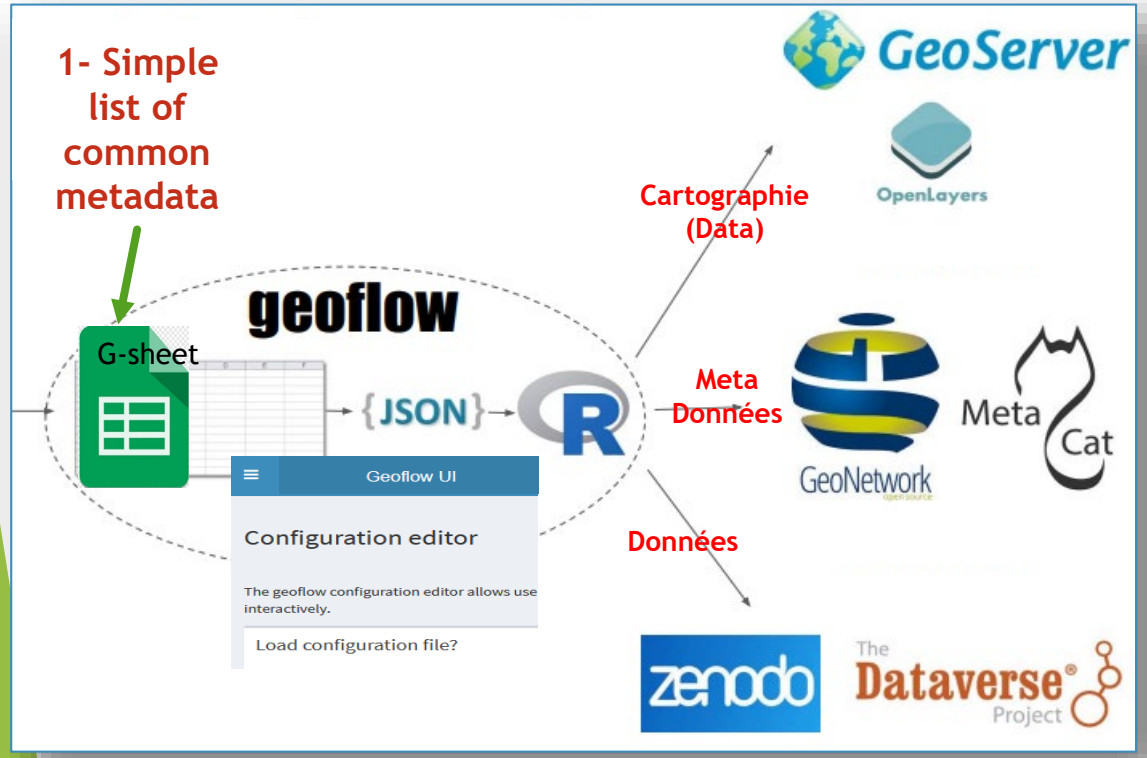
https://github.com/eblondel/geoflow



An example of tools that can help you

Geoflow a metadata management workflow for everyone !

Workflow



2- Text Paremters *to link Metadata & data to catalogues & repository*

R-coded metadata management tool, (supported by the R Consortium) ***to facilitate the production of standardized metadata*** *(DublinCore, Datacite, ISO 19115, Ecological Metadata Language)* ***through a very simplified interface*** *(R-shiny interface-test* ***(INRAE)****, enabling* ***collaborative metadata editing (Googlesheet,…)***

geo

Exemple of the geoflow guidline :

