

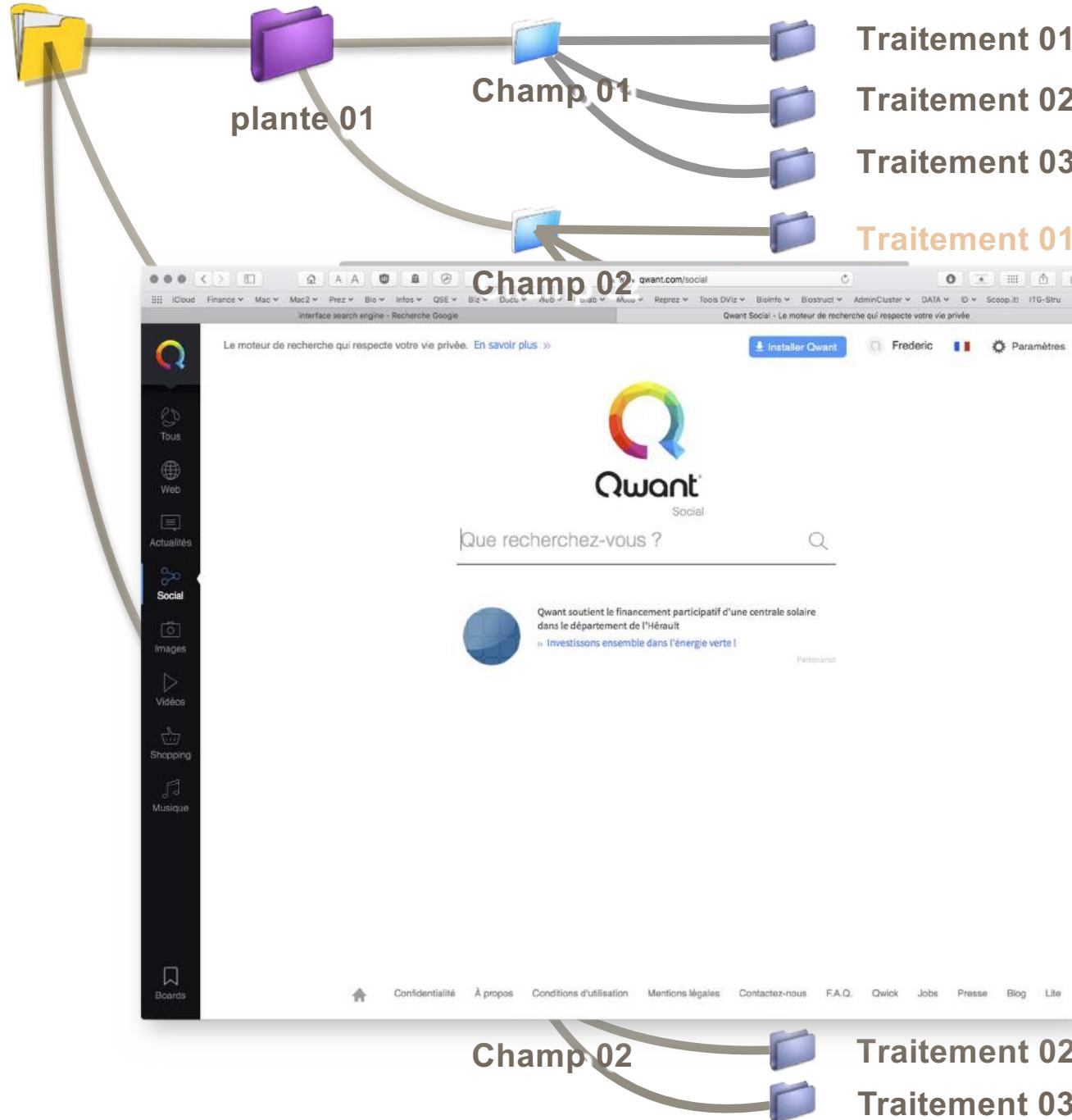
Introduction aux métadonnées

Valérie Cognat
&
Laurent Bouri



Pourquoi des métadonnées ?

