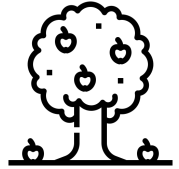


Module 3 : les métadonnées





Cyril Pommier - <https://orcid.org/0000-0002-9040-8733>



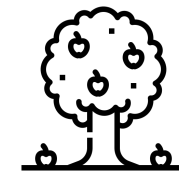
Thomas Denecker - <https://orcid.org/0000-0003-1421-7641>



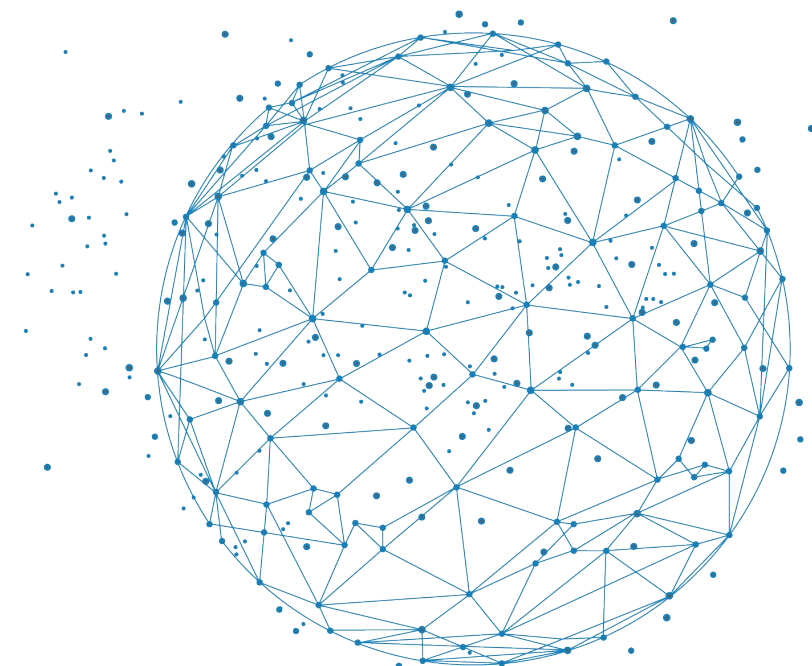
Hélène Chiapello - <https://orcid.org/0000-0001-5102-0632>

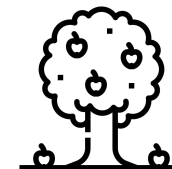


Special thanks to Frédéric de Lamotte - <https://orcid.org/0000-0003-4234-1172>

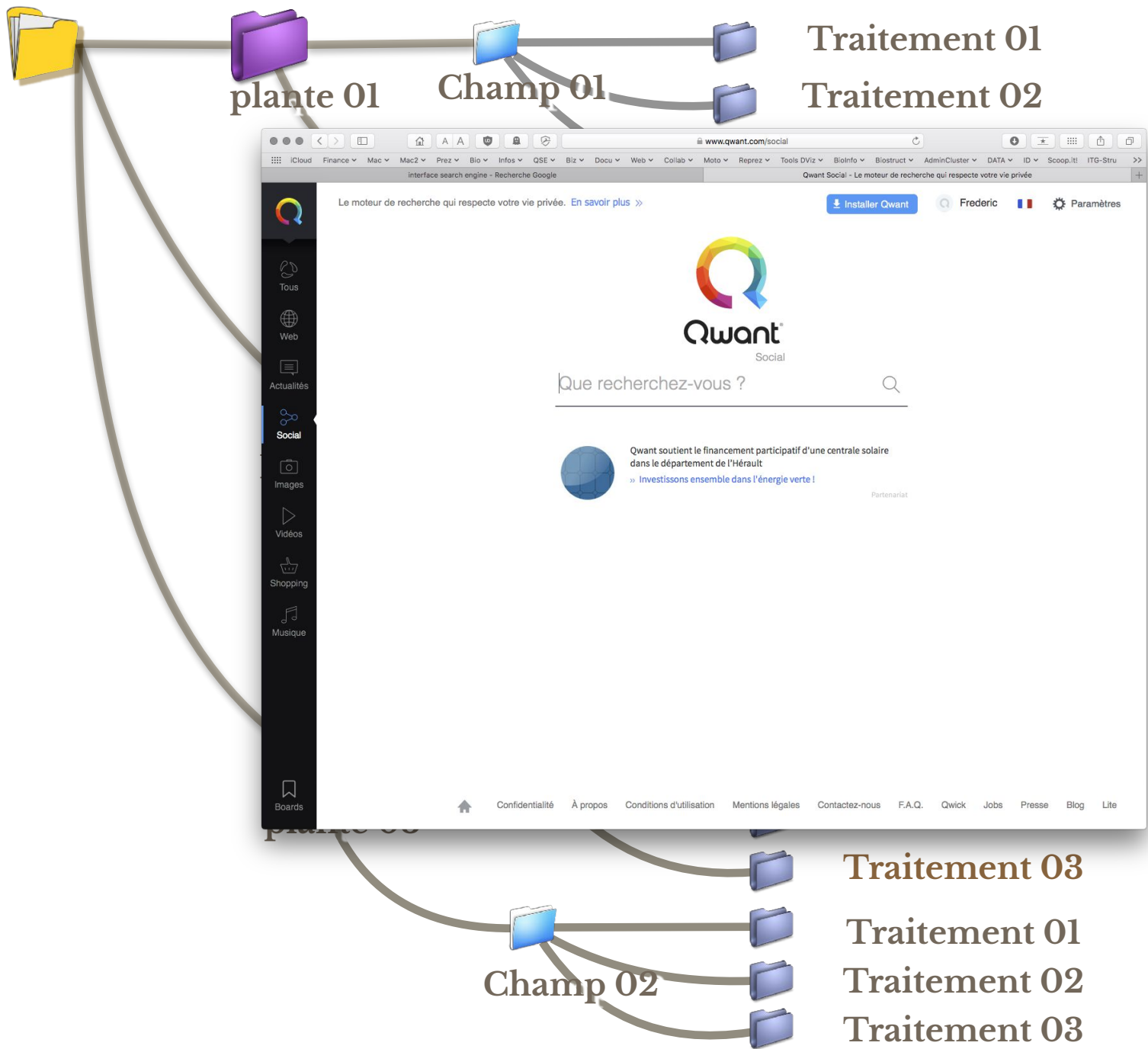
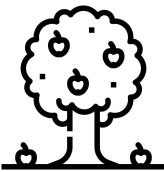


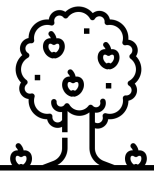
Introduction aux métadonnées





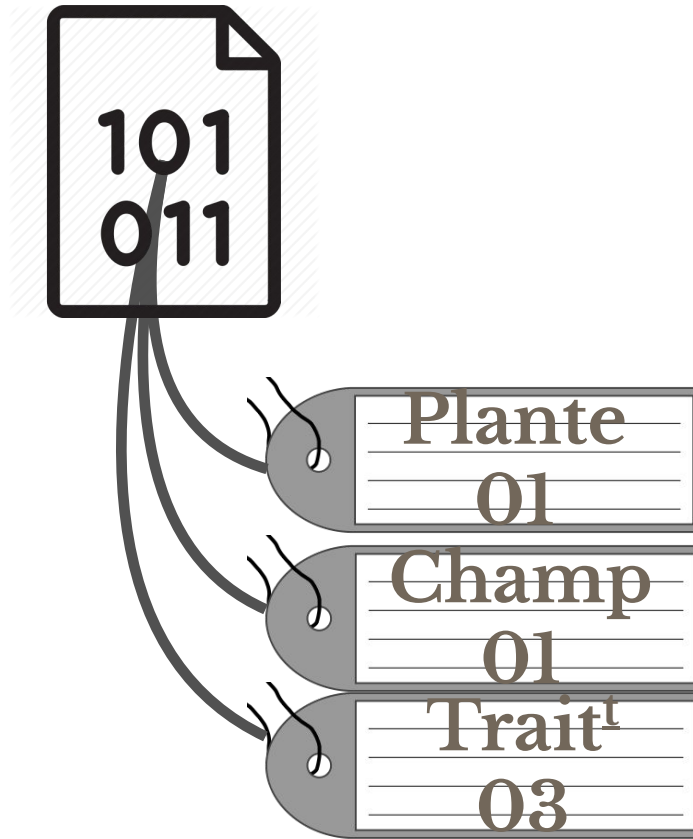
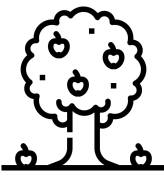
Définir une méthode commune et efficace pour retrouver et comprendre nos données (FAIR)



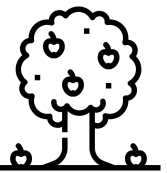


Les métadonnées





Tags
MetaData



atengineeringsolutio... • S'abonner
AT Engineering Solutions

atengineeringsolutions A Table we Designed in Collaboration with @northernbespoke and one of there customers. A bespoke table made to the exact specifications the customer requested before hand. #engineering #welding #scarborough #uk #metal #metalwork #furniture #table #workshop #weld #weldporn #welder #weldernation #house #home #instagram



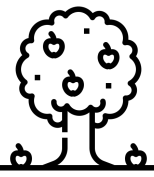
donnersteel, james.mcgregor.52, stephaxil, noah4444, didsy_, _edwardjones, markbulmer_photography, rtedgar_boroondara, thundrgram et beastmotivationfitness aiment ça.

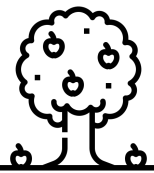
IL Y A 20 MINUTES

Ajouter un commentaire... ...



Un vocabulaire contrôlé



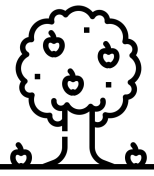


Genre	Espece	Sous espece	Groupe	Nom
Oryza	Sativa		japonica	PENTHE BLANC
Oryza	Sativa		japonica	PENTHE NOIR
Oryza	Sativa		indica	ZOGO
Oryza	Glaberrima			GBAI-GBAI
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Musa	acuminata	banksii	wild	Banksii H09

Germplasma	Origine	Collection
AG0003	Guinea	prospection 1979
AG0004	Guinea	prospection 1979

Vocabulaire contrôlé
Défini par la
communauté
Evolutif

Type de séquençage	Taille insert	Longueur de read	type de machine	Lieu du séquençage
illumina		1*150	HiSeq3000	Genotoul
illumina		1*150	HiSeq3000	Genotoul



What is Metadata?

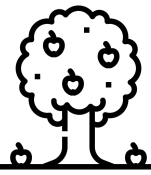
Metadata is: Data 'reporting'

- **WHO** created the data?
- **WHAT** is the content of the data?
- **WHEN** were the data created?
- **WHERE** is it geographically?
- **HOW** were the data developed?
- **WHY** were the data developed?



Photo by Michelle Chang. All Rights Reserved

Un standard de métadonnées



Le regard métier

Un jeu de métadonnées ne sera pas formalisé de la même manière selon les standards employés. Un même **objet** peut ne pas être **décrit** de la manière selon la perspective « **métier** » portée sur lui. **Le regard « métier » structure la donnée**. A titre d'exemple le traitement documentaire appliqué à une collection de cartes postales ne sera pas le même selon que celui-ci est opéré par un musée ou un service d'archives. Les **archivistes** s'attacheront à retrouver les **toponymes** là où les **musées** relèveront plutôt des détails ayant trait à l'**histoire de l'art**. (mais aussi les divergences entre les 2 communautés « climat »)

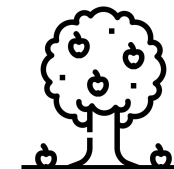
Ouverture et interopérabilité

Le traitement documentaire doit s'inscrire dans des logiques d'ouverture et d'interopérabilité. La **qualité des données et métadonnées conditionne les réutilisations possibles**, il en est de même pour le degré d'ouverture des ressources et de leurs métadonnées.

Le choix des métadonnées qui seront produites dans le cadre d'un projet de numérisation peut répondre à des **usages clairement identifiés** en amont du projet. Le fait de s'appuyer sur des **standards** favorise l'interopérabilité et peut permettre des **usages autres que ceux attendus**.

Approches participatives

Une approche participative peut venir compléter le traitement documentaire. Cette approche participative peut prendre diverses formes : collecte, enrichissement de métadonnées, annotations, transcriptions collaboratives... Il est préférable d'envisager cette approche collaborative en amont ou en parallèle du projet. Le porteur de projet doit être conscient des risques induits pour la qualité et la fiabilité des données et la nécessité de **gérer et animer les communautés** d'utilisateurs selon le type d'approche choisi.



In this section

[Briefing Papers](#)
[How-to Guides & Checklists](#)
[Developing RDM Services](#)
[Curation Lifecycle Model](#)
[Curation Reference Manual](#)
[Policy and legal](#)
[Data Management Plans](#)
[Tools](#)
[Case studies](#)
[Repository audit and assessment](#)

Standards

[Disciplinary Metadata](#)
[DIFFUSE](#)
[Publications and presentations](#)
[Roles](#)
[Curation journals](#)
[Informatics research](#)
[External resources](#)
[Online Store](#)

Digital curation standards

For digital curation and data preservation initiatives to be successful, activities must be based upon sound and tested standards that promote best practice.

The DCC is committed to providing a standards watch that will play a vital role in the testing and certification of new tools and trusted digital repositories.

Disciplinary Metadata

The issue of disciplinary metadata standards - what they are, who's using them, how to use them - has been gaining attention in the RDM community. To support this, we have created a [Disciplinary Metadata page](#) for those who need help figuring out what standards might address their own needs.

Rather than archival metadata standards, the resource focuses on descriptive standards that aid data discovery and re-use; this is the information a repository manager might give a researcher curious about what his or her discipline has decided should be the minimum information kept alongside their data sets. The initial focus has been on metadata standards for tab-delimited data.

If we're missing your favourite standard, please let us know! We're also particularly interested in hearing about your own experiences in implementing

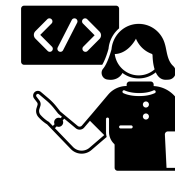
Curation Reference Manual



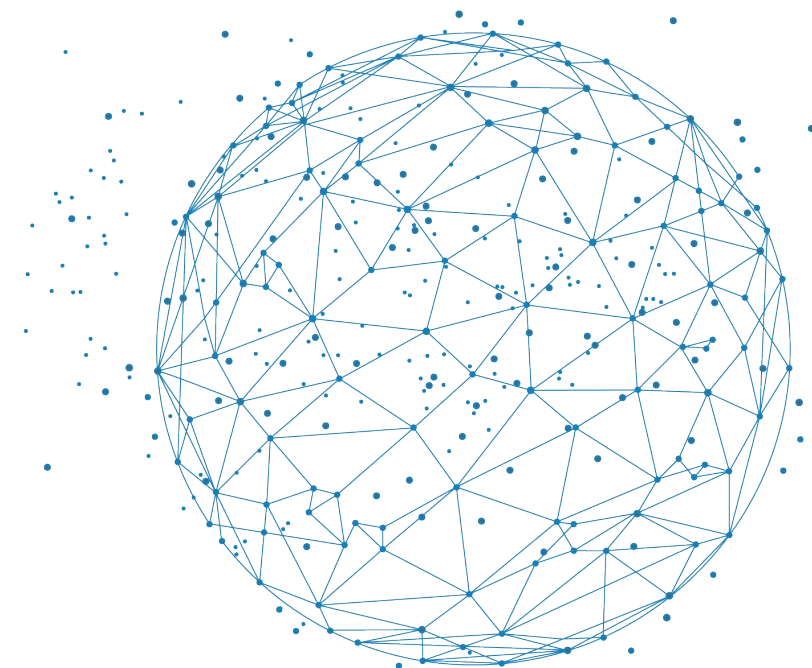
Advice, in-depth information and criticism on current techniques and best practice.

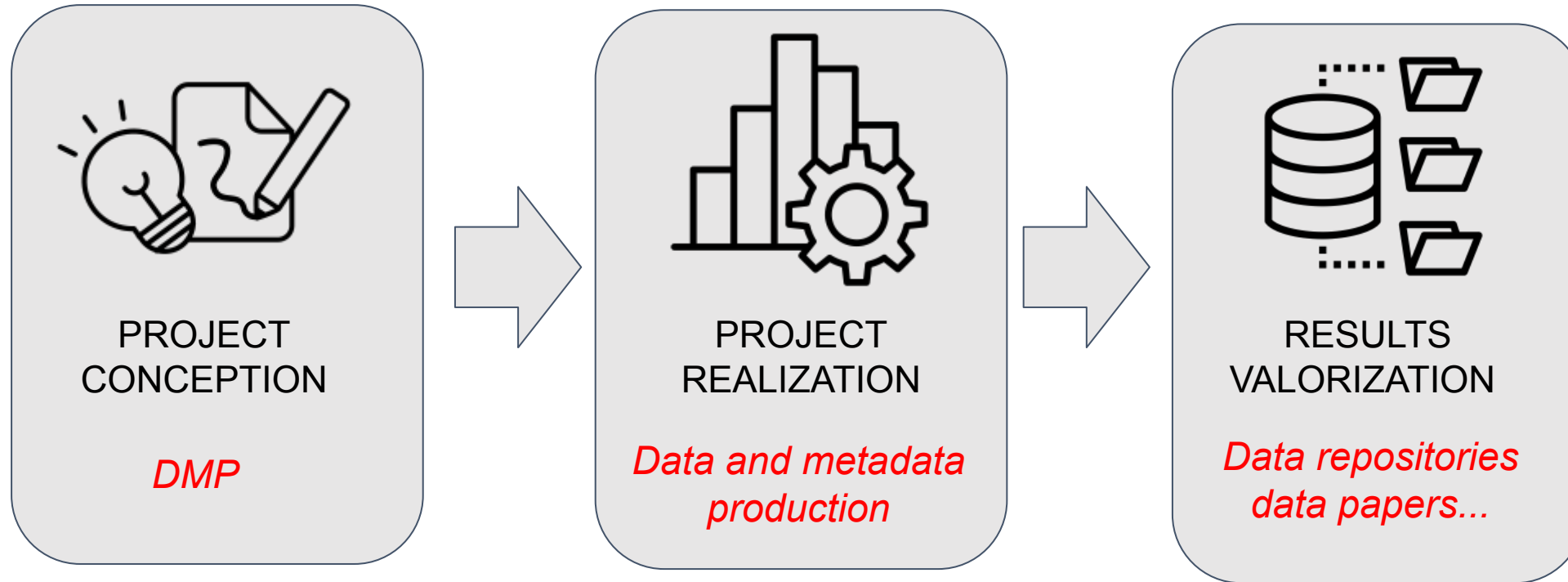
Contributions are made by our extended network of specialist partners and associates.

