



Fabien Arnaud<sup>1</sup>, Cécile Pignol<sup>1</sup>, Bruno Galabertier<sup>1</sup>, Xavier Crosta<sup>2</sup>, Isabelle Billy<sup>2</sup>, Elodie Godinho<sup>3</sup>, Karim Bernardet<sup>3</sup>, Pierre Sabatier¹, Anne-Lise Develle¹, Rosalie Bruel¹,⁴, Julien Penguen⁵, Pascal Calvat⁵, Pierre Stéphan6, and Mathias Rouan6

# The pyramid of data: general case

**Publication** 2) Data described in 1) Data included supplementary files and explained in associated to articles the article **Processed data** 4) Data signaled in and 3) Data papers: Data representations describe available

Collections of data

and structured databases

Raw data

1 Environment Dynamics and Territories of the Mountain (EDYTEM), Université Savoie Mont-Blanc, CNRS, Chambéry, France (fabien.arnaud@univ-savoie.fr) 2 Environnements et paléoenvironnements océaniques et continentaux (EPOC), Université de Bordeaux, CNRS, Bordeaux.France 3 Division Technique de l'Instiut des Sciences de l'Univers (DT-INSU), CNRS, La Seyne-sur-Mer, France 4 Rubenstein Ecosystem Science Laboratory, University of Vermont, Burlington, VT, USA 5 Observatoire Aquitain des Sciences de l'Univers (OASU), Université de Bordeaux, CNRS, Bordeaux, France 6Littoral, Environnement, Télédétection, Géomatique (LETG), Université de Bretagne

Occidentale, Brest, France

articles and in data centers or repositories

> 5) Data in drawers, hard disks, floppy disks, notebooks, clay tablet...

Adapted from Report on integration of data and publications. Opportunities for data exchange (Relly et al., 2011)



datesets







### Publications are just the emerged tip of a huge iceberg of « frozen » data



Findable . Accessible . Interoperable . Reusable 2) Data described in supplementary files associated to articles

3) Data papers: describe available datesets



1) Data included and explained in the article

4) Data signaled in articles and in data centers or repositories

5) Data in drawers, hard disks, floppy disks, notebooks, clay tablet...

Adapted from Report on integration of data and publications. Opportunities for data exchange (Relly et al., 2011)

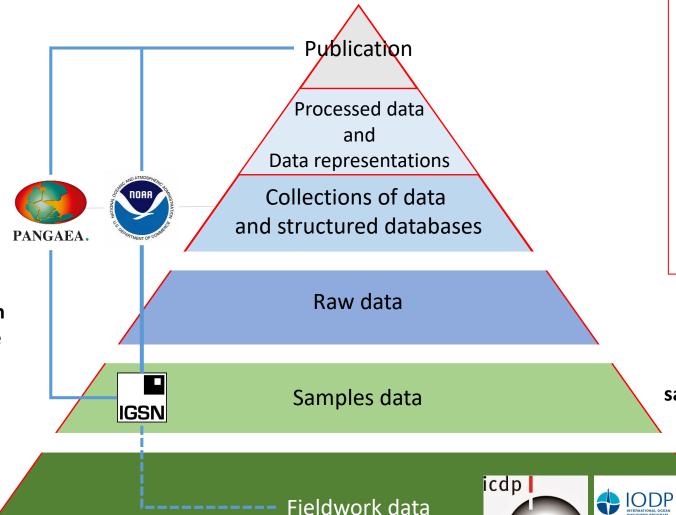




### The pyramid of data: case of field-based (geo)sciences

PANGAEA or NOAA portals permit to store and find published datasets

There is no raw data repository nor standard in sample-based geoscience



Samples and field data are the absolutely necessary foundations of the whole pyramid

However, they are often poorly banked

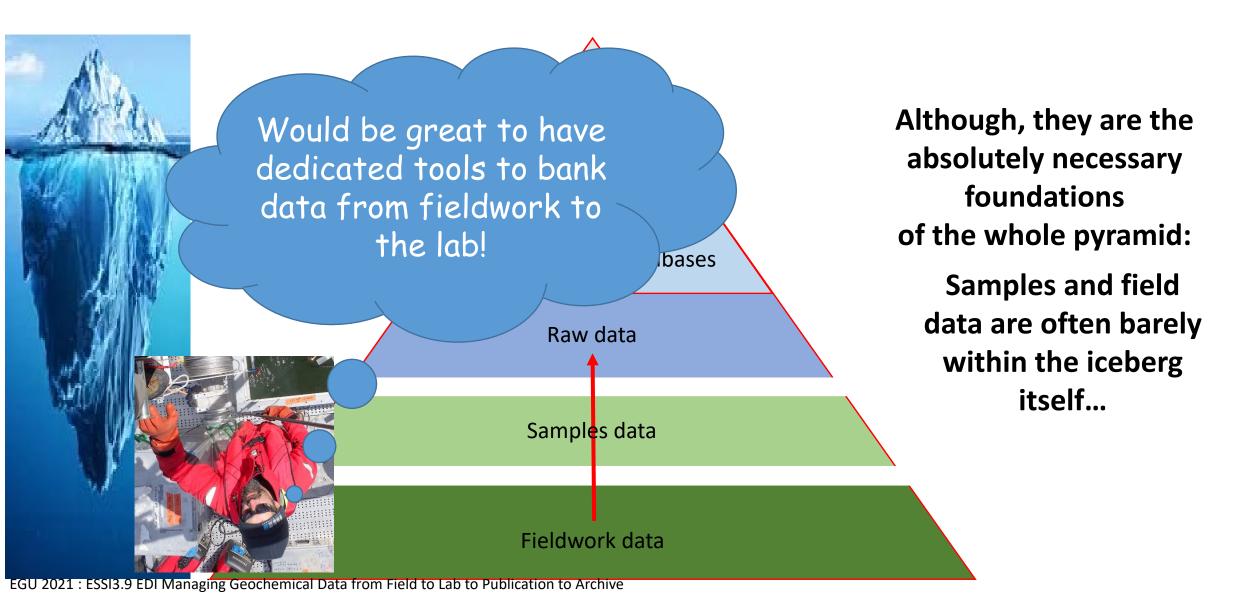
IGSN offers the opportunity to link samples and data but includes only poor fieldwork data