Part of metadata in the geoflow workflow

	A	B	C	D	E
1	**Colonne**	**Type de données**	**Définition**	**Besoin**	**Nomenclature clés - valeurs**
5	Subject	**Liste de mots-clé décrivant le jeu de données**	Sujets / matières associés au(x) jeu(x) de données. Un sujet se résume par un ou plusieurs mots-clés. Un sujet peut être composé de plusieurs mots-clés. On sépare les mots-clés d'un même sujet par une virgule. Plusieurs sujets peuvent décrire un jeu de données. On sépare les sujets par des underscores. Un mot clé peut être associé à un thésaurus. Dans ce cas on associe l'URL au mot clé avec @.	Recommandé	Pas de clé par défaut. Il faut créer soit même ses couples de "clé:valeur". Pour ajouter un sujet : utiliser comme clé le type de mot clé ou bien le thesaurus utilisé. On a ainsi des couples "clé:valeur" de cette forme : clé --> libre valeur --> [mot clé] ou [mot clé@url] clé-libre:mot-clé1,mot-clé2@url
6	Creator	**Noms des parties ayant un rôle dans la gestion des jeux de données.**	Noms des parties ayant un rôle dans la gestion des jeux de données. Chaque partie citée ici doit apparaitre dans un autre tableur : le tableur contact. Une personne peut être citée dans plusieurs catégories. Catégories : - owner pour les personnes auteures possédant les jeux de données - publisher pour les personnes responsables de la publication des jeux de données - metadata pour les personnes responsables de la gestion des métadonnées - pointOfContact pour les autres contributeurs, les personnes à contacter pour plus d'information. La catégorie owner est celle fortement recommandée.	Recommandé	owner:mail personne1,mail personne 2_ publisher:mail personne 1_ metadata:mail personne 3_ pointOfContact:mail personne 2,mail personne4
7	Date	**Date associée au jeu de données, généralement sa création.**	Dates associées au jeu de données.	Recommandé	creation:YYYY-MM-DD ou YYYY
8	Type	**Type de données (vecteur, texte, tabulaire,...)**	Type de données (vecteur, texte, tabulaire,...)	Recommandé. Si ce champ n'est pas renseigné "dataset" sera choisit par défaut.	generic:dataset
9	Language	**Langue des données et fiches de métadonnées.**	Langue des données et fiches de métadonnées. La meilleure pratique est d'utiliser l'ISO3 Ppour les deux cas les plus courant (anglais et français) indiquer en abréviation les trois première lettre de la langue en anglais (respectivement : eng et fre).	Recommandé. Si ce champ n'est pas renseigné "eng" est choisit par défaut.	Pas de clé. Mettre directement le code langue. Ex : eng
10	SpatialCoverage	**Couverture spatiale des données et système de coordonnées (SRID) correspondant.**	Couverture spatiale des données et système de coordonnées (SRID) correspondant. Attention ! On sépare le SRID du WTK avec un point virgule ; lien pour dessiner l'étendue géospatiale : https://arthur-e.github.io/Wicket/sandbox-gmaps3.html	Optionnel	SRID=code;WKT
11	TemporalCoverage	**Couverture temporelle des données. Peut être un moment donné ou une période.**	Couverture temporelle des données. Peut être un moment donné ou une période.	Optionnel	**Pour un moment donné :** YYYY-MM-DD ou YYYY 2021-08-01 ou 2021 **Pour une période :** YYYY/YYYY ou YYYY-MM-DDTHH:MinMin:SSZ/YYYY-MM-DDTHH:MinMin:SSZ 2010/2015 ou

-PART IV-

End of the trip !

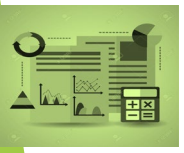
You have some keys to make a :

- Data Mangement Plan (DMP) for Project,

assisted by

Plateform,
Database,
…

datamanger
librarian



- Data Paper

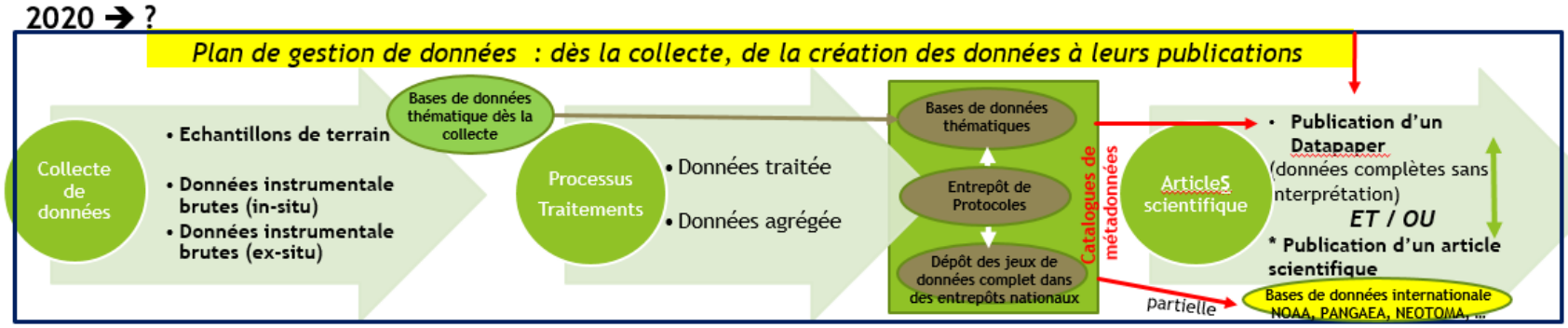
Researcher **make homogeneous collection of data on a purpose**

Researcher **don't have <u>time</u> (or <u>needs)</u> to exploit all mesurement data**

Data Mangement Plan (DMP) for Project

➔-It is a **DOCUMENT** *(deliverable follows a frame)* that **FEEDS** THROUGHOUT the PROJECT/THESE

➔- **LINKS** the overall *strategy* and *results* for **ALL DATAs** *(raw -> published)*



➔ So what's the plan ?