[Read more](#)
[Home](#)



Standards de métadonnées en Sciences de la Vie





Metadata concern all steps of a scientific project !



How do you describe the data?



With a set of metadata



How do you ensure you don't forget certain metadata?



With a **metadata standard**



Disciplinary standard



General standard



	A	B
1	Titre	
2	Auteur	
3	Date	
4	Résumé	
5	Mots-clés	
6	Identifiant	
7	Format	
8	Contexte de création	

Source: <https://www.pasteur.fr/fr/file/20615/download>



Question: Do you know any standard in life sciences ?

5 minutes to find an example of metadata standard and write a note in

https://scrumblr.ethibox.fr/metadata_standard



In essence, a standard is an agreed way of doing something. A standard provides the requirements, specifications, guidelines or characteristics that can be used for the description, interoperability, citation, sharing, publication, or preservation of all kinds of digital objects such as data, code, algorithms, workflows, software, or papers.

source: <https://fairsharing.org/educational>



Why do I have to use a **data standard**?

- To analyse, compare and exchange data
- To publish datasets in international resources

And a **metadata standard**?

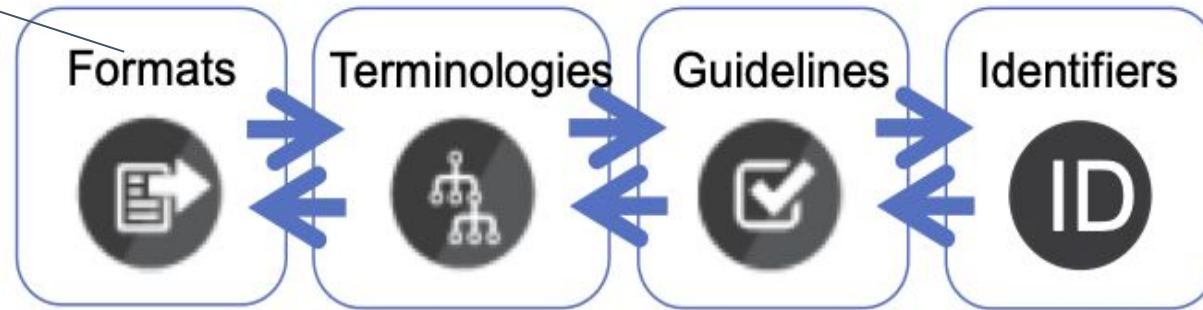
- To describe data richly and accurately, with the same vocabulary as the rest of your scientific community
- To make your metadata interoperable and to allow other systems to exploit them

The Gene Ontology is a **metadata** standard



Organization, Conceptual model, schema, exchange formats, etc...
e.g. SBML, FASTA, VCF, CSV
Biologist & Computer scientist driven

Good practices, how to, Minimum information reporting requirements, checklists...
e.g. MIAME guidelines
Biologist training



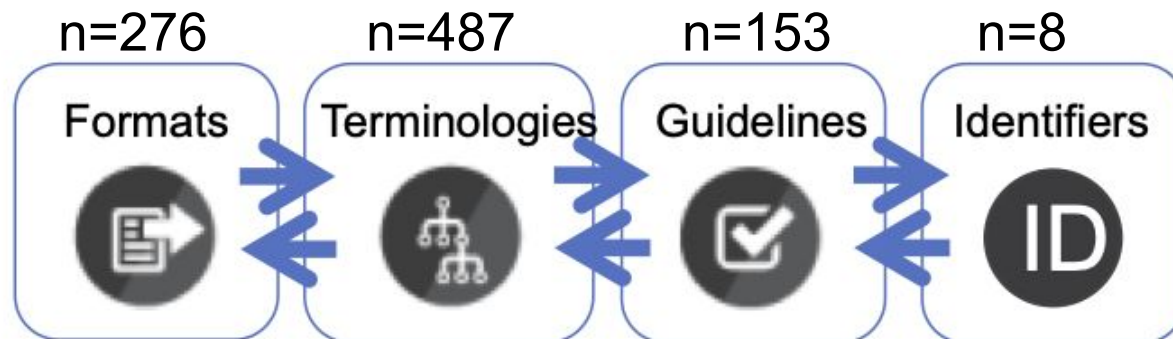
Controlled vocabularies, taxonomies, ontologies...
e.g. Gene Ontology, Community ontologies
Biologist driven

Unambiguous identifiers WWW level
For: datasets, genes, people, resources
e.g. DOI, URI, identifiers.org, ...



GENERIC STANDARDS

across life science : genomic, ...



COMMUNITY STANDARDS

for metadata and identifiers



FASTA FASTQ
GFF SBML
Newick MIAME
BAM EC number
MINSEQE VCF

Source: <https://fairsharing.org/standards/?q=life+sciences>



Two kinds of standard descriptors

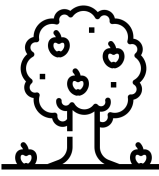
- Generic descriptors:
 - [Dublin core](#) for description of numerical resources
 - bioschema.org for description of life science resources (datasets, softwares, training material,...)
- Specific dataset descriptors:
 - [MIAME](#) (Minimum Information About a Microarray Experiment)
 - [MIAPPE](#) (Minimum Information About a Plant Phenotyping Experiment)
 - ...

Metadata standards often depend on the repository you will use to publish data

> It is helpful to decide at the beginning of the project what are the recommended repositories for your data types

> You can view ELIXIR repositories here:

<https://elixir-europe.org/platforms/data/elixir-deposition-databases>



Three text formats frequently used for metadata



JavaScript Object Notation

Comma Separated Values

```
Sample_alias, date, source
A, 20200802, blood
B, 20200802, feces
C, 20200802, skin
```

eXtensible Markup Language

```
<SAMPLE_SET>
  <SAMPLE alias="A">
    <date>20200802</date>
    <source>blood</source>
  </SAMPLE>
  <SAMPLE alias="B">
    <date>20200802</date>
    <source>feces</source>
  </SAMPLE>
  <SAMPLE alias="C">
    <date>20200802</date>
    <source>skin</source>
  </SAMPLE>
</SAMPLE_SET>
```

```
{
  "SAMPLE_SET": {
    "SAMPLE": [
      {
        "alias": "A",
        "date": "20200802",
        "source": "blood"
      },
      {
        "alias": "B",
        "date": "20200802",
        "source": "feces"
      },
      {
        "alias": "C",
        "date": "20200802",
        "source": "skin"
      }
    ]
  }
}
```



- 100 fastq files
- 1 VCF File
- 5 Metadata files
- 15 Phenotyping datafile
- Plus some Omics: Metabolomic, expression, etc...

⇒ How to organize all of that ?

The ISA model

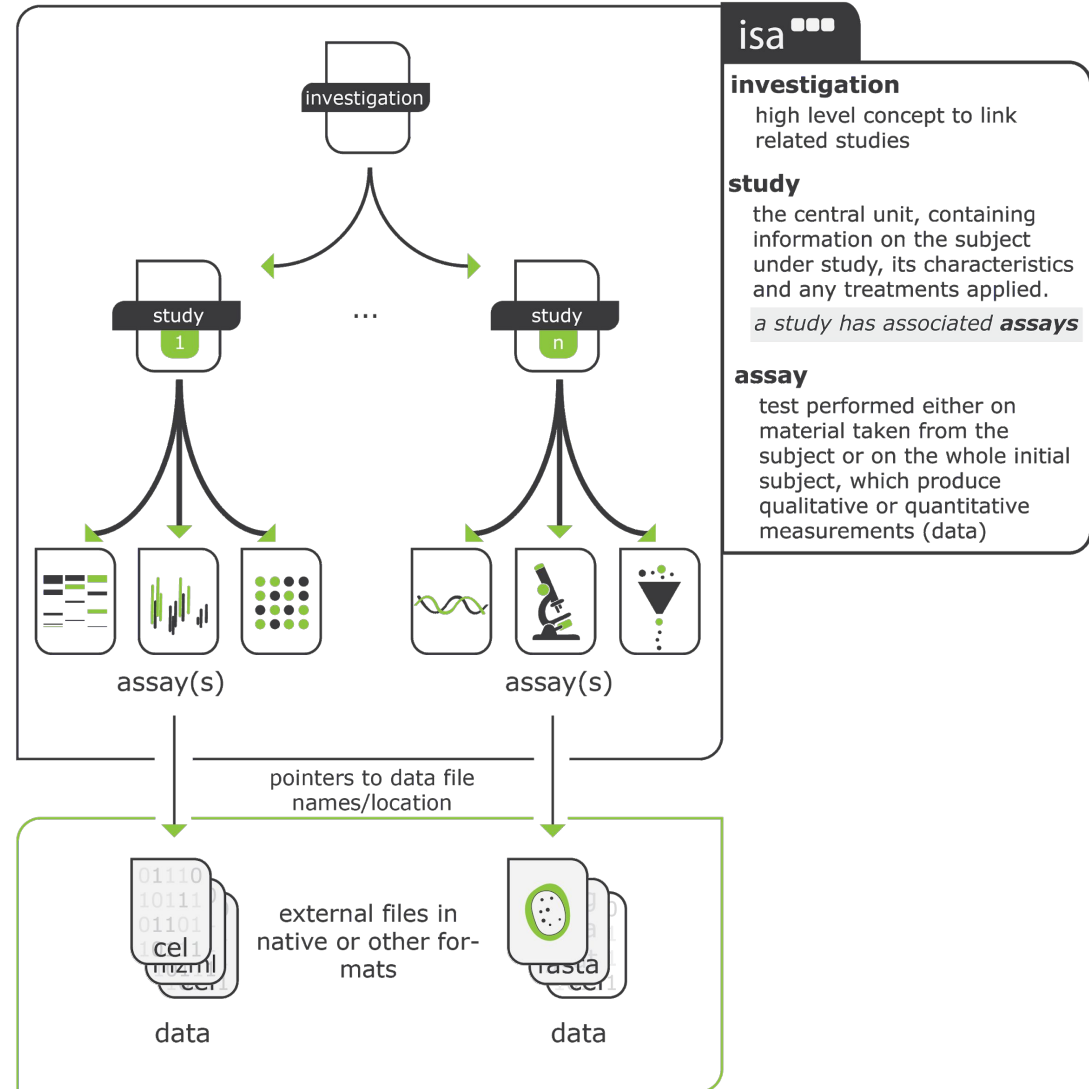
A standard for Life ScienceData

A model to capture **experimental metadata** through **3 core entities**:

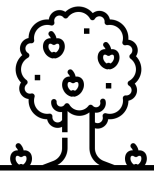
- **Investigation**: the project context
- **Study**: an experimentation in one location
- **Assay**: a specific measurement that targets a trait with a method and a scale

ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Rocca-Serra P et al. **Bioinformatics** 2010.

<https://doi.org/10.1093/bioinformatics/btq415>



Sources: <https://isa-tools.org> and :
<https://isa-specs.readthedocs.io/en/latest/isamodel.html>

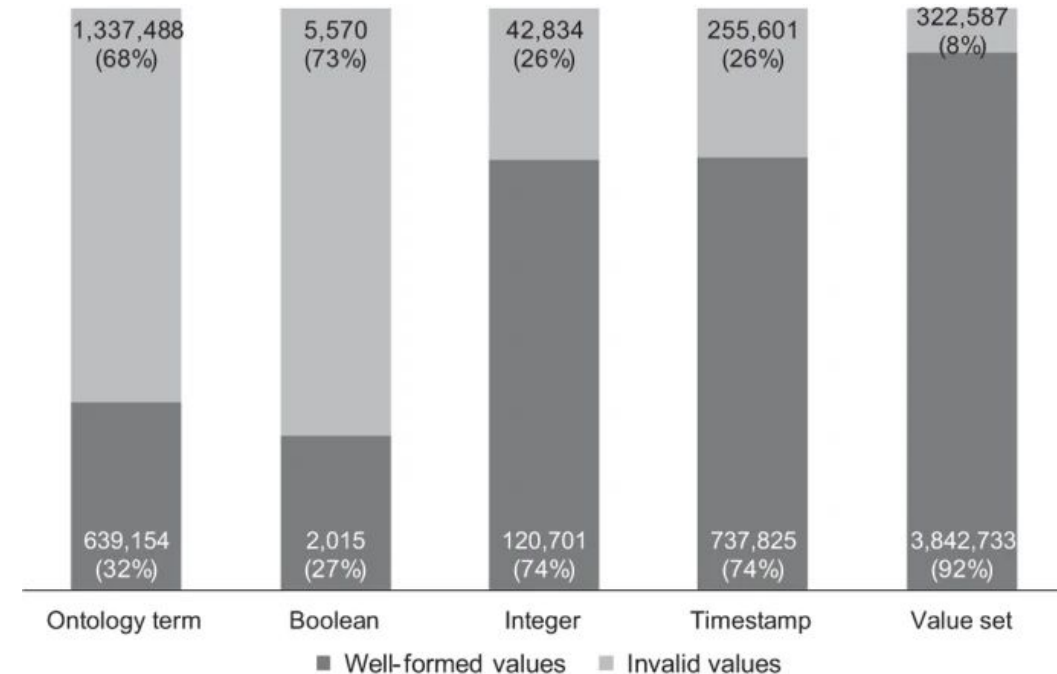




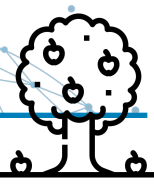
Submission in public resources is often a complex task

Submission procedures are heterogeneous

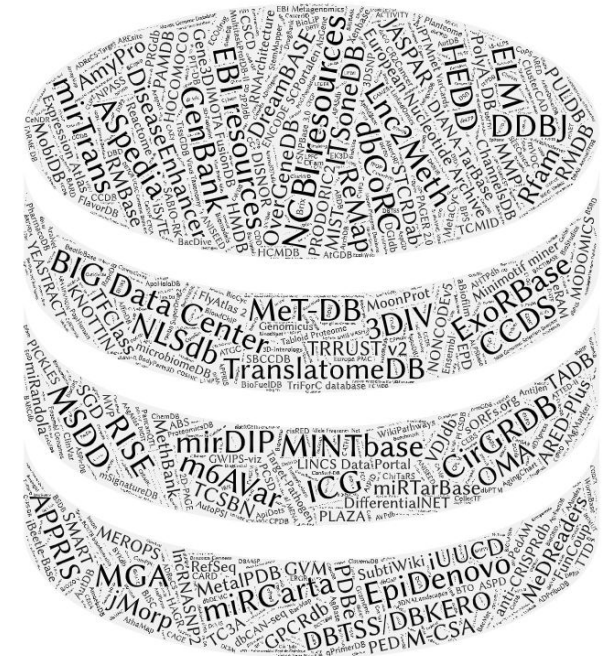
Metadata are often incomplete, inconsistent, redundant or not informative enough



Quality of dictionary attributes in NCBI BioSample according to their type, in [Gonçalves et al., 2019](#)

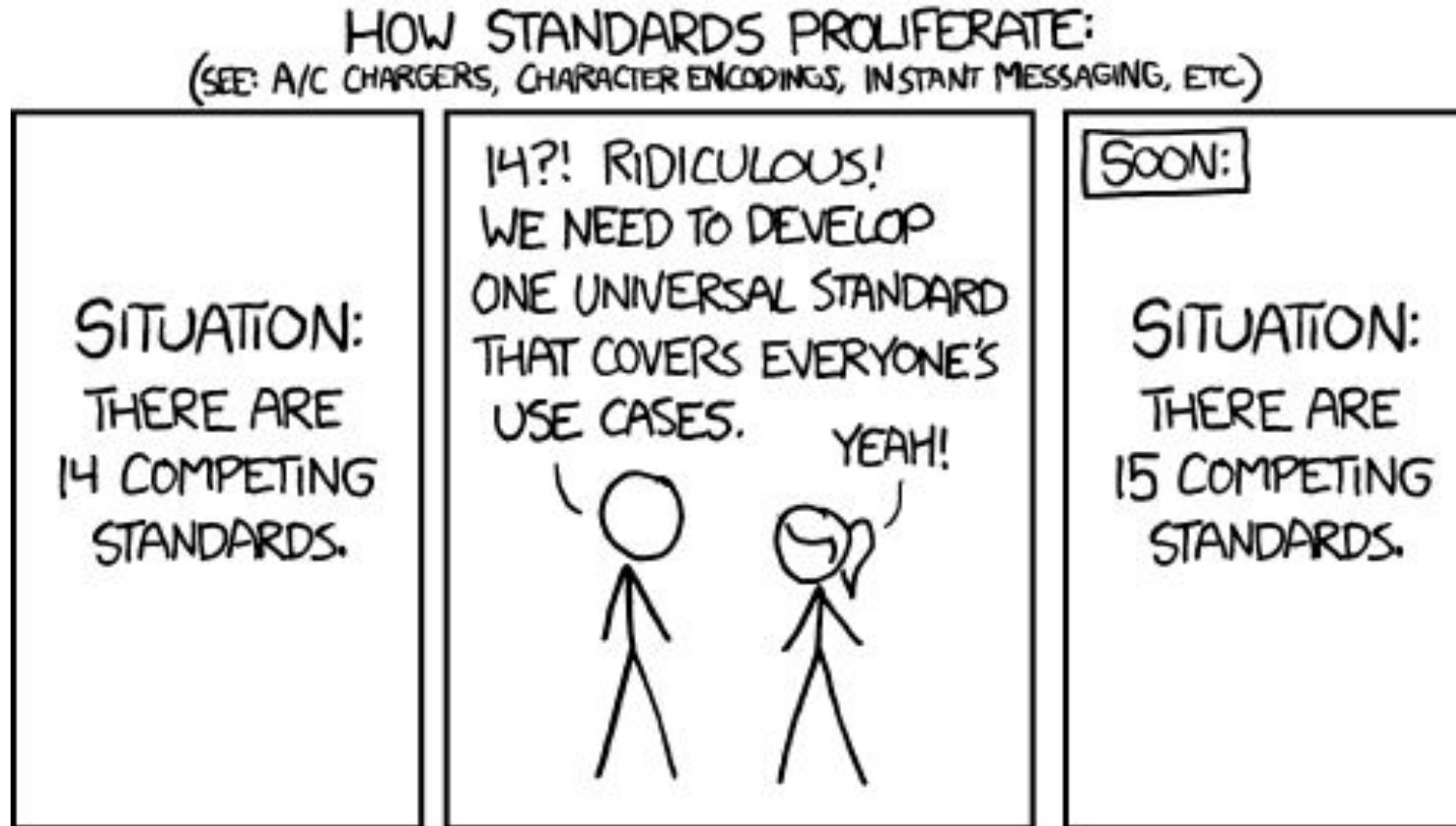


- There are thousand of databases, softwares and resources in biology with an **unequal level of standard adoption**
- It is not always easy for life scientists and bioinformaticians to identify and use the most appropriate standards

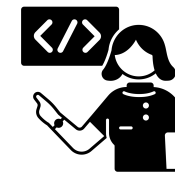


1641 databases in NAR Database 2021

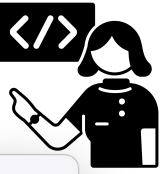
[Rigden et al, 2021](#)



Source: <https://xkcd.com/927/>



How do I find the standard I need?