Fieldwork data are sometimes banked at single-project scale, not for every researcher





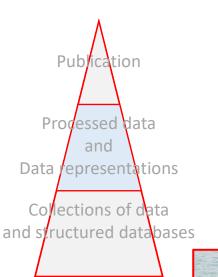
### The pyramid of data: case of field-based (geo)sciences







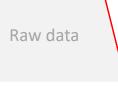
→ Catching fieldwork data at the source: CoreBook Android App.



Functions to help field operations (navigate, find older coring locations, placement of anchors, display of parallel cores)

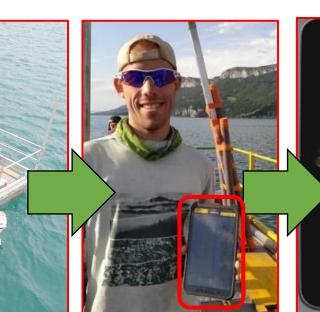
Operatorfriendly Data filling

XML exchange file Ready for input into a database



Samples data

Reusable Fieldwork data





xml





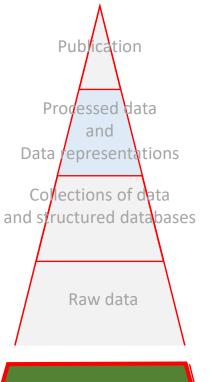
**→** Banking and displaying fieldwork data: the National Cyber-core repository **Cyber-carothèque nationale** 

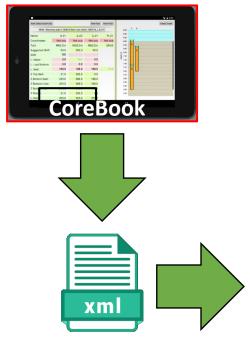


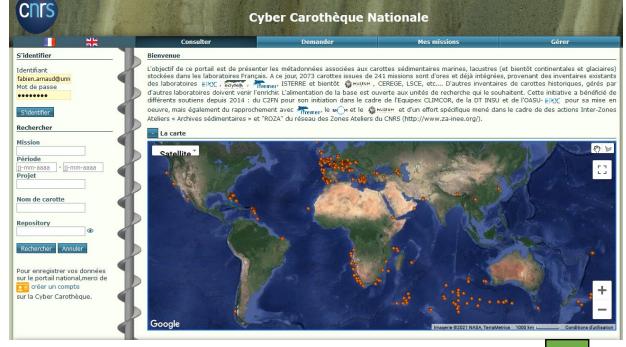












Fieldwork data

Samples data

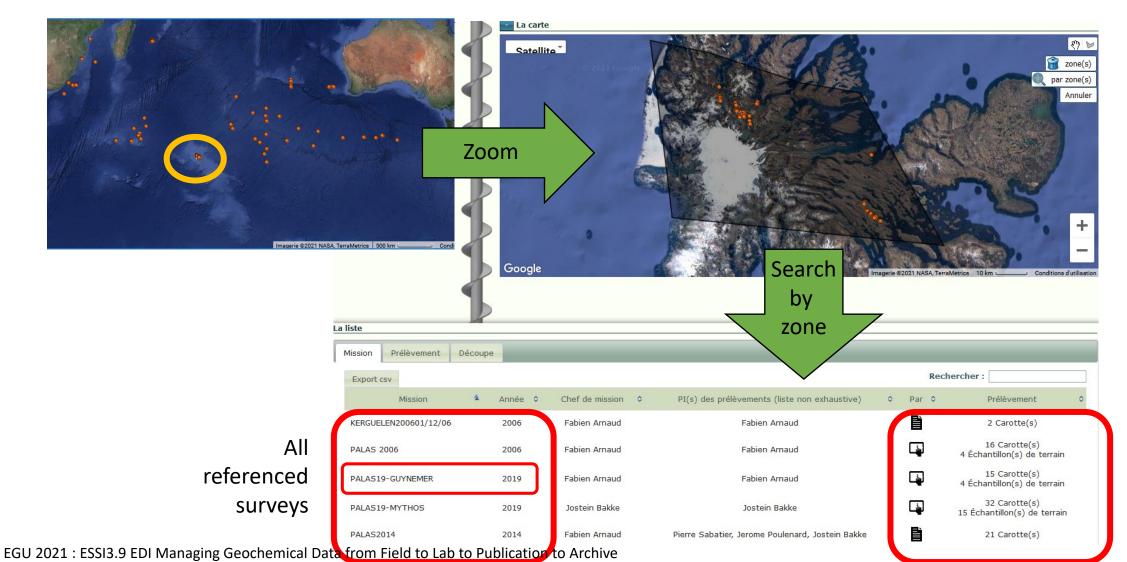
OGC Export geographical point & fielddata for GIS SYSTEM