Start at the begining, not at the middle or the end !

https://doranum.fr/plan-gestion-donnees-dmp/

https://coop-ist.cirad.fr/gerer-des-donnees/se-familiariser-avec-les-pgd/3-exemple-de-trame-d-un-plan-de-gestion-de-donnees

Data Mangement Plan (DMP) for Project

Differents Templates for the same result

https://coop-ist.cirad.fr/gerer-des-donnees/se-familiariser-avec-les-pgd/6-des-modeles-de-plans-de-gestion-de-donnees-en-anglais

- **In french / english**
  - modèle Horizon 2020 english
  - modèle français/anglais de l' ANR
  - modèle INRAe, IRD,…

DMP OPIDoR

Or
Template in clood collaborative

Collaborative work
PI project + 1 DMP Coordinator
+ All the participant of the project …

- **In English**

-DMP Horizon 2020 Guidelines on FAIR Data Management in Horizon 2020: Version 3.0. European Commission, 26 July 2016, 12 p. (version anglaise)

-Digital Curation Centre's DMPonline tool DMPonline (2018)

-University of California's DMPTool

-ESRC Data management plan: guidance for peer reviewers

-NERC Data Management planning

DMPONLINE

DMPTool

Data Paper ?

pair-rewiew publication (Rang A)
made aware for Raw, Cold, unpublished data

In DataPaper = NEVER data analysis !
- ✓ **No hypothesis**
- ✓ **No interpretation**
- ✓ **No derived conclusions**

▶ Data Paper « Data compilation » type

Synthesis / compilation of data scattered according to criteria

Work of grouping coherent (generally published) datas (on the same theme, of the same variables, on the same or similar site, etc…

➔ But some of them have an intention in the collection

▶ Data Paper « Observatory » type

Collection protocls, data organization

1 year of photography in a research site
Only 2 years of monitoting a site
Annual data from a plateform (with embargo)

▶ Data Paper « data collection » type

Provision of data collected but even used in a project, PhD, …

Hight metadata level LEVEL 2

avoid falling into « the long tail of » Darkdata

https://doranum.fr/data-paper-data-journal/
https://coop-ist.cirad.fr/gerer-des-donnees/rediger-un-data-paper/1-qu-est-ce-qu-un-data-paper

Data Paper (2)

➔ ***Nature*** **Data Repository Guidance** *Scientific Data* repository
➔ ***A list of peer-review Data papers :*** *PDF* *(CIRAD, 2018)*

NEVER ! data analysis
✓ No hypothesis
✓ No interpretation
✓ No derived conclusions

- ***Data paper*** inform the scientific community of the ***availability of these datasets*** and their ***potential*** for future use

Describes scientific data :

➔ **circumstances** and **methods** of its collection.

➔ **Quality** : technical and statistical analyses for validating data

- ***Data paper*** shows the ***originality and scope*** of the dataset it describes

Journals publishing data papers are particularly interested in

➔ **the scope** of the data submitted

➔ **The potential for reuse** by other scientists (major argument)

Hight metadata level LEVEL 2

➔ describes a set of scientific data
➔ using precise informations (metadataS generalist, interdisplinary & thematics, quality), scope, potential …
➔ data storage in a permanent warehouse (DOI)

Check-list « What can I do ? »

<u>At the beginning</u> (or whenever possible) of your PhD or Project

Check-list (1)

ORCID

Clean your personnel id network
➔ only one ORCID or one IDHal

1- Think beforehand about all this actions :

➔ good **hierarchical folder/file tree** (raw, processed, a aggregated). Think you could need *to share it* at the end

➔ **properly name** your folders/files with consistent name/date, version,…

➔ **search the metadata standards of your thematics** (sample, geospatial, climate, chemistry, etc.) and **vocabularies** (from generic to specific) to describe your variables

DON'T WAIT if :

- ERC, ANR(>2020)= DMP mandatory
➔Project manager take the template of a DMP and think about a plan !
(Take a good coordinator & collaborators to help you along the project)

- Anticipation of a Datapaper ("observatory" or "collection" type)
➔ prepare files as soon as they are collected in connection with the acquisition platform & metadata standard

-Senior Researcher
-Data Scientist
Help The young !