The FAIRsharing portal

A resource providing **curated descriptions** of standards, databases and policies



Sansone et al. Nat Biotech. 2019

<https://doi.org/10.1038/s41587-019-0080-8>

fairsharing.org

search through all content

STANDARDS DATABASES POLICIES COLLECTIONS ADD CONTENT STATS LOGIN

A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies.
We guide consumers to discover, select and use these resources with confidence, and producers to make their resource more discoverable, more widely adopted and cited.

RESEARCHERS DEVELOPERS & CURATORS JOURNAL PUBLISHERS LIBRARIANS & TRAINERS SOCIETIES & ALLIANCES FUNDERS

Researchers in academia, industry and government

Discover and cite the standards, databases or repositories that exist for your discipline when creating a data management plan, releasing data or submitting a manuscript to a journal...

Dev
Incr
acce
read

1583 Standards	
Terminology Artifact	830
Model/Format	504
Reporting Guideline	228
Identifier Schema	21
VIEW ALL	

1861 Databases	
Repositories	956
Knowledgebases	788
Knowledgebase/Repositories	117
VIEW ALL	

150 Policies	
Journal	89
Funder	23
Society	14
Project	13
VIEW ALL	

<https://fairsharing.org>



The FAIRsharing portal

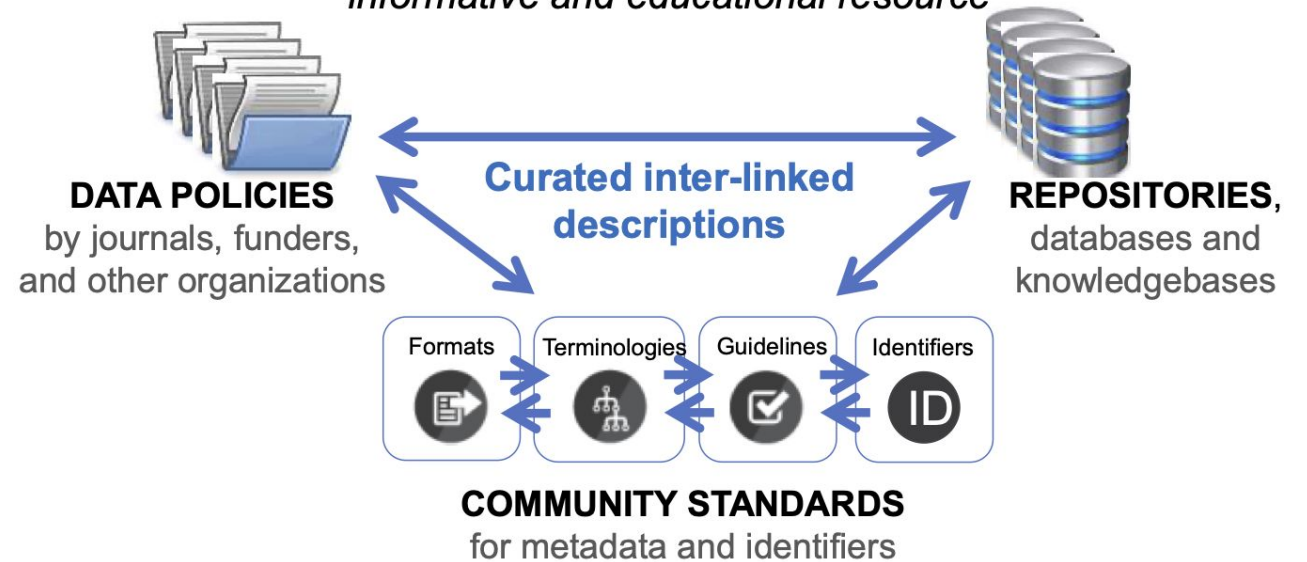
Citable *DOI* for all records

Accessible via *API* or *web interface*

Curation

FAIRsharing.org

informative and educational resource



RECORD STATUS

- R Ready for use, implementation, or recommendation
- Dev In development
- U Status uncertain
- D Deprecated as subsumed or superseded

All records are manually **curated in-house**, verified and claimed by the community behind each resource

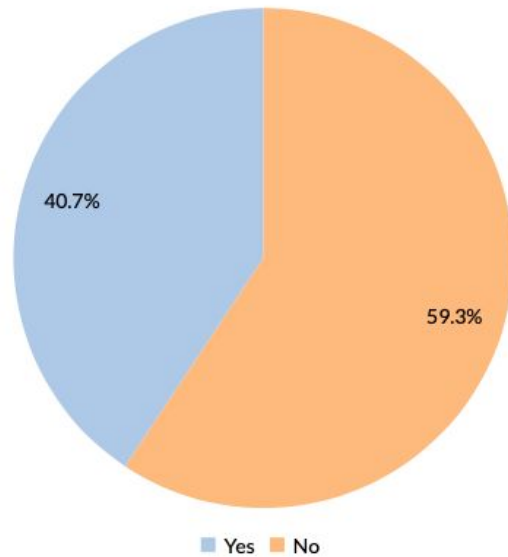


- R** Ready for use, implementation, or recommendation
- Dev** In development
- U** Status uncertain
- D** ~~Deprecated as subsumed or superseded~~

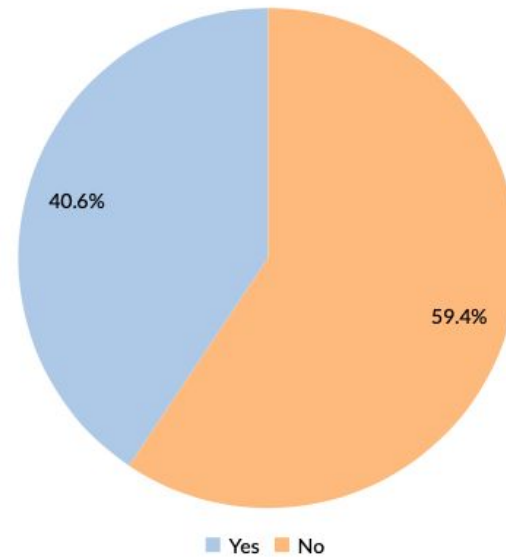
Please don't use "Uncertain" or "Deprecated" standards



Standard records that have maintainers



Standards that have a publication



59.3 % of standards have no maintainer

59.4% of standard has no publication

<https://fairsharing.org/summary-statistics/?collection=standards>

Collections in the FAIRsharing portal



A collection includes standards and/or databases grouped by domain, species or organization

Graph view to visualize relationship links between resources

The screenshot shows the FAIRsharing.org interface for a collection titled "COVID-19 Resources". The page features a navigation bar with tabs for Standards, Databases, Policies, Collections, Add/Claim Content, Stats, and Log In or Register. Below the navigation, there are subject filters (Biomedical Science, Clinical Studies, Epidemiology, Global Health, Health Science, Preclinical Studies, Public Health, Virology) and user-defined tags (Respiratory Disease). A "View as Graph" button is visible, along with a "Compare with collection/recommendation (Beta)" dropdown menu. The main content area displays "General collection/recommendation statistics" for "COVID-19 Resources (bsg-c000070)". A table of records is shown, with columns for Registry Name, Abbreviation, Type, Subject, Domain, Taxonomy, Related Database, Related Standard, and Related Policy. The table lists various registries such as "American Type Culture Collection Database" and "Australian New Zealand Clinical Trials Registry".

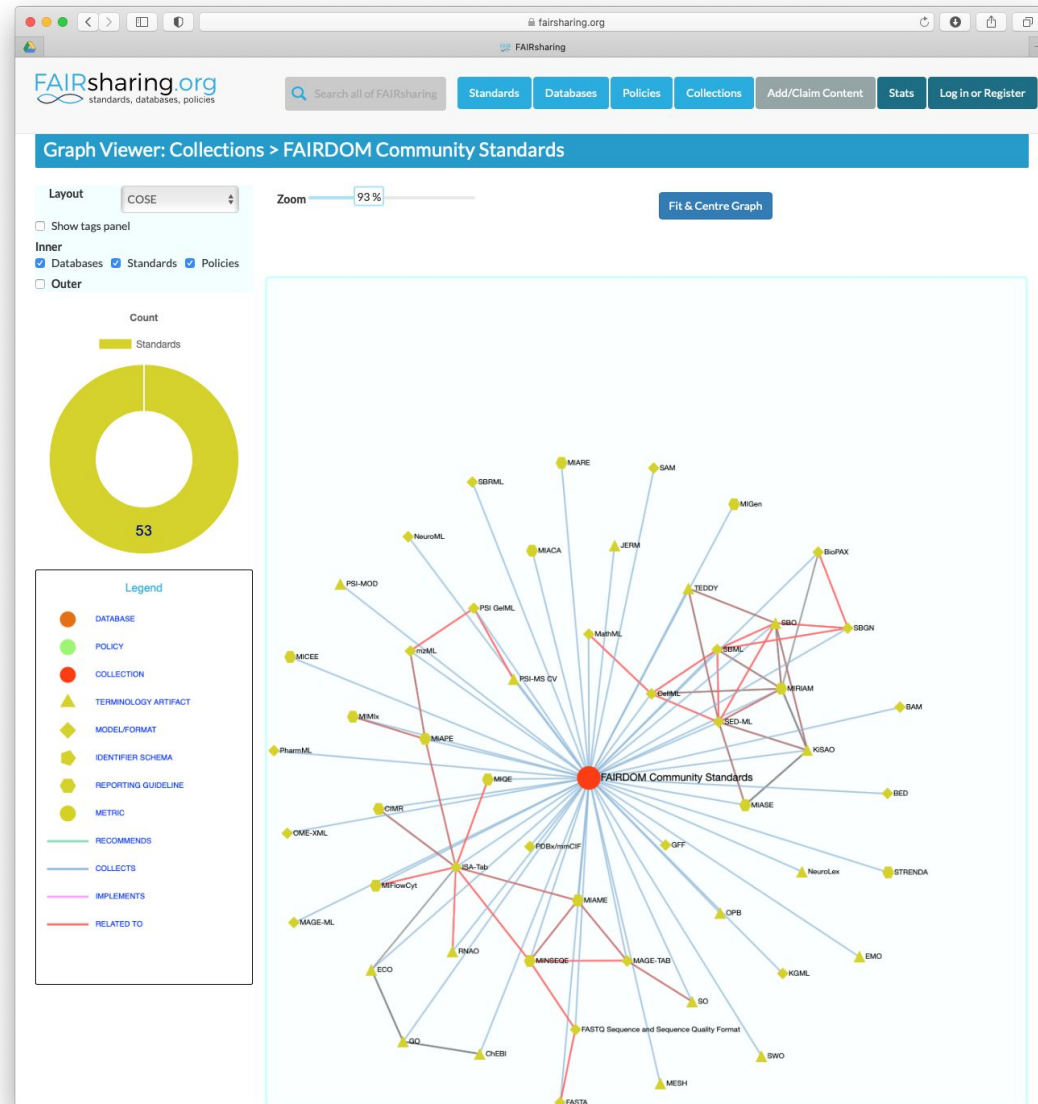
<https://fairsharing.org/collections/>



53 collections related to Life Science standards in FAIRsharing

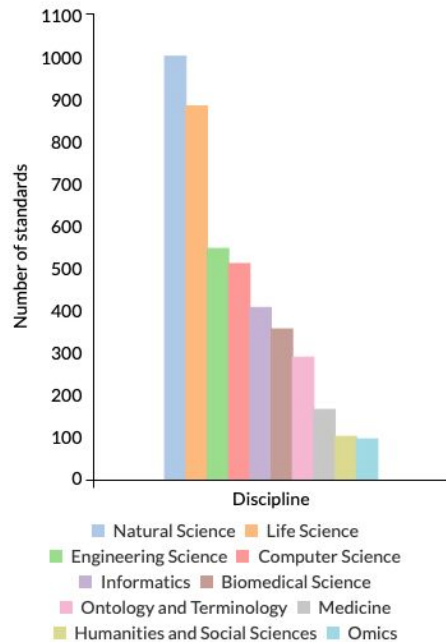
Example 1: the *FAIRdom community Standards collection* (System biology)

<https://fairsharing.org/collection/FAIRDOME>

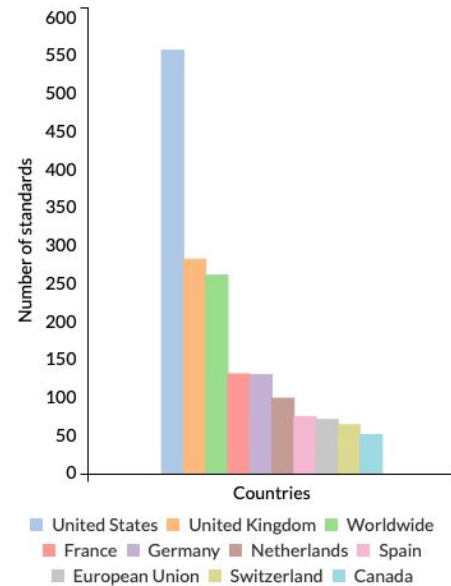




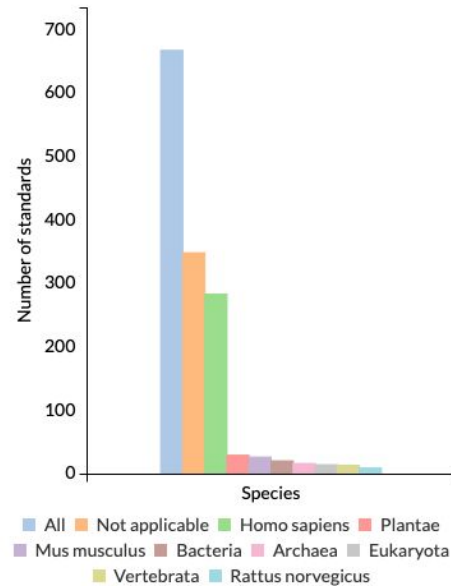
Top 10 disciplines covered by standards



Top 10 standard producing countries



Top 10 species covered by standards

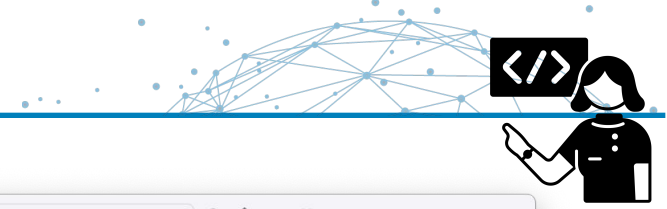


Life Science is one of the best covered discipline

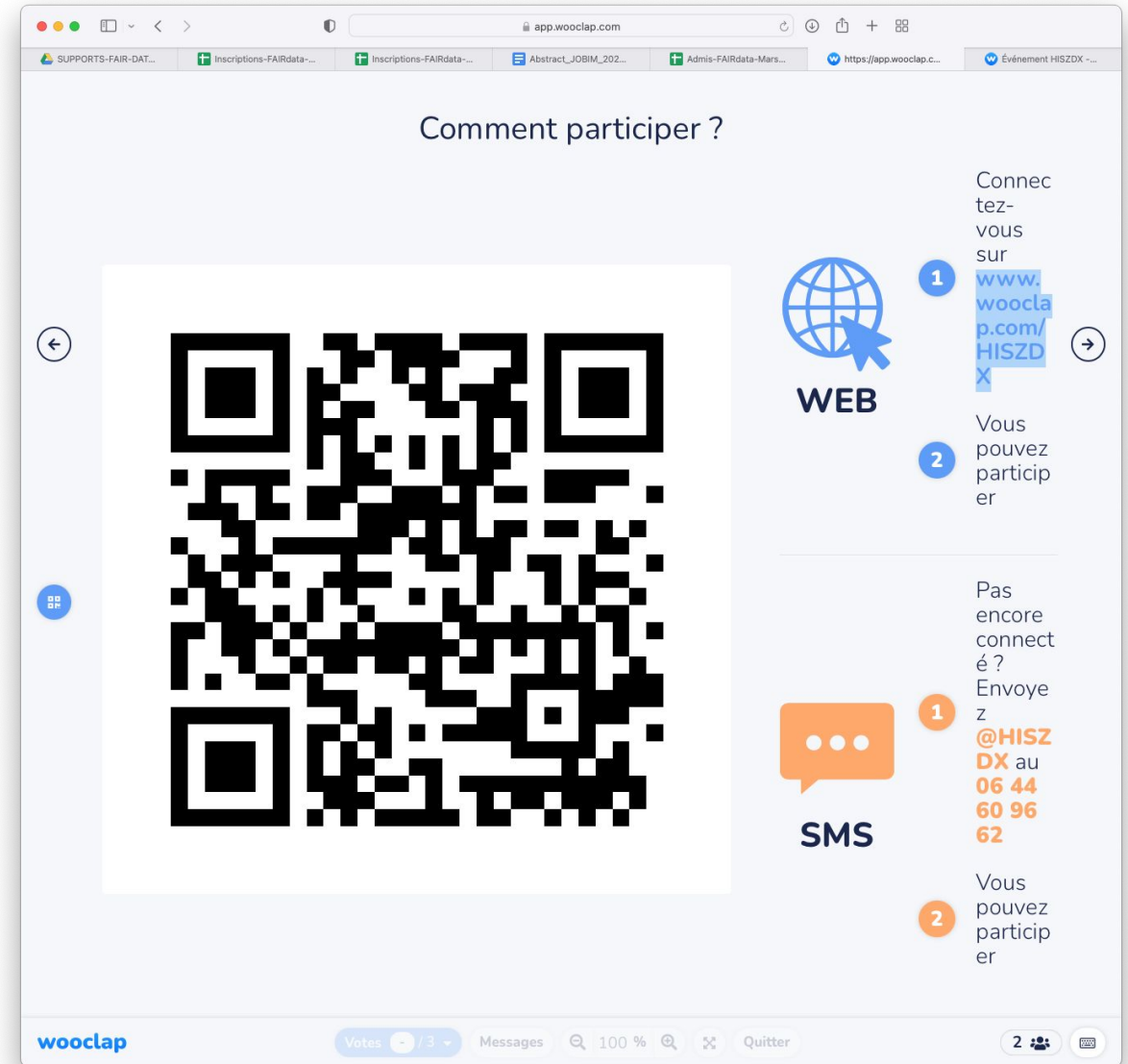
US and UK are the main standards producers