Flow





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



All referenced cores and samples





### 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** Acknowledgement of Mission PALAS19-GUYNEMER crew members with Description **Participants** determined role and **XML** Fichier source Corebook **ORCID** identification Chef foreur Arnaud Fabien source file Chef de mission Arnaud Fabien Code Début de mission 27-11-2019 06:00 PI(s) des prélèvements (liste non exhaustive) Fin de mission 27-11-2019 18:00 Arnaud Fabien Description Financeur Aucun(e) Projet PALAS Documents liés Contexte scientif GUY19-I-Pil-05-20191223-141634.jpg GUY19-I-Pil-05-20191223-141716.jpg Connaissances prélimina GUY19-I-Pil-05-20191223-141736.ipg GUY19-I.20190619-195534.Coring.Schema.png Link to project: permit GUY19-I-B-01-20191201-064036.mp4 Commentaire to display all EPGUY19-01-20191204-113555.jpg EPGUY19-01-20191204-113630.jpg surveys/cores within Confidentialité Pub the same project Remerciements Any additionnal When applicable: link document (pictures, DOI to the sketches of **Automatic** survey URL sounds, videos) taken **CSR** report with DOI parallel corings

with CoreBook App.

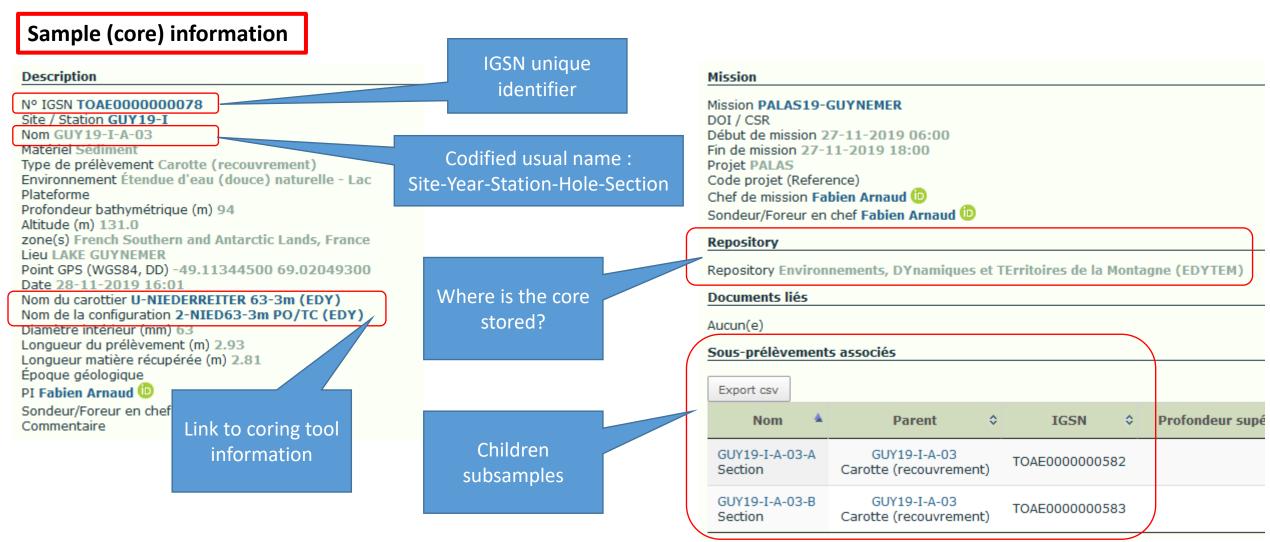
ited at each corebook synchronization. If a synchronization takes place te, this scheme may depend on new changes made by the operator s Corebook mission. Please take care of the scheme version.