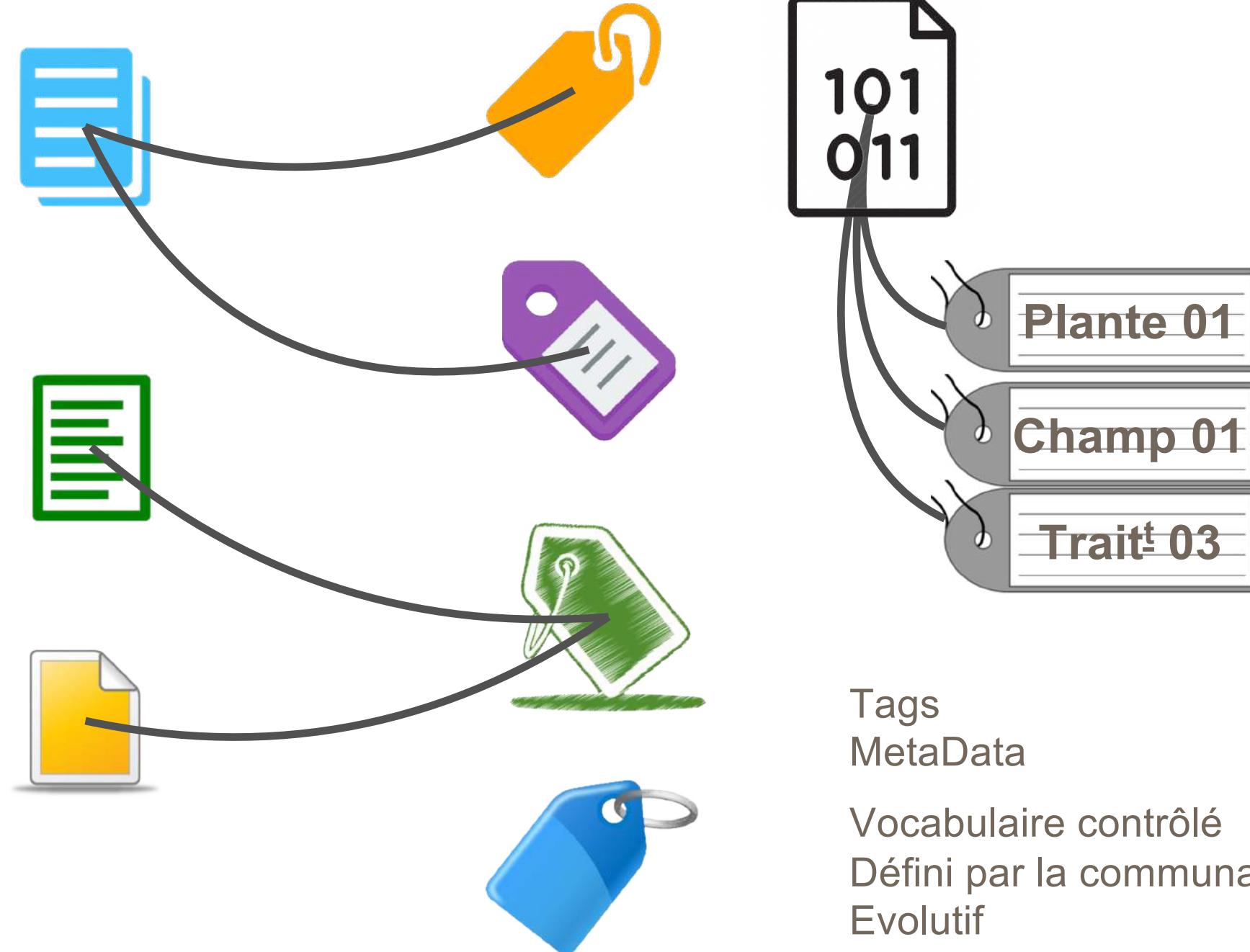
Définir une méthode commune et efficace pour retrouver nos données.