Human species is the best covered species

<https://fairsharing.org/summary-statistics/?collection=standards>



Connect on www.wooclap.com/HISZDX





Find the *Genomic Standards Consortium (GSC)* used by both *ENA* and *SRA* databases in the **FAIRsharing** collections

Use both the record summary and the Graph visualization to interpret and answer the questions in zoom:

1. How many records (*i.e.* standards) are associated to the GSC ? => **6**
2. What type of standard is *Minimum Information about any (x) Sequence (MiXS)* ? => **Reporting guideline**
3. What is the record status of the GAZ record ? => **Uncertain**

Source: <https://gensc.org>

The Genomic Standards Consortium (GSC)



The screenshot shows the FAIRsharing.org interface for the collection 'bsg-c000040'. The header includes the FAIRsharing.org logo and a navigation menu. Below the header, the collection ID 'bsg-c000040' is displayed. The main content area features the GSC logo and a description: 'The Genomic Standards Consortium (GSC) is an open-membership working body formed in September 2005. The aim of the GSC is making genomic data discoverable. The GSC enables genomic data integration, discovery and comparison through international community-driven standards.' It also lists the maintainer 'rwalls' with an ORCID link, and provides dates for when the record was added and updated. There are buttons for 'Homepage' and 'Reference'. A 'Taxonomic range' section shows 'All' selected. A 'Knowledge Domains' section shows 'Genome' selected. A 'Subjects' section shows 'Genomics' selected. There are buttons for 'View as Graph' and 'Show edit history'. A 'Compare with collection/recommendation (Beta)' section is present. At the bottom, there is a 'General collection/recommendation statistics' section with a 'Show Stats' button.

The screenshot shows the FAIRsharing.org Graph Viewer for the collection 'bsg-c000040'. The header includes the FAIRsharing.org logo and a navigation menu. Below the header, the collection ID 'bsg-c000040' is displayed. The main content area features a network graph titled 'Graph Viewer: Collections > Genomic Standards Consortium'. The graph shows a central node 'Genomic Standards Consortium' (red circle) connected to several other nodes: 'MIBIG' (yellow circle), 'MixS' (yellow circle), 'MixS - MIMARKS' (yellow circle), 'MixS - MIGS/MIMS' (yellow circle), 'GCDML' (yellow circle), and 'GAZ' (yellow triangle). A legend on the left side of the graph defines the symbols: red circle for DATABASE, green circle for POLICY, red circle for COLLECTION, yellow triangle for TERMINOLOGY ARTIFACT, yellow diamond for MODEL/FORMAT, yellow circle for IDENTIFIER SCHEMA, yellow circle for REPORTING GUIDELINE, yellow circle for METRIC, green line for RECOMMENDS, blue line for COLLECTS, purple line for IMPLEMENTS, and red line for RELATED TO. A donut chart on the left shows the count of 'Standards' as 6.

<https://fairsharing.org/collection/GSC>

<https://fairsharing.org/graph/#/collection/bsg-c000040>



- An international community-driven standard in **Genomics** producer of the ***MlxS: Minimum Information Standards about any(X) Sequence***
- MlxS includes **technology-specific checklists** (MIGS, MIMS, MIMARKS,...) and also allows **annotation of sample data** using environmental packages

Specification projects	MIGS	MIMS	MIMARKS	New checklists
Checklists	EU BA PL VI ORG	metagenomes	survey specimen	e.g., pan-genomes
Shared descriptors	collection date, environmental package, environment (biome), environment (feature), environment (material), geographic location (country and/or sea, region), geographic location (latitude and longitude), investigation type, project name, sequencing method, submitted to INSDC			
Checklist-specific descriptors	assembly, estimated size, finishing strategy, isolation and growth condition, number of replicons, ploidy, propagation, reference for biomaterial		target gene	
Applicable environmental packages (measurements and observations)	Air Host-associated Human-associated Human-oral Human-gut Human-skin Human-vaginal		Microbial mat/biofilm Miscellaneous natural or artificial environment Plant-associated Sediment Soil Wastewater/sludge Water	

Source: <https://gensc.org>

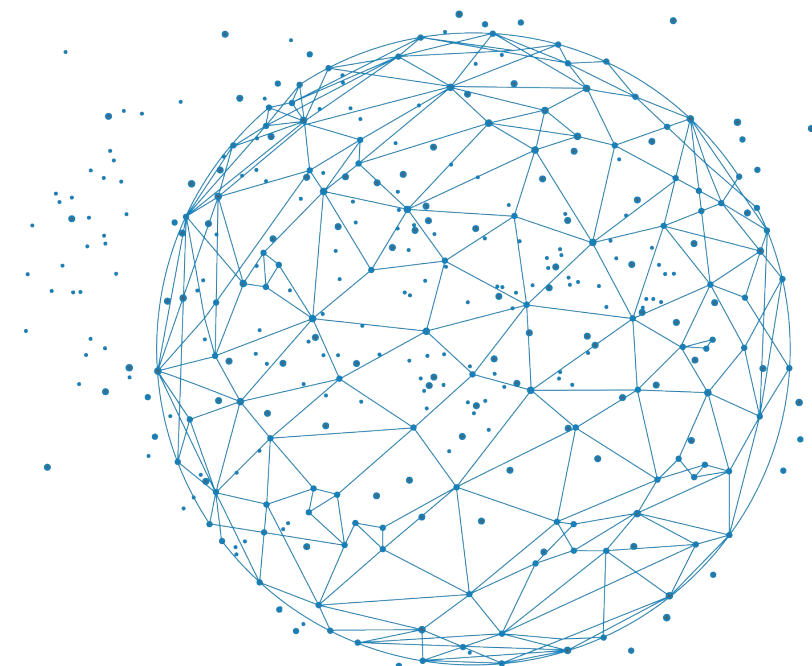
[Yilmaz et al, 2011](#)



Description	Name	URL
A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies.	FAIRsharing portal	https://fairsharing.org
Investigation, Study, Assay (ISA) resource: A standard model and a set of tools to capture experimental data in life sciences	ISAtools	https://isa-tools.org
Genomics Standard Consortium (GSC): An international consortium developing standards and checklists in genomics	GSC	https://gensc.org
RDMkit: Documentation and metadata	RDMkit documentation and metadata	https://rdmkit.elixir-europe.org/metadata_management.html



Retour d'expérience de soumission en banque de données internationales





Connectez-vous sur www.wooclap.com/TDSUTZ



- Open science
- La reproductibilité des expériences
- Donner accès à mes données
- Archiver mes données
- Publication d'articles
- Analyser mes données




3 bases de données





ENA



European Nucleotide Archive



Qui a déjà
soumis à
l'ENA ?



C'était facile ?



Plateforme ouverte pour la gestion, le partage, l'intégration, l'archivage et la diffusion des données de séquençage.

Connecté avec UniProt, RNACentral, EBI Metagenomics, Ensembl, Ensembl Genomes, ArrayExpress, ...

Des données variées: génomique animale, la biotechnologie marine, la biodiversité, la surveillance des agents pathogènes et la biologie des cellules souches



🏠 ENA Training Modules
latest

ENA DATA SUBMISSION

- General Guide On ENA Data Submission
- How to Register a Study
- How to Register Samples
- Preparing Files for Submission
- How to Submit Raw Reads
- How to Submit Assemblies
- How to Submit Targeted Sequences
- How to Submit Other Analyses

ENA DATA DISCOVERY & RETRIEVAL

- General Guide on ENA Data Retrieval
- How to Explore an ENA Project
- How to Download Data Files
- How To Perform An Advanced Search
- How to Access ENA Programmatically

ENA DATA UPDATES

- Updating Metadata Objects
- Updating Assemblies
- Updating Annotated Sequences

TIPS AND FAQs

- Data Release Policies
- Common Run Submission Errors
- Tips for Sample Taxonomy
- Requesting New Taxon IDs
- Metagenome Submission Queries
- Locus Tag Prefixes
- Archive Generated FASTQ Files
- Third Party Tools

Docs » ENA: Guidelines and Tutorials [Edit on GitHub](#)

ENA: Guidelines and Tutorials

Welcome to the guidelines for submission and retrieval for the European Nucleotide Archive. Please use the links to find instructions specific to your needs. If you're completely new to ENA, you can see an introductory webinar at the bottom of the page.

ENA Data Submission

- [General Guide On ENA Data Submission](#)
- [How to Register a Study](#)
- [How to Register Samples](#)
- [Preparing Files for Submission](#)
- [How to Submit Raw Reads](#)
- [How to Submit Assemblies](#)
- [How to Submit Targeted Sequences](#)
- [How to Submit Other Analyses](#)

ENA Data Discovery & Retrieval

- [General Guide on ENA Data Retrieval](#)
- [How to Explore an ENA Project](#)
- [How to Download Data Files](#)
- [How To Perform An Advanced Search](#)
- [How to Access ENA Programmatically](#)

ENA Data Updates

- [Updating Metadata Objects](#)
- [Updating Assemblies](#)
- [Updating Annotated Sequences](#)

Tips and FAQs

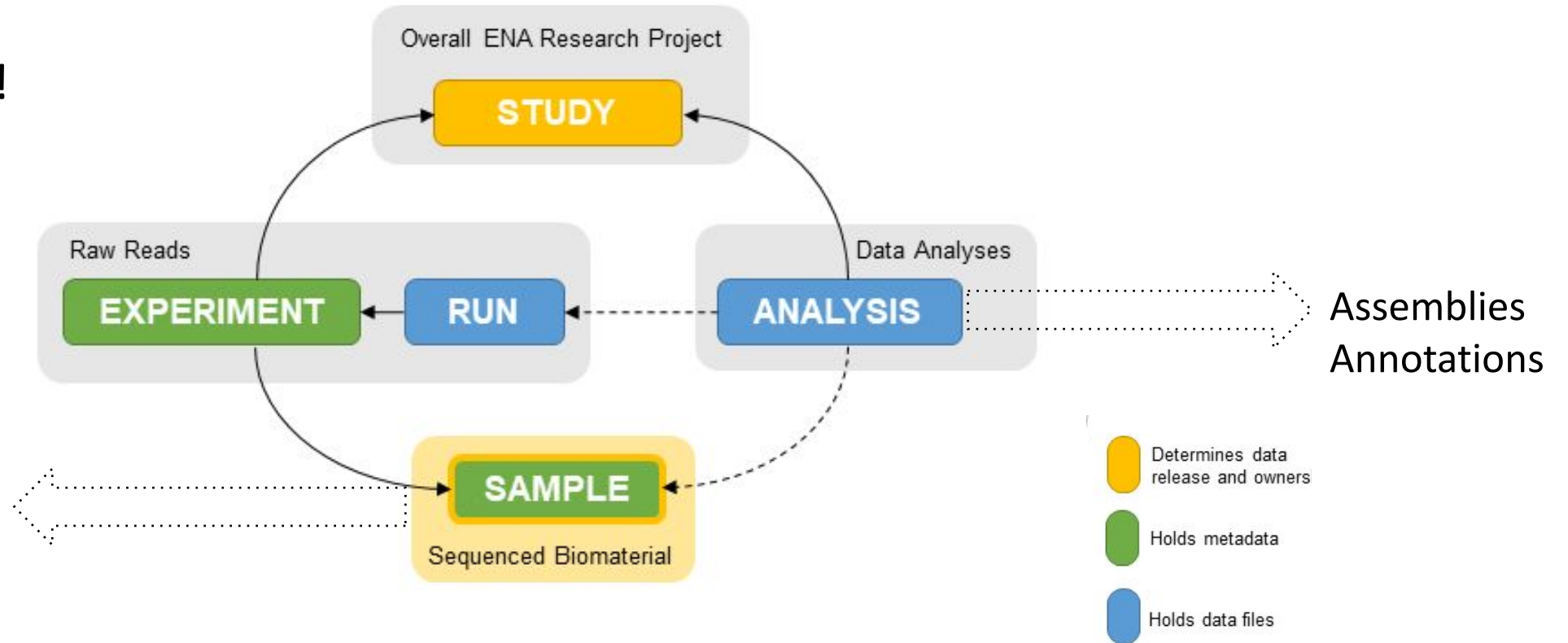
- [Data Release Policies](#)
- [Common Run Submission Errors](#)
- [Tips for Sample Taxonomy](#)
- [Requesting New Taxon IDs](#)
- [Metagenome Submission Queries](#)
- [Locus Tag Prefixes](#)
- [Archive Generated FASTQ Files](#)
- [Third Party Tools](#)

<https://ena-docs.readthedocs.io/en/latest/>



ISA compliant !

All **samples** submitted to ENA must conform to a **Checklist**



Source:

<https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html>



Metadata validation

Permitted values for platform

- LS454: 454 technology use 1-color sequential flows
- ILLUMINA: Illumina is 4-channel flowgram with 1-to-1 mapping between basecalls and flows
- PACBIO_SMRT: PacificBiosciences platform type for the single molecule real time (SMRT) technology.
- ION_TORRENT: Ion Torrent Personal Genome Machine (PGM) from Life Technologies.
- CAPILLARY: Sequencers based on capillary electrophoresis technology manufactured by LifeTech (formerly Applied BioSciences).
- OXFORD_NANOPORE: Oxford Nanopore platform type. nanopore-based electronic single molecule analysis.
- BGISEQ
- DNBSEQ

<https://ena-docs.readthedocs.io/en/latest/submit/reads/webin-cli.html?permitted-values-for-instrument>



- A **checklist** defines the **minimum and optional metadata** expected to describe biological samples
- ENA are based on the **Genomic Standards Consortium (GSC)** recommandations
- The **most suitable checklist** depends on the type of the sample:
<https://www.ebi.ac.uk/ena/browser/checklists>
- All ENA checklist are defined by an **access number** like ERCxxx (Ena R Checklist xxx)
 - example: GSC MixS plant associated
<https://www.ebi.ac.uk/ena/browser/view/ERC000020>



EMBL-EBI Services Research Training About us

ENA
European Nucleotide Archive

Enter text search terms Search

Examples: histone, BN000065

Enter accession View

Examples: Taxon:9606, BN000065, PRJEB402

Home Submit Search Rulespace About Support

Sample Checklists

There is a minimum amount of information required during ENA sample registration and all samples must conform to a defined checklist of expected metadata values. The most suitable checklist for sample registration depends on the type of the sample.

These sample checklists have been developed to meet the needs of different research communities. Different communities have different requirements on the minimum metadata expected to describe biological samples.

Accession	Name	Description
ERC000012	GSC MixS air	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000013	GSC MixS host associated	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000014	GSC MixS human associated	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000015	GSC MixS human gut	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000016	GSC MixS human oral	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000017	GSC MixS human skin	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...
ERC000018	GSC MixS human vaginal	Genomic Standards Consortium package extension for reporting of measurements and observations obtain...



ENA
European Nucleotide Archive

Enter text search terms Search

Examples: histone, BN000065

ERC000033 View

Examples: taxon:3606, BN000065, PRJEB402

Home Submit Search Rulespace About Support

Checklist: ERC000033

ENA virus pathogen reporting standard checklist

Minimum information about a virus pathogen. A checklist for reporting metadata of virus pathogen samples associated with genomic data. This minimum metadata standard was developed by the COMPARE platform for submission of virus surveillance and outbreak data (such as Ebola) as well as virus isolate information.

View: XML

Download: XML

Checklist Fields

Filter fields...

Filter by type:

- Human surveillance data
- Collection event information
- sample collection
- host disorder
- host description
- Virus isolate information
- General collection event information
- Serology detection
- Infraspecies information
- Associated host information
- host details
- Environmental information

Field Name	Field Format	(Field Restriction)	Requirement	(Units)
subject exposure	free text		optional	
subject exposure duration	free text		optional	
type exposure	free text		optional	
personal protective equipment	free text		optional	
hospitalisation	text choice	options	optional	
illness duration	free text		optional	
illness symptoms	free text		optional	
collection date	restricted text	regular expression	recommended	
geographic location (country and/or sea)	text choice	options	mandatory	
geographic location (latitude)	restricted text	regular expression	recommended	DD
geographic location (longitude)	restricted text	regular expression	recommended	DD
geographic location (region and locality)	free text		recommended	
sample collection status	text choice	options	recommended	



	Interactive	Webin-CLI	Programmatic
Study	Y	N	Y
Sample	Y	N	Y
Read data	Y	Y	Y
Genome Assembly	N	Y	N
Transcriptome Assembly	N	Y	N
Template Sequence	N	Y	N
Other Analyses	N	N	Y



Dashboard

Welcome to the Webin Submissions Portal

You can use this service for a range of submission activities as well as reports on your submissions. For help with submitting your data, including the use of this interface, please refer to our [Help Guides](#). Please familiarise yourself with the different submission interfaces and what can be submitted through each by reading our [General Guide on ENA Data Submission](#). All users are advised to take a moment to understand the [ENA Metadata Model](#). You may also like to review how the release of data is managed in our [Data Release FAQ](#).