Coring Technical Scheme for GUY19-I on LAKE GUYNEMER





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

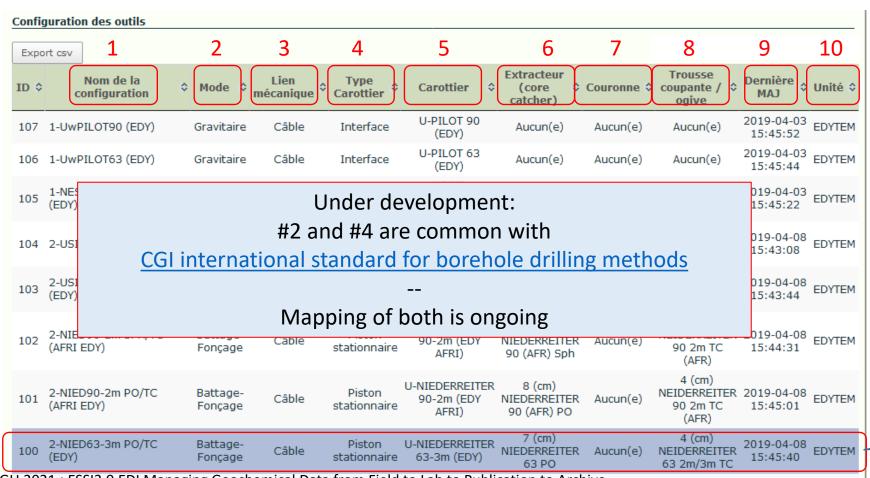






## Banking and displaying fieldwork data: the National Cyber-core repository (cyber-carothèque nationale)

### **Coring tools information**



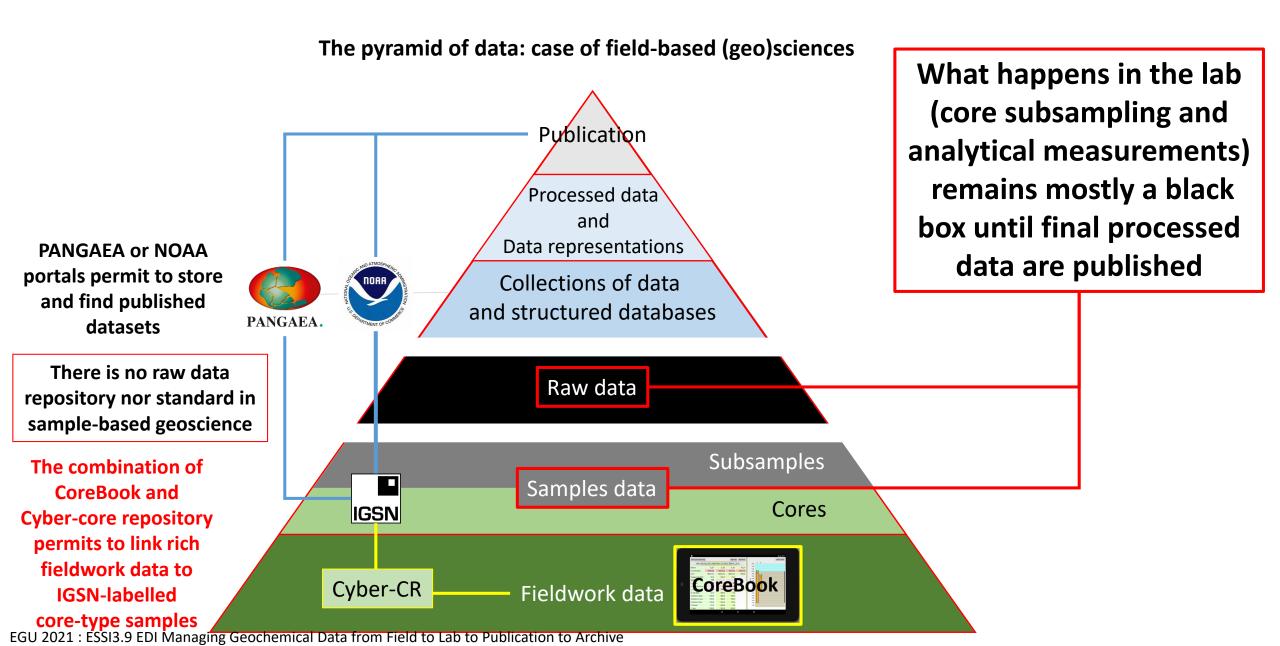
# Normalised description of any coring/drilling tool + its configuration

- 1- Name of the configuration
- 2- Mode (energy for penetration)
- 3- Link to the energy source
- 4- Type of corer
- 5- Usual name of the corer
- 6- Type of core catcher (if any)
- 7- Type coring bit (if any)
- 8- Type of cutting shoe (if any)
- 9- Latest update
- 10- Corer's owner