Workshop
with your expert community
to decide your specific
Guideline & Template data

*(Before see **thematic's** Groups in the Research Data Alliance)*

Thematics interest groups

RDA RESEARCH DATA ALLIANCE

https://coop-ist.cirad.fr/gerer-des-donnees/rediger-un-pgd/6-garantir-la-comprehension-et-l-accessibilite-des-donnees
https://doranum.fr/stockage-archivage/comment-nommer-fichiers/
https://dataservices.gfz-potsdam.de/portal/phd.html

Check-list (2)

2. **SAMPLE :** Properly **name and describe** your “Parent” and “Child” Samples (sub-samples)

➔ « Sampling plan” (and update if changed) => DMP

➔ Deposit in sample repository = get an IGSN identifier + search by catalogues

For Cores => French Cyber-core repository , For other samples CNRS AA IGSN, SESAR,

3. **PROTOCLES : well describe your measurement protocols**

(txt procols, sheet of laboratory notebook, Electronic labnotebook (ELN), publication of a protocols (article or Protocols repository)

4. **ANALYTICAL DATA :** well structure your data in tables ➔**MAKE TEMPLATES MODEL REPETABLE**

➔ ***Sheet Tab « Read-me »*** : with a **data dictionary** (variable name + unit of measure + vocabularies + definitions), define unmeasured values (9999, -9999, n/a, value below the measurement threshold ,…), suppressed value, ….

➔ ***Sheet Tab « rawdata »*** (if necessary, otherwise link to the permanent repository)

➔ ***Sheet Tab « data » : only variables values****, any information other than definine in the Read-me*

5. **Versioning** your data as your work progresses (with date : AAAA-MM-DD_xxxxx.xls)

Check-list (3) : 2 cases

Scientific Article Publication :

- ✓ I submit an article with the link to datas warehouse (samples, compaigns, protocls, data)

IN THE POSTPRINT PAPER (some editors ask links in the papier abstract)

- ✓ Data deposit with complete data dictionary and metadata (standard & vocabulary)

Datapaper Publication ? Why ?

-My research is an observatory (eg. without Database)

-Part of my research will lead the creation of an cross-checking of data

-I don't have time & needs & expertise to use all the dataset of an instrumental acquisition

-I don't have time to write an article

Don't aliment the long tail of data anymore !

➔ **I make a Datapaper**

Supply the infrastructure at your disposal *(IT Infrastructures available in France)*

- **Metadata catalogues :** OSUs Cat.(1), Réseau des Zones Ateliers (LTER) (1), Pôles DataTERRA (ODATIS, OZCAR-THEIA, FORM@Terre, AERIS) (1), Pole national de la Biodiv. (PNDB) (2), CNRS-INEE (IndoRES-MD Catalogue) (1)
- **Thematic Database (OSU, International,…)**
- **National Data repository :** Dataverses IRD, INRAe, OSU (Loraine), InDORES-data (INEE) *Nov2021,* Entrepôt national fédéré de la Recherche *(mars 2022)*

(1) Iso19115
(2) EML
(3) mixte

Conclusion of this trip ?

Faut pas trop s'en F.A.I.R ?

A lot of bibliography, websites' tools throughout the présentation and,

1. Gil, Y. *et al*. Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science* **3**, 388–415 (2016).

2. Khider, D. *et al*. PaCTS 1.0: A Crowdsourced Reporting Standard for Paleoclimate Data. *Paleoceanography and Paleoclimatology* **34**, 1570–1596 (2019).

3. Lehnert, K., Elger, K. & Ulbricht, D. IGSN – More than Unique Identification of Samples. (2020) doi:10.5281/zenodo.3776495.

4. McNutt, M. *et al*. Liberating field science samples and data. *Science* **351**, 1024–1026 (2016).

5. Morrill, C. *et al*. The Paleoenvironmental Standard Terms (PaST) Thesaurus: Standardizing Heterogeneous Variables in Paleoscience. *Paleoceanography and Paleoclimatology* **36**, e2020PA004193 (2021).

6. Peng, G. *et al*. *International Community Guidelines for Sharing and Reusing Quality Information of Individual Earth Science Datasets*. https://osf.io/xsu4p/ (2021) doi:10.31219/osf.io/xsu4p.

7. Monya Baker . Five keys to writing a reproducible lab protocol
Effective sharing of experimental methods is crucial to ensuring that others can repeat results. An abundance of tools is available to help.
Monya Baker. (*2021) https://www.nature.com/articles/d41586-021-02428-3