A dedicated submission API for COVID-19 genomes is available [here](#).

Studies (Projects)

- + Register Study
- + Submit XMLs (advanced)
- Studies Report

Samples

- + Register Samples
- + Register Novel Taxonomy
- + Submit XMLs (advanced)
- Samples Report

Raw Reads (Experiments and Runs)

Raw reads can also be submitted using [Webin-CLI](#)

- + Submit Reads
- + Submit XMLs (advanced)
- Runs Report
- Run Files Report
- Run Processing Report
- Unsubmitted Files Report

Data Analyses

Assemblies and annotated sequences must be submitted with [Webin-CLI](#). Other analyses can be submitted as XMLs.

- + Generate Annotated Sequence Spreadsheet
- + Submit XMLs (advanced)
- Analyses Report
- Analysis File Report
- Analysis Processing Report



v4.2.1

Latest

Compare ▾

Rajkumar-D released this 26 days ago v4.2.1 0d34c7a

- sequence context: Added support for BioSample accessions, SRA Sample accessions and SRA Sample aliases in the ORGANISM field in addition to the already supported NCBI taxonomy names and IDs.

▼ Assets 4

webin-cli-4.2.1-sources.jar	109 KB
webin-cli-4.2.1.jar	61.5 MB
Source code (zip)	
Source code (tar.gz)	





- **SUBMISSION** (XML Schema)
- **STUDY** (XML Schema)
- **SAMPLE** (XML Schema)
- **EXPERIMENT** (XML Schema)
- **RUN** (XML Schema)
- **ANALYSIS** (XML Schema)
- **DAC** (XML Schema)
- **POLICY** (XML Schema)
- **DATASET** (XML Schema)
- **PROJECT** (XML Schema)

Exemple : submission.xml

```
<SUBMISSION>
  <ACTIONS>
    <ACTION>
      <ADD/>
    </ACTION>
  </ACTIONS>
</SUBMISSION>
```



Submit new data

Information on how to submit COVID-19 data

We have a new [drag-and-drop data submission tool](#), suitable for viral sequence submissions. We are inviting volunteers to try it out - please register your interest below.

Data types

Viral, non-human and cell line sequence data

Human molecular biology data

Linked viral and human molecular biology data

Viral and non-human proteomics data

Structural biology data

Viral and non-human molecular interaction data

Viral and non-human metabolomics data

Viral and other non-human molecular biology data

Compound and target data

Clinical and epidemiological data

Non-biological data

Viral, non-human and cell line sequence data

This class includes sequence data from studies targeting virus alone or with co-occurring species. It also includes sequencing from non-human host species (such as from species acting as models for infection) and human cell lines (where data are consented for full open publication). All sequencing library types, all platforms, all library methods and all levels of processing (from raw data to assembled sequences) are included in this class.

Deposition actions:

Users should submit data to ENA
 Specific deposition instructions are available for viral data submission
 Users are encouraged to contact ENA at virus-dataflow@ebi.ac.uk

General depositions and those from users who are managing their data in SARS-CoV-2 Data Hubs are also included in this class.

Drag and Drop viral sequence submission tool

We have a new [drag-and-drop data submission tool](#), which is suitable for many viral sequence submissions. Please register your interest and we will be in contact to assess the suitability of the tool for your data set.

[Register](#)



Checklist: ERC000033

ENA virus pathogen reporting standard checklist

Minimum information about a virus pathogen. A checklist for reporting metadata of virus pathogen samples associated with genomic data. This minimum metadata standard was developed by the COMPARE platform for submission of virus surveillance and outbreak data (such as Ebola) as well as virus isolate information.

[View: XML](#)
[Download: XML](#)

Checklist Fields

Filter fields...	Field Name	Field Format	(Field Restriction)	Requirement	(Units)
	subject exposure	free text		optional	
	subject exposure duration	free text		optional	
	type exposure	free text		optional	
	personal protective equipment	free text		optional	
	hospitalisation	text choice	options	optional	
	illness duration	free text		optional	
	illness symptoms	free text		optional	
	collection date	restricted text	regular expression	recommended	
	geographic location (country and/or sea)	text choice	options	mandatory	
	geographic location (latitude)	restricted text	regular expression	recommended	DD
	geographic location (longitude)	restricted text	regular expression	recommended	DD
	geographic location (region and locality)	free text		recommended	

<https://www.ebi.ac.uk/ena/browser/view/ERC000033>



Tools & Data Resources

🔍 Search all tools & data resources

Tools

Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.



Web API Multiple sequence alignment

InterProScan

InterProScan searches sequences against InterPro's predictive protein signatures.



Web API Protein feature detection
Sequence motif recognition

BLAST [protein]

Fast local similarity search tool for protein sequence databases.



Web API Sequence similarity search

BLAST [nucleotide]

Fast local similarity search tool for nucleotide sequence databases.



Web API Sequence similarity search

HMMER

Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.



Web API Sequence similarity search
Protein function prediction

See all tools >

Data resources

Ensembl

Genome browser, API and database, providing access to reference genome annotation



Web API

UniProt

A comprehensive resource for protein sequence and functional annotation.



Web API

PDBe

The European resource for the collection, organisation and dissemination of 3D structural data (from PDB and EMDB) on biological macromolecules and their complexes.



Web API

Europe PMC

A database to search the worldwide life sciences literature



Web API

Expression Atlas

An added-value database that shows which genes/proteins are expressed under which conditions, and how expression differs between conditions.



Web API

ChEMBL

An open data resource of binding, functional and ADMET bioactivity data.



Web API

See all data resources >

Getting started

Search by

Name, biome, or keyword

Text search

Sequence similarity

Sequence search

Or by data type

XXX	354951 amplicon	3745 studies
	27960 assemblies	326190 samples
	2050 metabarcoding	434691 analyses
	33933 metagenomes	
	2217 metatranscriptomes	

Or by selected biomes

Human (141734)	Digestive system (94341)	Aquatic (45990)	Marine (33451)	Digestive system (32651)
Plants (26768)	Soil (23684)	Skin (10501)	Wastewater (3858)	Food production (2805)

[Browse all biomes](#)

Request analysis of

Your data A public dataset

Submit and/or Request Request

Latest studies

[EMG produced TPA metagenomics assembly of the Microbial composition of samples from infant gut \(human gut metagenome\) data set](#)

The human gut metagenome Third Party Annotation (TPA) assembly was derived from the primary whole genome shotgun (WGS) data set PRJNA63661. This project includes samples from the following biomes : Human gut.
[View more - 325 samples](#)

[EMG produced TPA metagenomics assembly of PRJNA274897 data set \(Oil droplet biodegradation Trondheimsfjord Metagenome\).](#)

The Third Party Annotation (TPA) assembly was derived from the primary whole genome shotgun (WGS) data set PRJNA274897, and was assembled with metaSPAdes v3.13.0. This project includes samples from the following biomes: root:Engineered:Lab enrichment...
[View more - 14 samples](#)

PMC 728.11_cyano

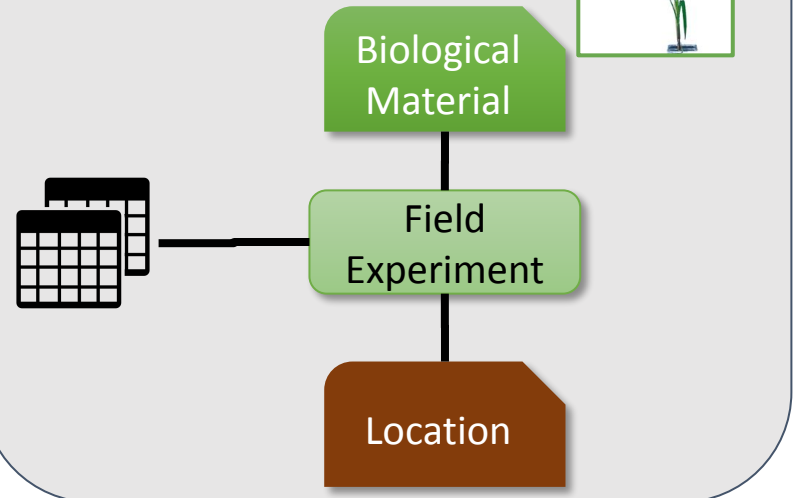
[Microcystis aeruginosa PMC 728.11_cyano metagenome sequencing](#) [View all studies](#)



Data Integration between silos, From Phenotyping to Genotyping

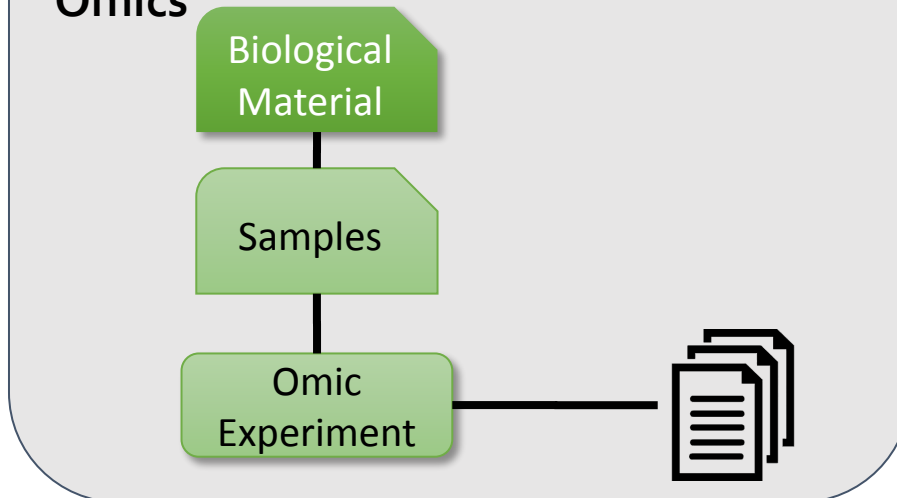
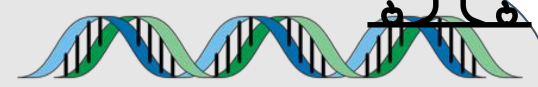


Phenomics

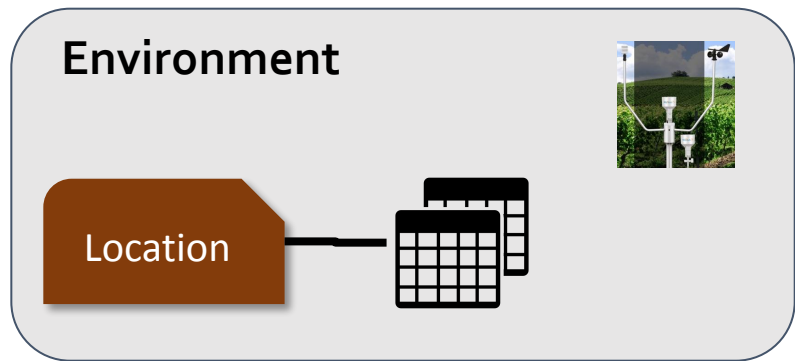
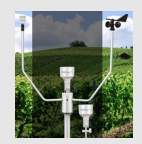


Identifying key resources/pivot objects

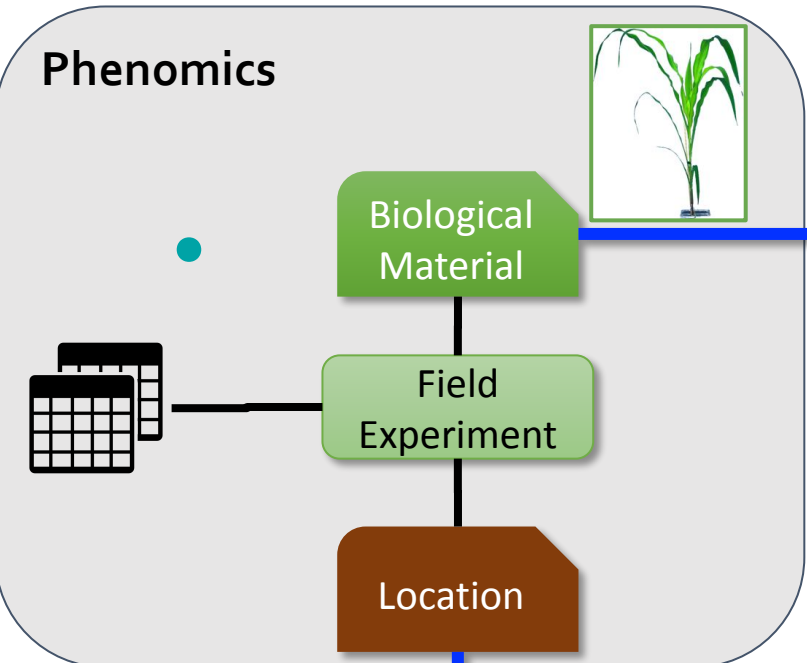
Genetics Genomics Omics



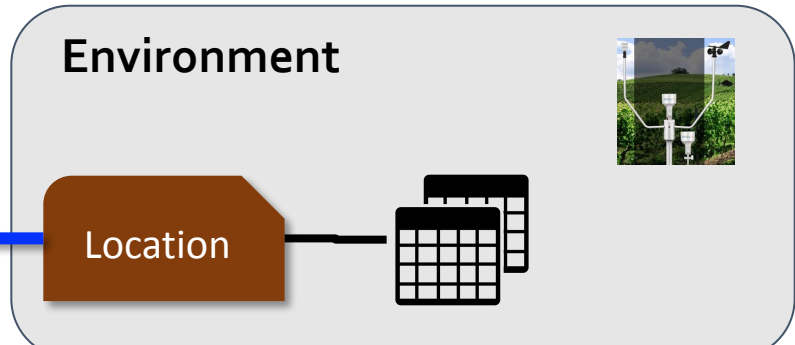
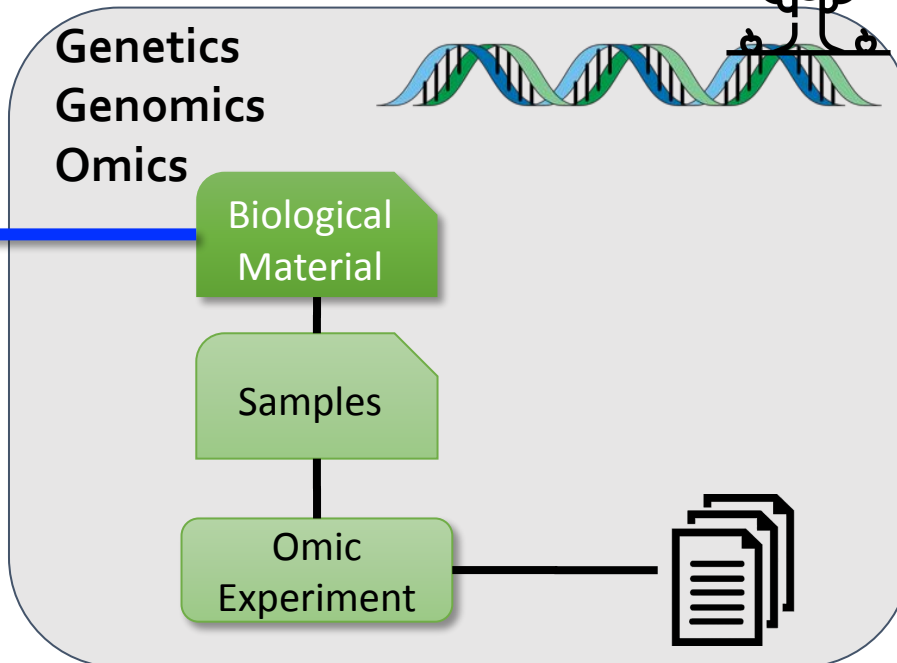
Environment



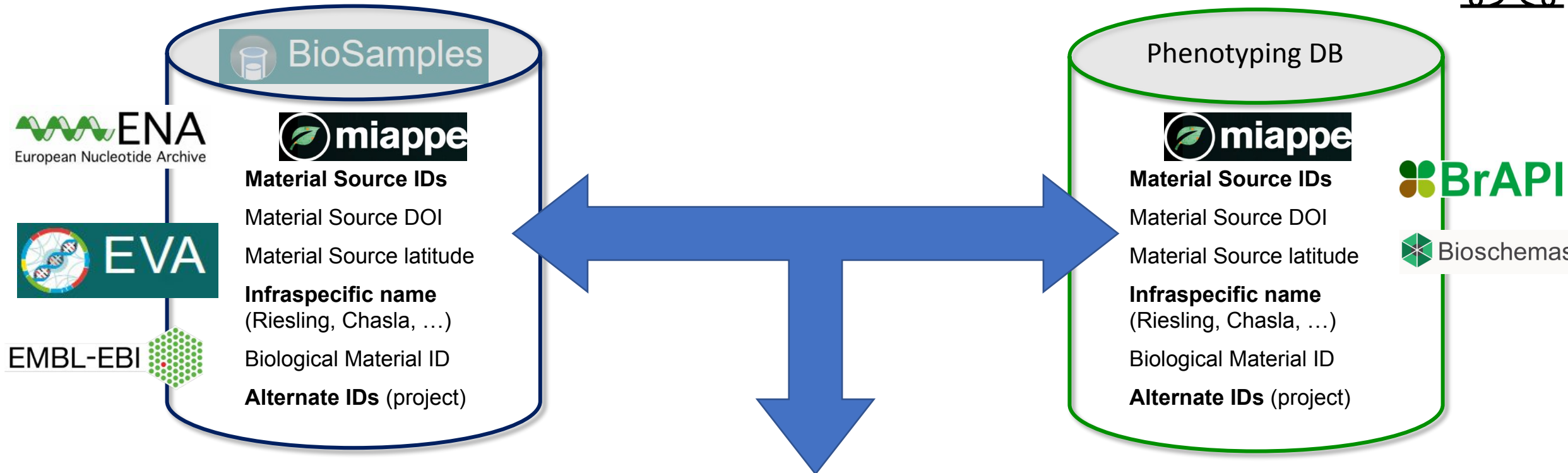
Data Integration between silos, From Phenotyping to Genotyping