Configuration used to take GUY19-I-A-03













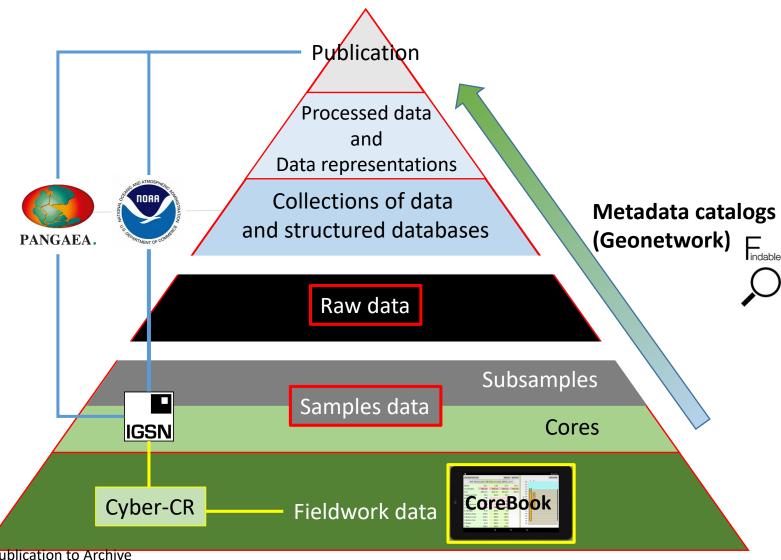
### Remaining challenges are:

- 1) Curating core handling and subsampling data
- 2) Keeping track of previous operations (coring, subsampling, sample treatement) while analytical measurements are performed, together with analysis metadata

### **Ongoing attempts**

- 1) Attempt to prolongate the cyber-core repository using <u>Collec-Science software</u> (not discussed here)
- 2) Prototype of an organised <u>Paleo-</u>database for basic data : **ROZA**

### The pyramid of data: case of field-based (geo)sciences





All Marines, continentals and

ice 's cores

by harvesting WFS

From Cyber Core Repository

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



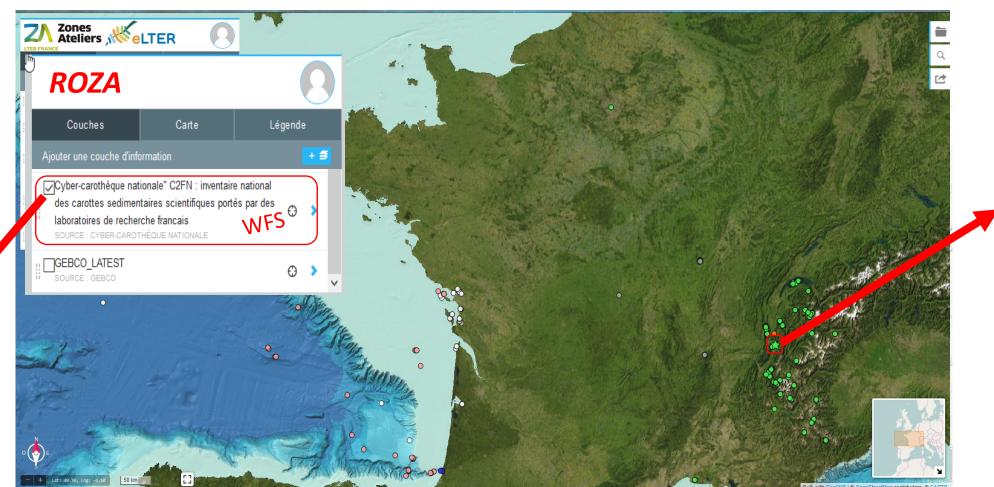
★ = Basic

Data available

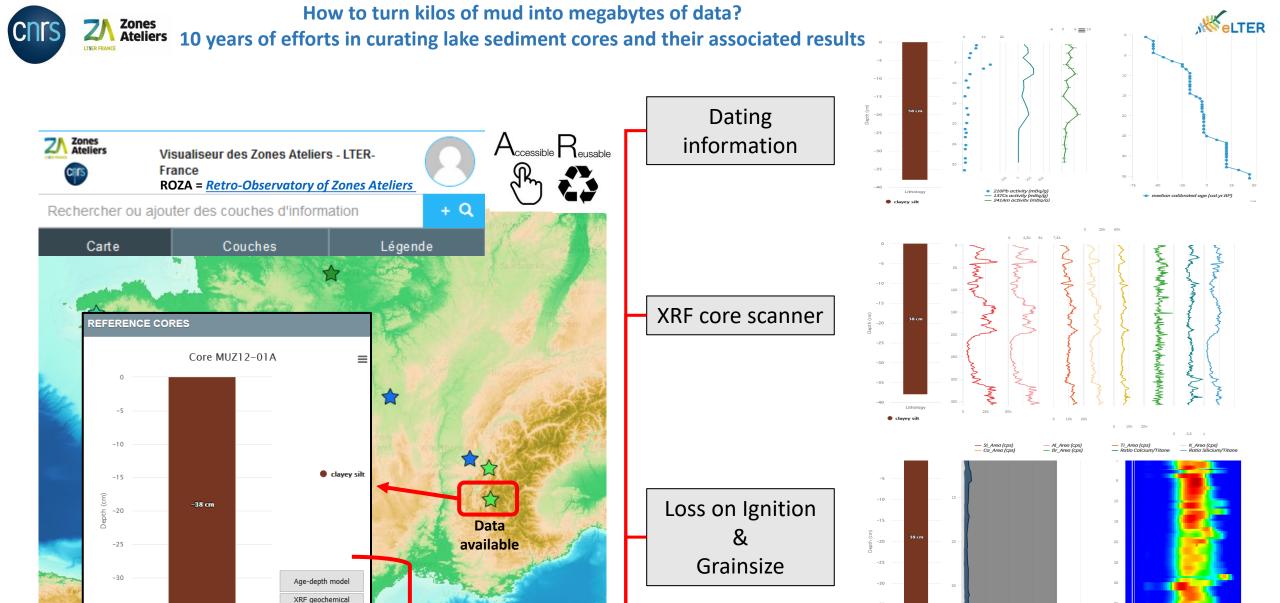
« French Zones Ateliers » are the French contribution to the international Long Term Environmental Research (eLTER) network