Interoperability pivot
Key shared resources



Data Integration between silos, From Phenotyping to Genotyping



ENA
European Nucleotide Archive

EVA

EMBL-EBI

BrAPI

Bioschemas

URGI ▾ Data providers ▾ More... ▾ <https://urgi.versailles.inrae.fr/aidare/>

FAIR Data-finder for Agronomic REsearch

Sources

- URGI GnpIS (81,335)
- EBI European Nucleotide Archive (44,975)
- CIRAD TropGENE (722)
- VIB PIPPA (692)
- IBET BioData (67)
- IWGSC@GnpIS (18,814,632)
- Evoltree@GnpIS (5,354)
- OpenMinTeD@GnpIS (3,392)
- EBI Ensembl Plants

Search keywords

Germplasm
Trait
Reset all

Crops
(common name, species, genus, subtaxa & synonyms)

Search crops

Germplasm list
(panel, collection & population)

Search germplasm lists

Community data discovery portals



General recipe for plant submission @ ENA

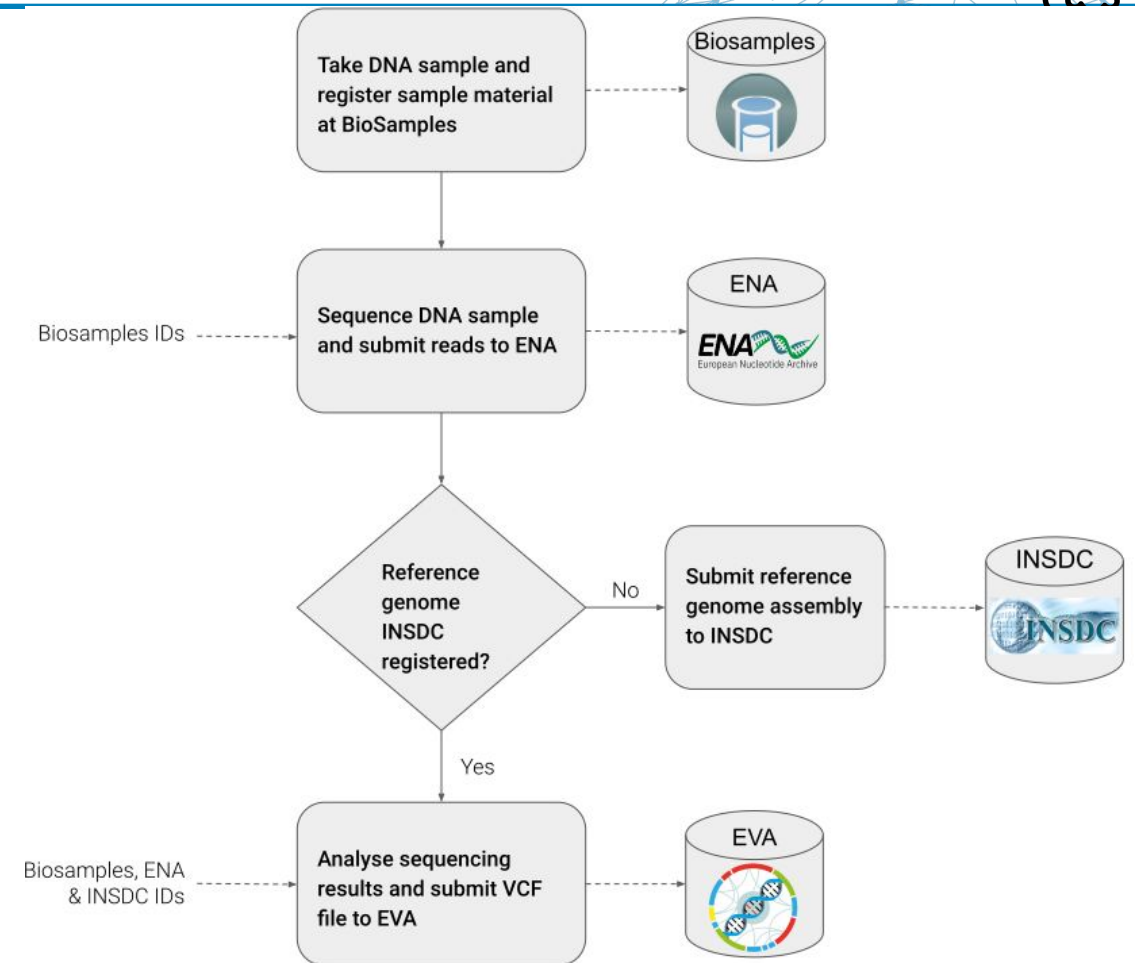
https://github.com/FAIRplus/the-fair-cookbook/blob/plant_miappe/content/recipes/reusability/miappe.md

Plant Checklist




<https://www.ebi.ac.uk/biosamples/schemas/certification/plant-miappe.json>

Plus a validator

<https://www.ebi.ac.uk/biosamples/docs/guides/validation>





<input type="checkbox"/>		00_Pheno-Geno-Plant-research-dataset-at-data.inrae.fr-URGI.pdf application/pdf - 363.9 Ko - 20 sept. 2021 - 35 téléchargements MD5: fe3837e8d2769cbcd6776fa6c73756e0
<input type="checkbox"/>		BiologicalMaterial.xlsx application/vnd.openxmlformats-officedocument.spreadsheetml.sheet - 15.2 Ko - 20 sept. 2021 - 14 téléchargements MD5: c96ce678c08f63472e5b0f571ae1ba37
<input type="checkbox"/>		ObservedVariables.xlsx application/vnd.openxmlformats-officedocument.spreadsheetml.sheet - 16.9 Ko - 20 sept. 2021 - 15 téléchargements MD5: 54a807d0706e3caa618d63bf3674c450



De
•
Metadata templates

Field	Accession_Number	accession_holding	Material source DOI	Material source ID (Holding institute/stock centre, accession)	Biological material ID*	Organism*
Definition			Digital Object Identifier (DOI) of the material source	An identifier for the source of the biological material, in the form of a key-value pair comprising the name/identifier of the repository from which the material was sourced plus the accession number of the repository for that material. Where an accession number has not been assigned, but the material has been derived from the crossing of known	Code used to identify the biological material in the data file . Should be unique within the Investigation. Can correspond to experimental plant ID, seed lot ID, etc... This material identification is different from a BiosampleID which corresponds to Observation Unit or Samples sections below.	An identifier for the organism species level. Use of the NCE ID is recommended.
Example			doi:10.15454/1.4658436467893904E12	INRA:W95115_inra ICNF:PNB-RPI	INRA:W95115_inra_2001; INRA:inra_kemel_2351; Rothamsted:res_GK090847	NCBITAXON:4577
Format			DOI	Unique identifier	Unique identifier	Unique identifier
	A3_H	inra		inra:A3	A3_H	NCBITAXON:4577
	A310_H	inra		inra:A310	A310_H	NCBITAXON:4577
	R73_H	inra		inra:R73	R73_H	NCBITAXON:4577



GEO

Gene Expression Omnibus



Qui a déjà
soumis à
GEO ?

C'était facile ?





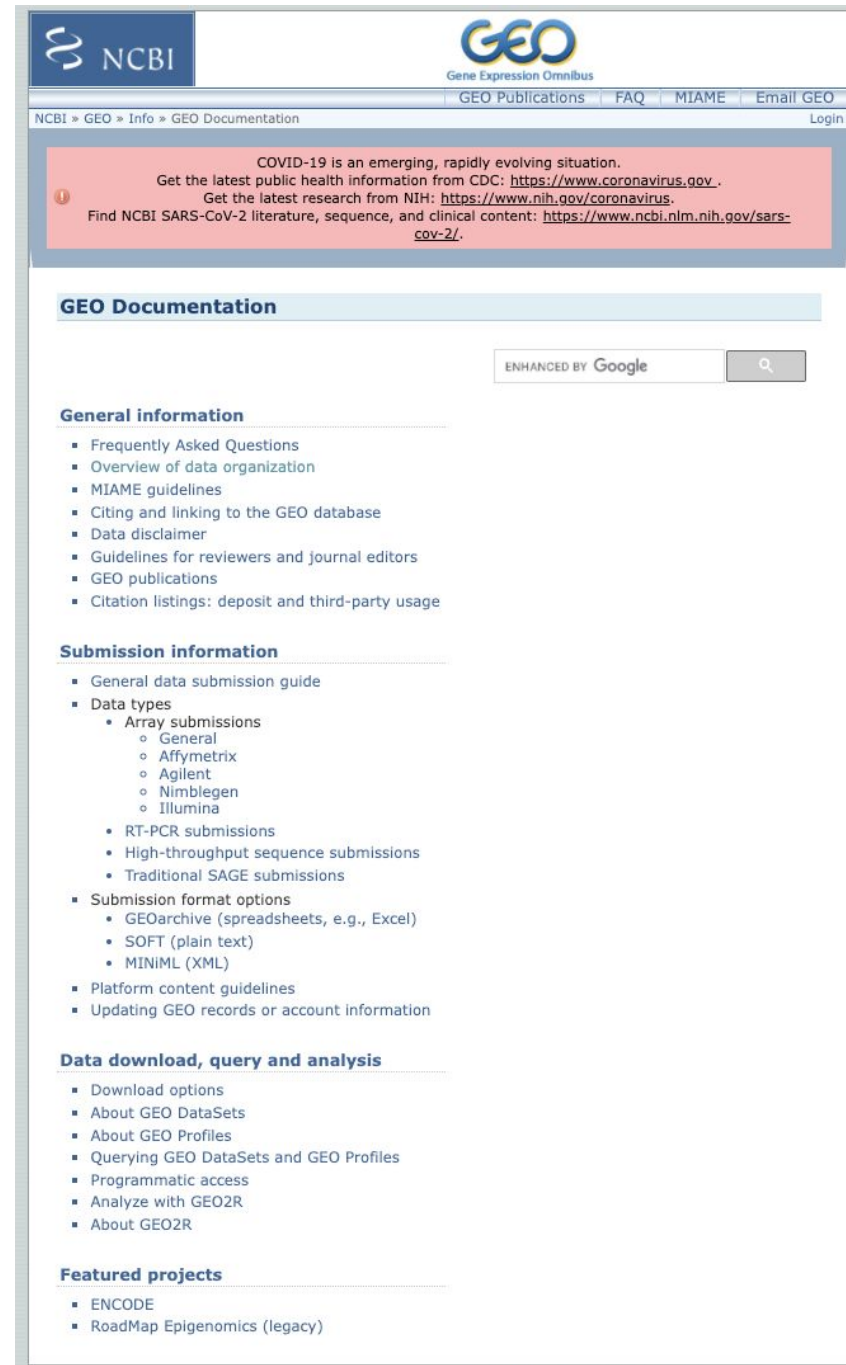
GEO est un dépôt public international qui archive et distribue librement des données de:

- microarray ;
- de NGS ;
- et d'autres formes de données de génomique fonctionnelle à haut débit .

soumises par la communauté des chercheurs.

Documentation

<https://www.ncbi.nlm.nih.gov/geo/info/>



NCBI GEO Info GEO Documentation Login

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.coronavirus.gov>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

GEO Documentation

ENHANCED BY Google

General information

- Frequently Asked Questions
- Overview of data organization
- MIAME guidelines
- Citing and linking to the GEO database
- Data disclaimer
- Guidelines for reviewers and journal editors
- GEO publications
- Citation listings: deposit and third-party usage

Submission information

- General data submission guide
- Data types
 - Array submissions
 - General
 - Affymetrix
 - Agilent
 - Nimblegen
 - Illumina
 - RT-PCR submissions
 - High-throughput sequence submissions
 - Traditional SAGE submissions
- Submission format options
 - GEOarchive (spreadsheets, e.g., Excel)
 - SOFT (plain text)
 - MINIML (XML)
- Platform content guidelines
- Updating GEO records or account information

Data download, query and analysis

- Download options
- About GEO DataSets
- About GEO Profiles
- Querying GEO DataSets and GEO Profiles
- Programmatic access
- Analyze with GEO2R
- About GEO2R

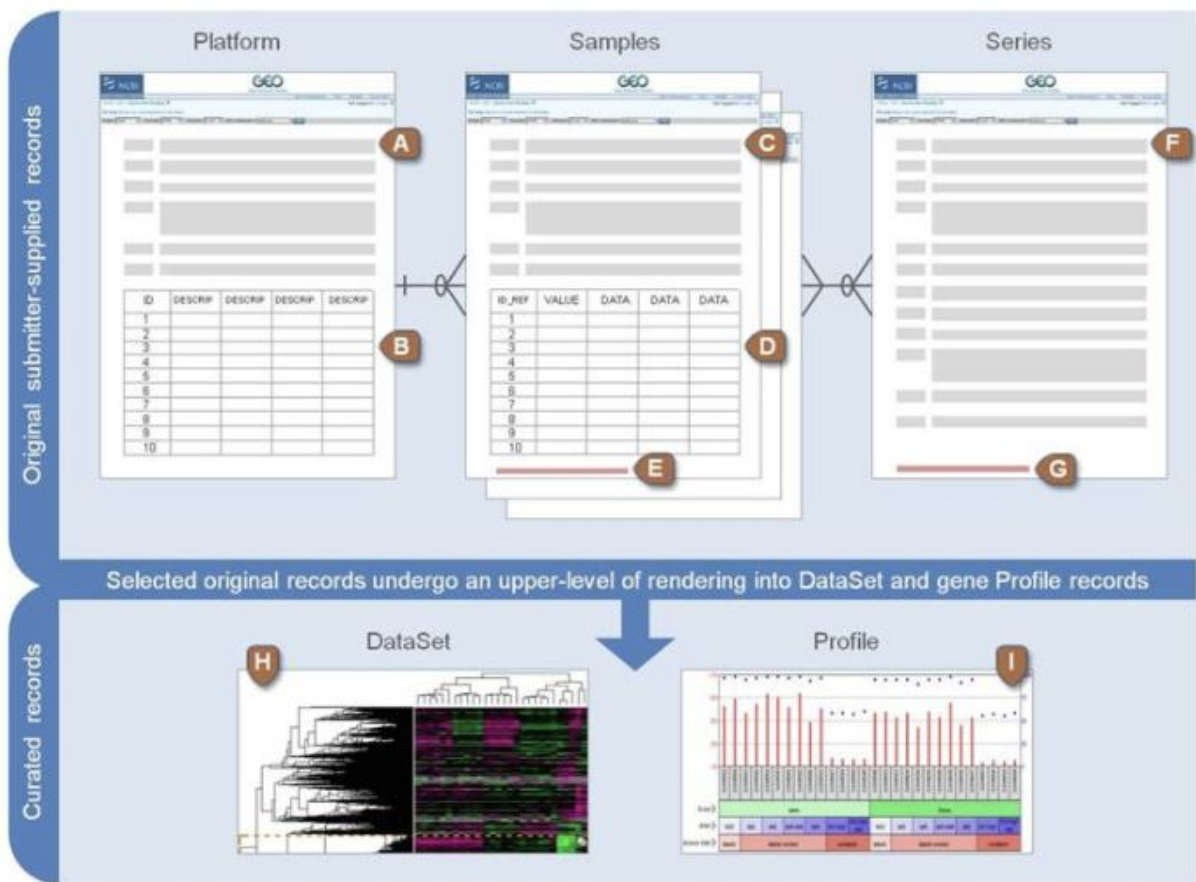
Featured projects

- ENCODE
- RoadMap Epigenomics (legacy)





Organisation des données



Platform	<p>Platform records are supplied by submitters</p> <p>A Platform record is composed of a summary description of the array or sequencer and, for array-based Platforms, a data table defining the array template. Each Platform record is assigned a unique and stable GEO accession number (GPLxxx). A Platform may reference many Samples that have been submitted by multiple submitters. Example Platform record »</p>	<p>A Text description of the array or sequencer</p> <p>B Text tab-delimited table of the array template</p>
Sample	<p>Sample records are supplied by submitters</p> <p>A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it. Each Sample record is assigned a unique and stable GEO accession number (GSMxxx). A Sample entity must reference only one Platform and may be included in multiple Series. Example Sample record »</p>	<p>C Text description of the biological sample and protocols to which it was subjected</p> <p>D Text tab-delimited table of processed hybridization result (may optionally include raw data columns)</p> <p>E Original raw data file, or processed sequence data file</p>
Series	<p>Series records are supplied by submitters</p> <p>A Series record links together a group of related Samples and provides a focal point and description of the whole study. Series records may also contain tables describing extracted data, summary conclusions, or analyses. Each Series record is assigned a unique and stable GEO accession number (GSExxx). Example Series record »</p>	<p>F Text description of the overall experiment</p> <p>G Tar archive of original raw data files, or processed sequence data files</p>



GEOarchive format

GEOarchive is a flexible spreadsheet-based submission format useful for batch deposit of experiments. GEOarchive submissions can be created in any spreadsheet software, usually Microsoft Excel.

A GEOarchive submission consists of several parts as follows:

Metadata spreadsheet	'Metadata' refers to descriptive information and protocols for the overall experiment and individual Samples. This information is supplied by completing all fields of the appropriate metadata spreadsheet template which can be downloaded from the GEOarchive templates and examples section below.
Matrix table	The matrix table is a spreadsheet containing the final, normalized values that are comparable across rows and Samples, and preferably processed as described in any accompanying manuscript. A complete data matrix should be supplied, not a summary subset. It is possible to include additional data columns in the table, for example, Affymetrix Detection calls and P-values, or background or flag columns. See the Affymetrix template for an example.
Raw data files	In addition to the normalized data provided in the Matrix table, submitters are required to provide raw data, usually in the form of supplementary raw data files. This facilitates the unambiguous interpretation of the data and potential verification of the conclusions as described in the MIAME and MINSEQE standards. Affymetrix submissions must include CEL files. Non-Affymetrix GEOarchive submissions should include the original software-generated scan quantification files, for example, GenePix GPR files. Next-generation sequence submissions must include files containing reads and quality scores.
Platform	If your experiments are performed using a commercial array (e.g., Affymetrix GeneChip) or other array already deposited in GEO, please use the FIND PLATFORM tool to find the GEO accession number (GPLxxxx) for inclusion in the 'platform' column in the <i>SAMPLES</i> section of the metadata spreadsheet. If your array does not already exist in GEO, please include a <i>PLATFORM</i> section in your metadata spreadsheet and include Platform annotation columns in your matrix table. The Platform data must include meaningful, trackable, sequence identifiers (e.g. GenBank/RefSeq accessions, locus tags, clone IDs, oligo sequences, chromosome locations, etc - see the Platform content guidelines for full list). References to in-house databases or top BLAST hits are not sufficient. Platform submission is not necessary for SAGE or next-generation sequence submissions.

Bundle all parts (Excel file containing the metadata spreadsheet and matrix spreadsheet, raw data files) together into a .zip, .rar, or .tar archive using a program like WinZip, and transfer to GEO using the 'Transfer files to GEO with web form' option on the [Submit to GEO](#) page. Incomplete submissions will result in processing delays.

Submit

GEOarchive templates and examples

The first step in creating your GEOarchive submission is to download the appropriate template (Excel spreadsheet) from the list below. Each Excel file consists of several worksheets, including a metadata template, and examples of metadata and matrix tables. Click the tabs at the bottom of the worksheet window to switch between worksheets. Mouse over field names in the templates to view content guidelines.

Microarray

For the following microarray vendors, please download templates from the vendor-specific instructions pages:

- [Affymetrix submissions](#)
- [Agilent submissions](#)
- [Nimblegen submissions](#)
- [Illumina submissions](#)

For microarrays not from the vendors above, please use a 'Generic' template. For generic microarray submissions where the Platform is already deposited in GEO, please download the most appropriate template:

- [Generic single channel submission template](#)
- [Generic dual channel submission template](#)
- [Generic merged dye-swap submission template](#)
- [Generic tiling ChIP-chip submission template](#)

For generic microarray submissions where the Platform is not deposited in GEO, please download the most appropriate template:

- [Generic single channel submission template, including Platform](#)
- [Generic dual channel submission template, including Platform](#)
- [Generic merged dye-swap submission template, including Platform](#)
- [Generic tiling ChIP-chip submission template, including Platform](#)

To submit only a Platform, please download the following template (this option is appropriate only if you have no hybridization or sequence data to deposit):

- [Platform-only template](#)

High-throughput sequencing

For high-throughput sequence submissions, please refer to full instructions at:

- [High-throughput sequence submissions](#)

Other data types

For NanoString submissions, please use one of the 'Generic single channel' templates as appropriate:

- [Generic single channel submission template](#)
- [Generic single channel submission template, including Platform](#)

For high-throughput RT-PCR submissions, please refer to full instructions at:

- [RT-PCR submissions](#)

For traditional SAGE submissions, please refer to full instructions at:

- [Traditional SAGE submissions](#)



GA_illumina_expression.xls [Mode de compatibilité]

Accueil Insertion Dessin Mise en page Formules Données Révision Affichage

Arial 10 A A Standard

Mise en forme conditionnelle Mettre sous forme de tableau Styles de cellule

Insérer Supprimer Mise en forme

Trier et filtrer Rechercher et sélectionner

F7

	A	B	C	D	E	F	G	H	I	J	K	L
1	SERIES											
2	title	Genome-wide analysis of mechano-responsive gene expression by tenocytes in fascicles subjected to cyclic tensile strain										
3	summary	Analysis of mechano-regulation of tenocyte metabolism at gene expression level. The hypothesis tested in the present study was that cyclic tensile strain influence the balance of anabolism/catabolism of tenocytes. Results provide important information of the response of tenocyte										
4	overall design	Total RNA obtained from isolated tendon fascicles subjected to 1 or 24 hours in vitro cyclic tensile strain compared to unstrained control fascicles.										
5	contributor	Jane,Doe										
6	contributor	John,A,Smith										
7												
8	SAMPLES											
9		# The corresponding example matrix table is included in the next worksheet.										
10	Sample name	title	source name	organism	idat file	characteristics: Strain	characteristics: age	characteristics: tiss	molecule	label	description	platform
11	Sample 1	Fascicle Strained 24h rep1	Rat tail tendon	Rattus norvegicus	4307579061_B_Grn_Gras	Wistar	5 months	tail tendon	total RNA	biotin	replicate 1	GPL6101
12	Sample 2	Fascicle Unstrained 24h rep1	Rat tail tendon	Rattus norvegicus	4307579072_A_Grn.idat	Wistar	5 months	tail tendon	total RNA	biotin	replicate 1	GPL6101
13	Sample 3	Fascicle Strained 1h rep2	Rat tail tendon	Rattus norvegicus	4307579062_B_Grn.idat	Wistar	5 months	tail tendon	total RNA	biotin	replicate 2	GPL6101
14												
15	PROTOCOLS											
16	extract protocol	RNA was extracted with Trizol reagent, followed by clean-up and DNase I treatment with QIAGEN RNeasy mini kit in accordance with the prescribed protocol provided with the kit. Quality control was performed with Agilent Bioanalyser.										
17	label protocol	Biotinylated cRNA were prepared with the Ambion MessageAmp kit for Illumina arrays										
18	hyb protocol	Standard Illumina hybridization protocol										
19	scan protocol	Standard Illumina scanning protocol										
20	data processing	The data were normalised using quantile normalisation with IlluminaGUI in R										
21	value definition	quantile normalized										
22												
23												
24												
25												

Metadata Template Matrix normalized Matrix non-normalized Metadata Example Matrix normalized Example Matrix non-normalized Example +

Prêt 100 %



exemple : GSE25724

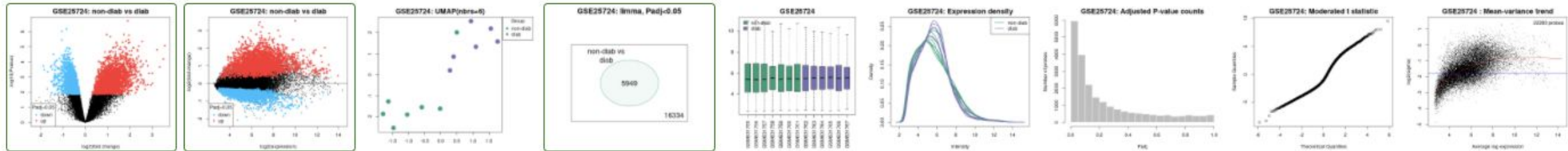
GEO accession Set Expression data from type 2 diabetic and non-diabetic isolated human islets

▼ Samples Define groups Selected 13 out of 13 samples

Group	Accession	Title	Source name	Tissue	Disease state	Age	Gender	Characteristics
non-diab	GSM631755	Non-diabetic islets, rep1	human islets, non-diabetic	pancreatic islets	non-diabetic	47 yrs	male	bmi (kg/m2): 27.7
non-diab	GSM631756	Non-diabetic islets, rep2	human islets, non-diabetic	pancreatic islets	non-diabetic	33 yrs	male	bmi (kg/m2): 22.9
non-diab	GSM631757	Non-diabetic islets, rep3	human islets, non-diabetic	pancreatic islets	non-diabetic	47 yrs	male	bmi (kg/m2): 28.4
non-diab	GSM631758	Non-diabetic islets, rep4	human islets, non-diabetic	pancreatic islets	non-diabetic	54 yrs	male	bmi (kg/m2): 23.1
non-diab	GSM631759	Non-diabetic islets, rep5	human islets, non-diabetic	pancreatic islets	non-diabetic	76 yrs	female	bmi (kg/m2): 25.9
non-diab	GSM631760	Non-diabetic islets, rep6	human islets, non-diabetic	pancreatic islets	non-diabetic	77 yrs	female	bmi (kg/m2): 23.8
non-diab	GSM631761	Non-diabetic islets, rep7	human islets, non-diabetic	pancreatic islets	non-diabetic	73 yrs	female	bmi (kg/m2): 22
diab	GSM631762	Type 2 diabetic islets, rep1	human islets, diabetic	pancreatic islets	type 2 diabetes	79 yrs	male	bmi (kg/m2): 27.5
diab	GSM631763	Type 2 diabetic islets, rep2	human islets, diabetic	pancreatic islets	type 2 diabetes	76 yrs	male	bmi (kg/m2): 26
diab	GSM631764	Type 2 diabetic islets, rep3	human islets, diabetic	pancreatic islets	type 2 diabetes	73 yrs	female	bmi (kg/m2): 29
diab	GSM631765	Type 2 diabetic islets, rep4	human islets, diabetic	pancreatic islets	type 2 diabetes	75 yrs	female	bmi (kg/m2): 26.5
diab	GSM631766	Type 2 diabetic islets, rep5	human islets, diabetic	pancreatic islets	type 2 diabetes	54 yrs	female	bmi (kg/m2): 23.9
diab	GSM631767	Type 2 diabetic islets, rep6	human islets, diabetic	pancreatic islets	type 2 diabetes	66 yrs	male	bmi (kg/m2): 23.1



Visualization ?





GISAID



Qui a déjà
soumis à
GISAID ?



C'était facile ?



Données de tous les virus de la grippe et du **coronavirus à l'origine du COVID-19** : séquence génétique et les données cliniques et épidémiologiques associées aux virus humains, ainsi que les données géographiques et spécifiques aux espèces associées aux virus aviaires et autres virus animaux, pour aider les chercheurs à comprendre comment les virus évoluent et se propagent pendant les épidémies et les pandémies.

GISAID le fait en surmontant les obstacles et les restrictions dissuasifs, qui découragent ou empêchent le partage des données virologiques avant la publication officielle.

L'Initiative garantit que le libre accès aux données de GISAID est fourni gratuitement à toutes les personnes qui ont accepté de **s'identifier et de respecter le mécanisme de partage de GISAID régi par son accord d'accès à la base de données.**



Fichier excel

20210222_EpiCoV_BulkUpload_Template.xls [Mode de compatibilité]

Accueil Insertion Dessin Mise en page Formules Données Révision Affichage

C44 x ✓ fx e.g. CLC Genomics Workbench 12, Geneious 10.2.4, SPAdes/MEGAHIT v1.2.9, UGENE v. 33, etc.

EpiCoV hCoV-19 bulk upload

Version: 2021-02-24

Instructions:

- Enter your data into the sheet "Submissions"
- The mandatory columns are indicated in color.
- Do not change the content of the two first rows (1 & 2)
- Delete, overwrite the examples given in row 3
- your sequences must be in one single FASTA-File to compliment this spreadsheet with your metadata
- EXCEL extension must remain .xls (not .xlsx). Always save in EXCEL 97 - 2003 Format.
- Provide for every row/virus the filename of the FASTA-File that contains the corresponding sequence.
- "FASTA Filename" must match exactly the actual filename without any directory prefixed. ("all_sequences.fasta" is OK, "c:/users/meier/docs/all_sequences.fasta" is not)
- FASTA-Headers in the .FASTA-File must exactly match the values of "Virus name" (e.g. >hCoV-19/Netherlands/Gelderland-01/2020)
- Do not change the type of the columns (Collection Date must be formatted as "text" not "date")
- Always use the newest bulk-upload-XLS-Template
- Use "unknown" written in lower case if no value is available
- The user should name the XLS-Sheet as follows prior sending to the curation team: "YYYYMMDD_a_descriptive_name_metadata.xls"

Upload your completed Excel sheet together with the FASTA-File through the Batch Upload interface

In the event you experience any difficulties with your upload, please contact us for assistance at hCoV-19@gisaid.org

What happens next?

EpiCoV Curators across different timezones will be alerted and review your data. Only if necessary, will you be contacted, before your data are released


You will receive an eMail alert informing you that your data has been released.

Column information		
Submitter	mandatory	enter your GISAID-Username
FASTA filename	mandatory	the filename that contains the sequence without path (e.g. all_sequences.fasta not c:/users/meier/docs/all_sequences.fasta)
Virus name	mandatory	e.g. hCoV-19/Netherlands/Gelderland-01/2020 (Must be FASTA-Header from the FASTA file all_sequences.fasta)
Type	mandatory	default must remain "betacoronavirus"
Passage details/history	mandatory	e.g. Original, Vero
Collection date	mandatory	Date in the format YYYY or YYYY-MM or YYYY-MM-DD
Location	mandatory	e.g. Europe / Germany / Bavaria / Munich
Additional location information		e.g. Cruise Ship, Convention, Live animal market
Host	mandatory	e.g. Human, Environment, Canine, Manis javanica, Rhinolophus affinis, etc
Additional host information		e.g. Patient infected while traveling in ...
Sampling Strategy		e.g. Sentinel surveillance (ILI), Sentinel surveillance (ARI), Sentinel surveillance (SARI), Non-sentinel-surveillance (hospital), Non-sentinel-surveillance (GP network), Longitudinal sampling on same patient(s), S gene dropout
Gender	mandatory	Male, Female, or unknown
Patient age	mandatory	e.g. 65 or 7 months, or unknown
Patient status	mandatory	e.g. Hospitalized, Released, Live, Deceased, or unknown
Specimen source		e.g. Sputum, Alveolar lavage fluid, Oro-pharyngeal swab, Blood, Tracheal swab, Urine, Stool, Cloacal swab, Organ, Feces, Other
Outbreak		Date, Location e.g. type of gathering, Family cluster, etc.
Last vaccinated		provide details if applicable
Treatment		include drug name, dosage
Sequencing technology	mandatory	e.g. Illumina MiSeq, Sanger, Nanopore MiniON, Ion Torrent, etc.
Assembly method		e.g. CLC Genomics Workbench 12, Geneious 10.2.4, SPAdes/MEGAHIT v1.2.9, UGENE v. 33, etc.
Coverage		e.g. 70x, 1,000x, 10,000x (average)
Originating lab	mandatory	Where the clinical specimen or virus isolate was first obtained
Address	mandatory	
Sample ID given by the originating laboratory		
Submitting lab	mandatory	Where sequence data have been generated and submitted to GISAID
Address	mandatory	
Sample ID given by the submitting laboratory		
Authors	mandatory	a comma separated list of Authors with complete First followed by Last Name
Comment	leave empty	do not use this column
Comment icon	leave empty	do not use this column

Instructions Submissions +

Prêt




© 2008 - 2021 | [Terms of Use](#) | [Privacy Notice](#) | [Contact](#)

You are logged in as **Thomas Denecker** - [logout](#)

Registered Users
EpiFlu™
EpiCoV™
My profile

EpiCoV™
 Search
 Downloads
 Upload

Single Upload

Enter and upload genetic sequence and metadata, available clinical and epidemiological data, geographical as well as species-specific data. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.

Virus detail

Virus name*

Accession ID

Type

Passage details/history*

Sample information

Collection date*

Location*

Additional location information

Host*

Additional host information

Outbreak Detail

Sampling strategy

Gender*

Patient age*

Patient status*

Specimen source

Last vaccinated

Treatment

Sequencing technology*

Assembly method

Coverage

Institute information

Originating lab*



GISAID © 2008 - 2021 | [Terms of Use](#) | [Privacy Notice](#) | [Contact](#)
You are logged in as **Helene Chiapello** - [logout](#)

Registered Users | EpiFlu™ | EpiCoV™ | My profile

EpiCoV™ | Search | Downloads | Upload

GISAID hCoV-19 Batch Upload

Upload genetic sequence as single FASTA-File and metadata, available clinical and epidemiological data, geographical as well as species-specific data as XLS or CSV. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.

Metadata as Excel or CSV*

max size: 5M [Choisir le fichier](#) aucun fichier sélé.

Sequences as FASTA*

max size: 32M [Choisir le fichier](#) aucun fichier sélé.

Confirmation options (Default) Notify me about ALL DETECTED FRAMESHIFTS in this submission for reconfirmation of affected sequences

Report upload XLS/CSV and FASTA.

[Download Instructions and Template](#) [Contact Curator](#) [Verify and Submit](#)

Important note: In the GISAID EpiFlu™ Database Access Agreement, you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be bound by the terms of the GISAID EpiFlu™ Database Access Agreement in respect of such data in the same manner as if they were data relating to influenza viruses.