ROZA = Retro-Observatory of Zones Ateliers portal





GeoCMS



→ ROZA aims at providing homogenised basic data to promote the re-use of sites, samples and data within the ZA network

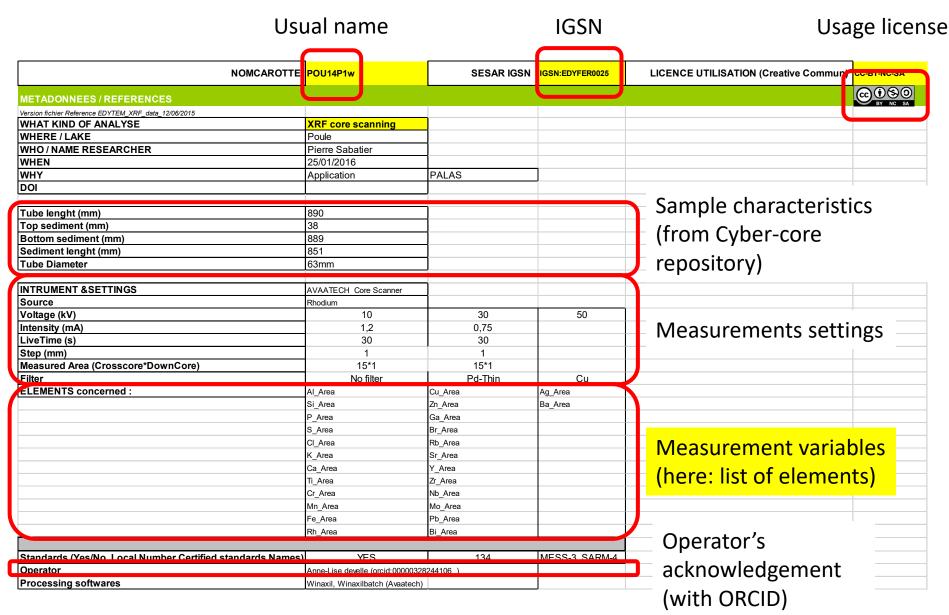
Basic data

Grain size











### **Ongoing effort**

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

→ Referenced in metadata catologs



### Promote visibility & interoperability

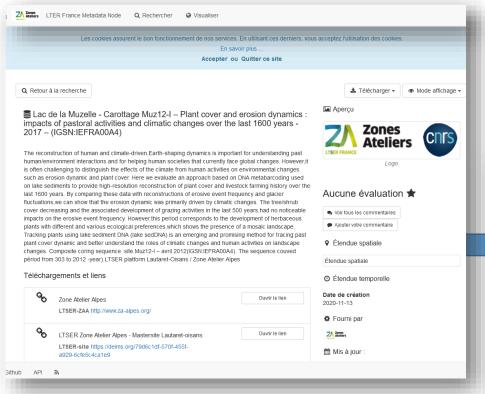
GENERAL

GEMET

**ENVTHES** 

by metadata geocatalogue (ISO19.115)\*





Link with vocabulary

### Thesaurus / Vocabulary (multi-language)

Agroecology Q

- Biodiversity Q
  Farming practices Q
  Catchment Q
  Hydrology Q
- INSPIRE

   Géology Q

   Habitats\_and\_biotopes Q

   Land use Q
  - recherche scientifique Q
     donnée environnementale Q
  - recherche sur l'environnement et développement Q
  - étude de milieu Q
  - données géoréférencées Q
  - banque d'échantillons de l'environnement Q
     échantillonnage Q
  - échantillonnage Q
  - technique d'échantillonnage Q
  - sédimentologie Q
    sédimentation Q
  - glacier Q
  - vase Q
  - argile Q
  - datation Q
  - spectroscopie Q
  - élément chimique Q
  - environmental analysis Q
     experimental research Q
  - biodiversity Qcore Q
  - cores Q
  - sampling date and time Q
  - sampling parameter Q
     sediment Q
  - sediment chemistry Q
  - chemical analysis Q
  - Pinus Q
     Rumex Q
  - Plantago Q
- EGU 2021: ESSI3.9 EDI Managing Geochemical Data from Field to Lab to Publication to Archive