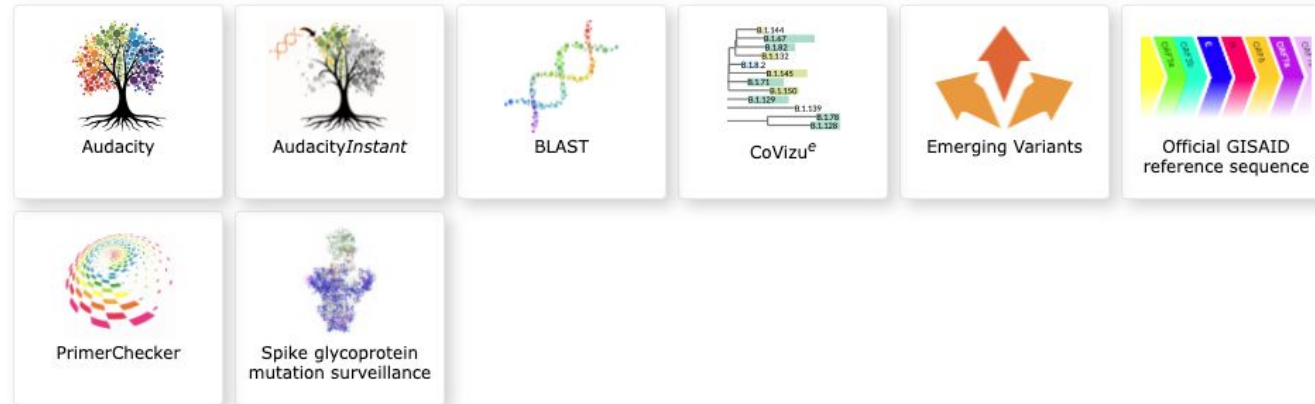
Version 2 Command Line Interface (CLI) for batch uploading

```
usage: cli2 upload [-h] [--database {EpiCoV,EpiFlu,EpiRSV}] [--token TOKEN] --metadata METADATA --fasta FASTA
                  [--frameshift {catch_all,catch_novel,catch_none}] [--failed FAILED] [--proxy PROXY] [--debug] [--log LOG]
```

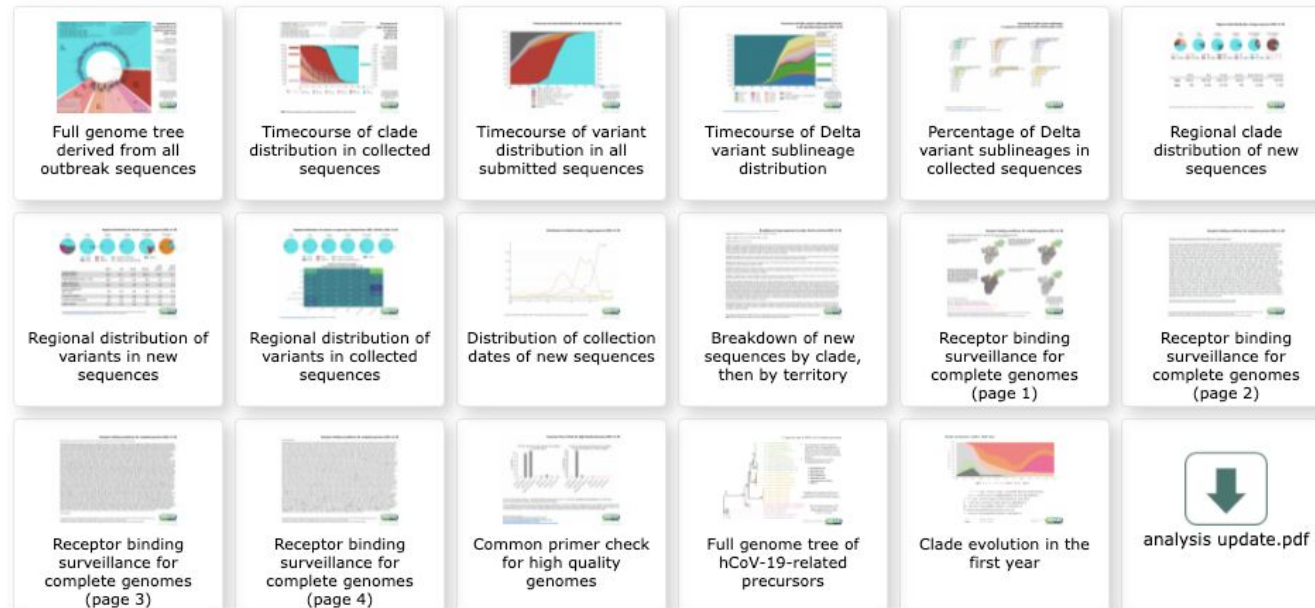
Perform upload of sequences and metadata to GISAID's curation zone.

optional arguments:

```
-h, --help          show this help message and exit
--database {EpiCoV,EpiFlu,EpiRSV}
                    Target GISAID database. (default: EpiCoV)
--token TOKEN       Authentication token. (default: ./gisaid.authtoken)
--metadata METADATA The csv-formatted metadata file. (default: None)
--fasta FASTA       The fasta-formatted nucleotide sequences file. (default: None)
--frameshift {catch_all,catch_novel,catch_none}
                    'catch_none': catch none of the frameshifts and release immediately; 'catch_all': catch all frameshifts and require email
                    confirmation; 'catch_novel': catch novel frameshifts and require email confirmation. (default: catch_all)
--failed FAILED     Name of CSV output to contain failed records. (default: ./failed.out)
--proxy PROXY       Proxy-configuration for HTTPS-Request in the form: http(s)://username:password@proxy:port. (default: None)
--debug             Switch off debugging information (dev purposes only). (default: True)
--log LOG           All output logged here. (default: ./upload.log)
```

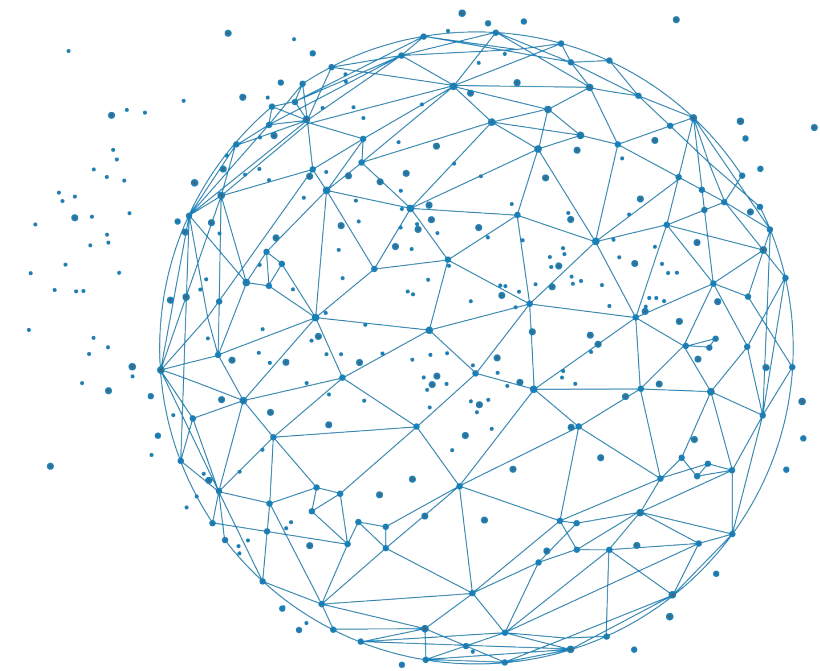


Analysis Update (2021-11-05)





Data brokering à l'IFB





Constat

- Les soumissions sont souvent complexes et difficiles à réaliser par les équipes expérimentales.
- Les métadonnées sont souvent mal comprises, ce qui entraîne des soumissions incomplètes, redondantes et incohérentes.

L'ENA a demandé à l'IFB de devenir le data broker français

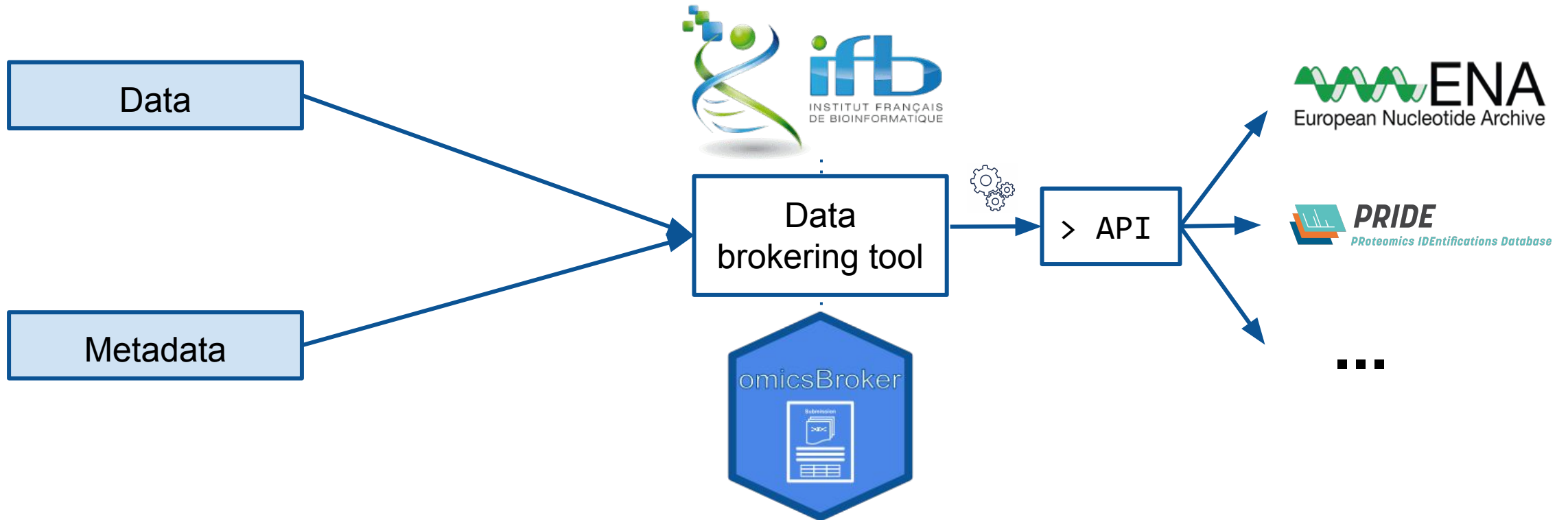
Idée principale : offrir un service national de data brokering à IFB pour **simplifier** et **rationaliser** les échanges de données entre les ressources internationales et le nœud Elixir français IFB.

3 types d'activités : le développement d'outils, la formation et le support aux utilisateurs.



IFB services to manage and centralize data and metadata of a project

IFB services to submit data and metadata of a project to international resources





omicsBroker is a tool to easily annotate and submit **omics data** to **international repositories**

Prototype disponible (soumission dans la zone de test de l'ENA)

- Développé en Django
- Disponible en Docker

Futurs développements

- Gestionnaire de soumission,
- API,
- ...



Metadata table

Excel

	Experience name	Organism	Platform	Instrument	Library layout	Insert size
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

Descriptions

Search

Platform

Definition

Platform name. Permitted values : <https://ena-docs.readthedocs.io/en/latest/submit/reads/webin-cli.html#permitted-values-for-platform>

Value

LS454 ; ILLUMINA ; PACBIO_SMRT ; ION_TORRENT ; CAPILLARY ; OXFORD_NANOPORE ; DNBSEQ

Harmonized Name

PLATFORM

* Mandatory



Le projet EMERGEN

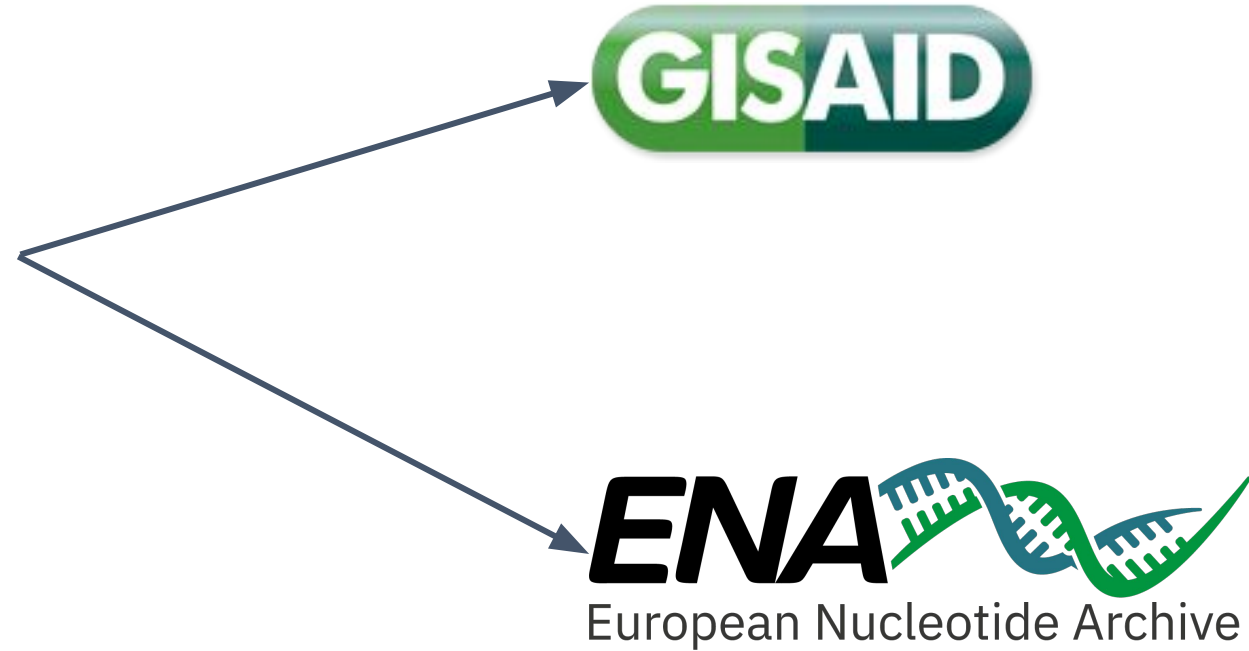
La pandémie de COVID-19 et l'émergence de variants du SARS-CoV-2 dont les caractéristiques de transmissibilité sont susceptibles de modifier la dynamique de l'épidémie en France ont souligné la nécessité de renforcer les capacités de surveillance génomique du SARS-CoV-2.

Cette surveillance a pour objectifs de détecter l'émergence et de suivre la distribution spatio-temporelle de virus présentant des mutations susceptibles d'avoir des conséquences fonctionnelles, comme par exemple l'infectiosité, la contagiosité, la virulence ou l'échappement immunitaire. Ces connaissances sont essentielles pour renforcer la maîtrise du risque infectieux en population et éclairer les décisions publiques.



EMERGEN-DB

The French COVID-19 database





Gestionnaire de soumissions

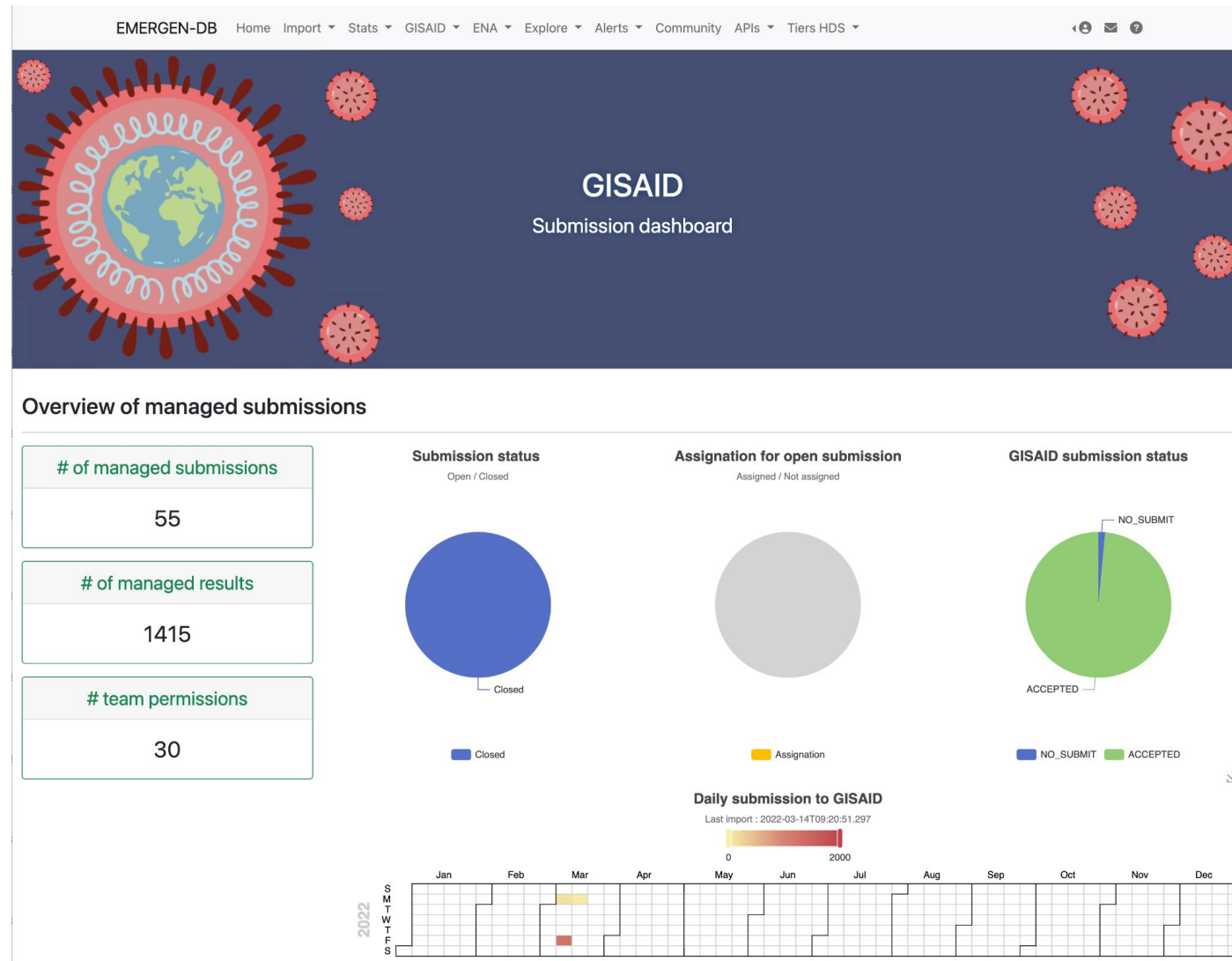
The screenshot shows the 'Submission to GISAID' page with the following sections:

- Data information:** A large empty box for data details.
- Submission information:** Database: GISAID, Created on: 11 mars 2022 11:14, To be submitted by the IFB?, Submission status: **Disputed**, and a calendar of activities.
- Metadata:** Summary cards for Total (46), To submit (42), No submit (2), Accepted (2), and History (1).
- Data brokering:** Assigned to: --Please choose an option--.
- Activity:** A timeline showing 'Modifications by IDeeacker'.

Outil de curation des données

The screenshot shows the 'Metadata to GISAID' page with the following sections:

- Download raw data:** Sequences (.fasta) and Metadata (.xlsx).
- GISAID metadata:** Download the displayed table.
- Filters:** A list of filters including To submit (45), Accepted (1), Already exists (0), Validation error (0), Upload error (0), Rejected (0), To revise (0), No submit (0), Submit by lab (0), and History (1).





<https://www.gfbio.org/>

COVID-19 is an emerging, rapidly evolving situation

- Get the latest public health information from CDC: <http://www.coronavirus.gov>
- Get the latest research information from NIH: <https://www.nih.gov/coronavirus>

[Learn to Publish COVID-19 Data to SRA](#)

METAGENOTE is a quick and intuitive way to annotate data from genomics studies including microbiome.

[Start Here!](#)

Why use METAGENOTE?



Annotate

Fully describe samples from which genomic sequences have been obtained.



Use Standards

Follow guidelines from the Genomics Standards Consortium (GSC) standards for ease of reproducibility.



Store & Search

Organize metadata into studies, projects and sample groups. Browse existing metadata.



Publish

Validate metadata and automatically publish to the NCBI Sequence Read Archive (SRA).