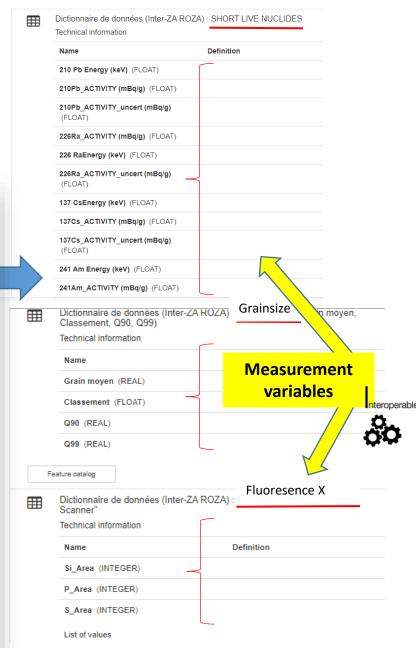
\* Create by R workflow « Geoflow »

flood frequency Q
 age modele Q

### Feature variables catalogue



Feature catalog

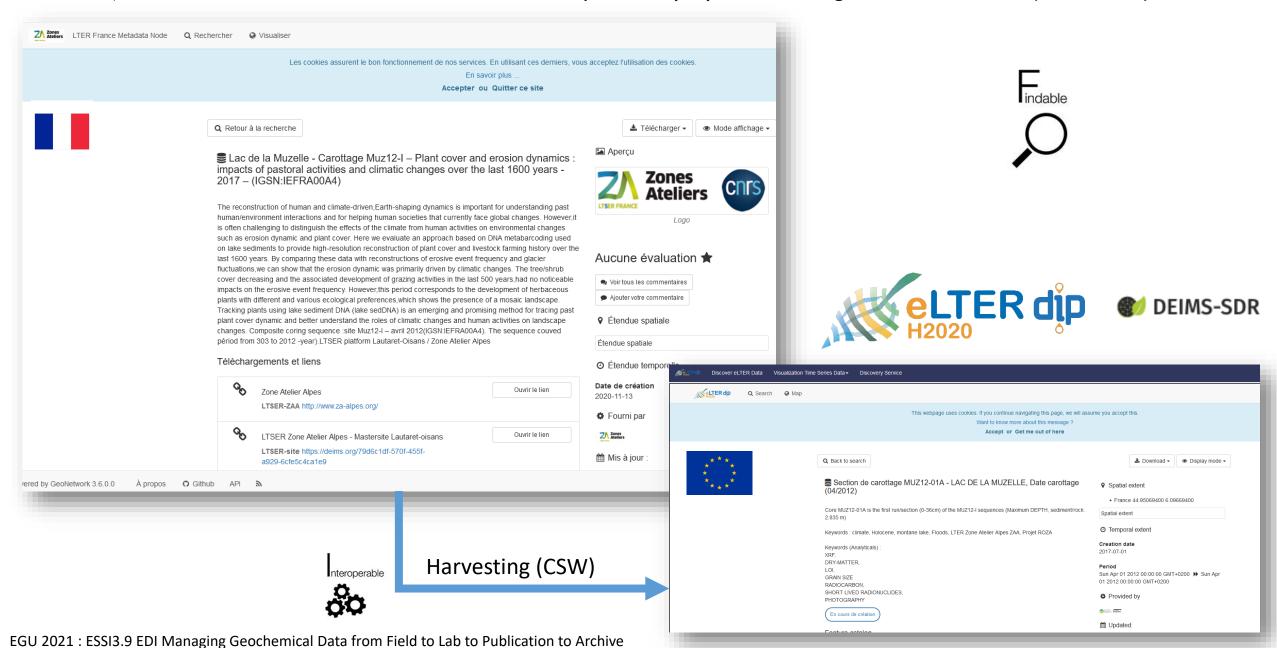




### Discover eLTER Data



### Interoperability by Geocatalogue métadonnées (ISO19115)





### How to turn kilos of mud into megabytes of data?

### 10 years of efforts in curating lake sediment cores and their associated results

### **Achievements**

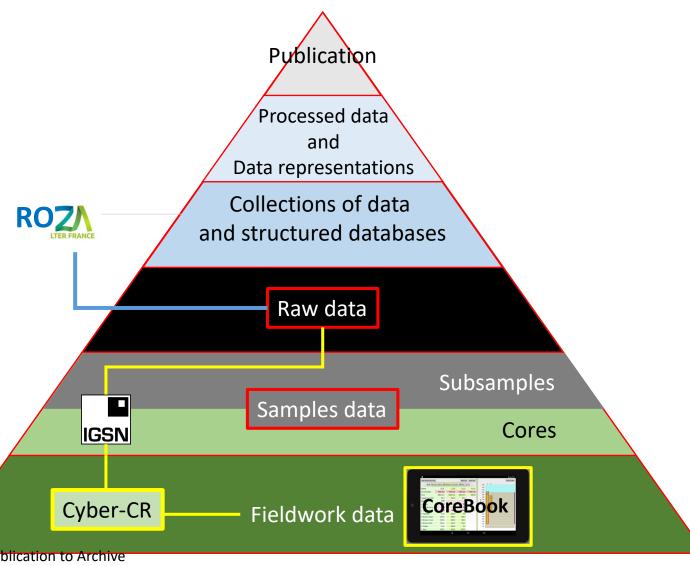
### With ROZA, it is possible to get a dataset that is linked to its:

- measurement metadata (LTER catalogue)
- sample data (IGSN)
- Fieldwork data (Cyber-Core repository)

### Remaining challenges are:

- **ROZA is just a prototype** (with lot of flaws!)
- Propose an user-friendly workflow to inject data
- Deal with sub-samples (currently, only the core level is treated)
- Extend the database (only 4 cores...)
- Include treated data: age-models instead of dating data → SERAC

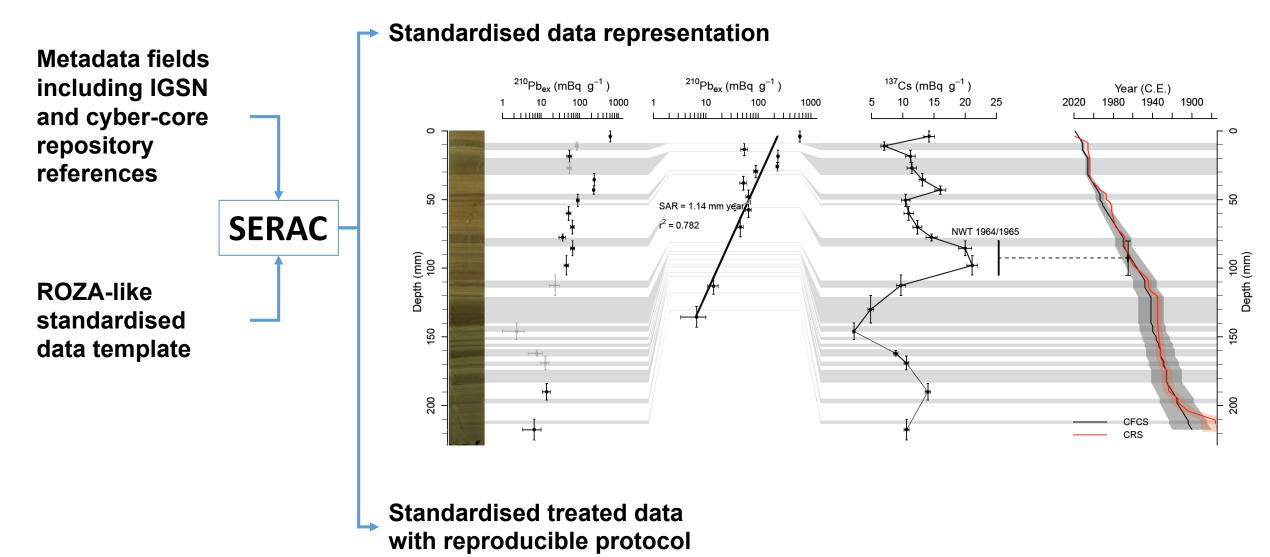
The pyramid of data: case of field-based (geo)sciences





### SERAC: a R package for Short livEd RAdionuclide Chronology of recent sediment cores





Bruel and Sabatier, 2020, Journal of Env. Radioactivity









**Limits** 





On the road to be



### Corebook

Gathers standardised fieldwork data

### **Cyber-core repository**

- Banks fieldwork data
- Links rich fieldwork data with IGSN

#### **IGSN**

Provides unique identifier for upper levels

#### **ROZA**

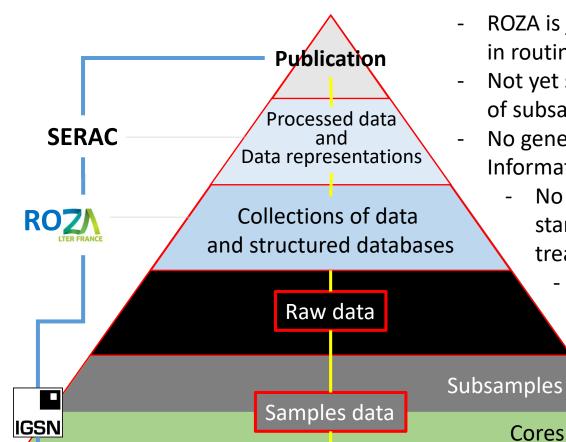
- Standardised dataset linked with:
- measurement metadata (LTER catalogue)
- sample data (IGSN)
- Fieldwork data (Cyber-Core repository)

### **SERAC** provides:

- A publishable standardised representation of radionuclide data
- A standardised age model file with a reproducible protocol
- Links to underlying levels of the pyramid



In the case of radionuclide dating, it is possible to link all levels of the pyramid



Fieldwork data

- ROZA is just a protoype, not used in routine
  - Not yet satisfying management of subsamples
- No generalised use of Lab. Information systems
  - No ongoing projects to standardize other types treated of paleo-data
    - Needs more integration within international projects

Cyber-CR