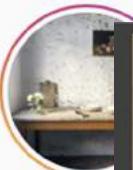


Les métadonnées : définition

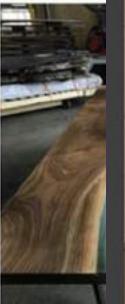
Les métadonnées







Meilleures Publi



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



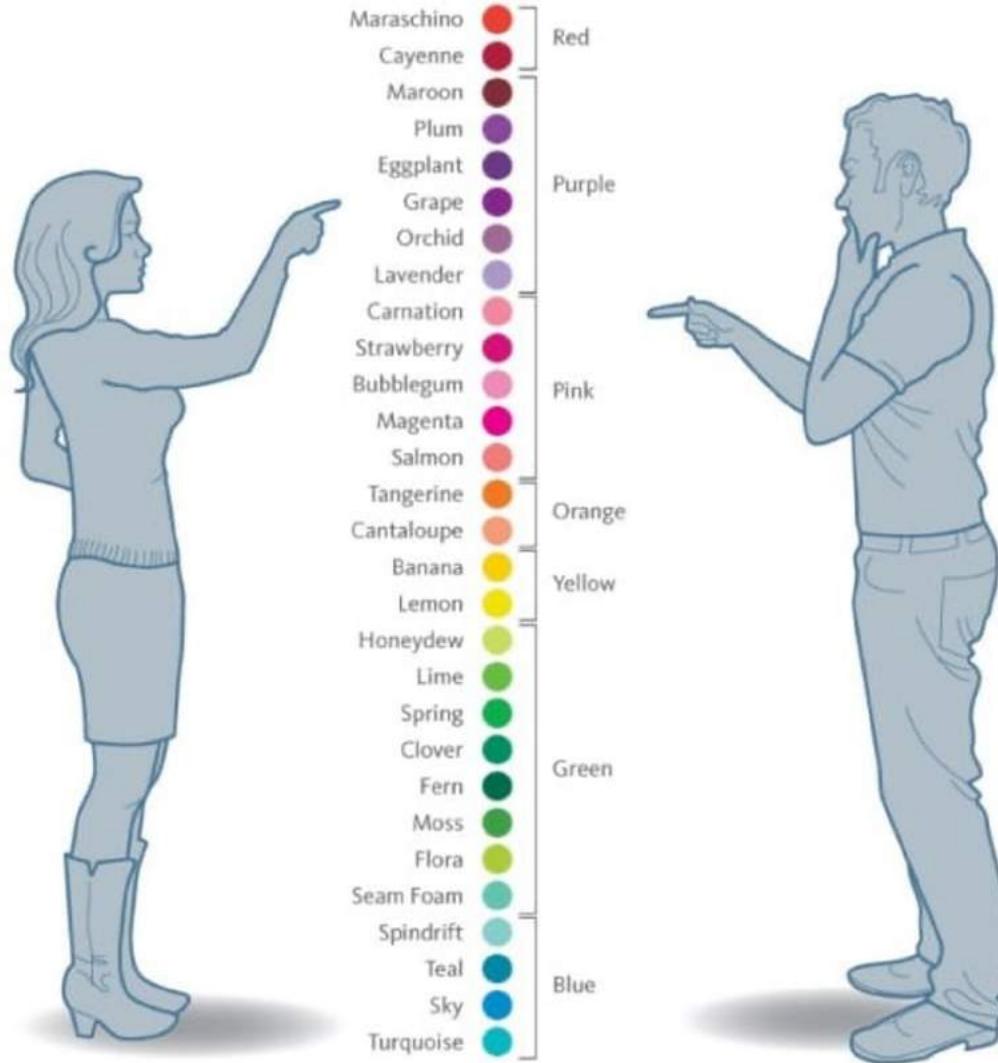
donnersteel, james.mcgregor.52,
stephaxil, noah4444, didsy_...
_edwardjones, markbulmer_photography,
rtedgar_boroondara, thundrgram et
beastmotivationfitness aiment ça.

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

Germplasme	Origine	Collection
AG0003	Guinea	prospection 1979
AG0004	Guinea	prospection 1979

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

Qu'est ce que les métadonnées

C'est du **data reporting** :

- WHO : Qui a créé les données
- WHAT : Quel est le contenu des données ?
- WHEN : Quand ont-elles été créées ?
- WHERE : Où ont-elles été créées ?
- HOW : Comment ont-elles été créées ?
- WHY : Pourquoi ont-elles été créées ?

Un standard de métadonnées

RECOMMANDATIONS TECHNIQUES POUR LES MÉTADONNÉES ET STANDARDS

VERSION N°1 – 2017

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 entre 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.



Standards de métadonnées en Sciences de la Vie

Valérie Cognat, IBMP, Strasbourg
<https://orcid.org/0000-0001-9337-2767>

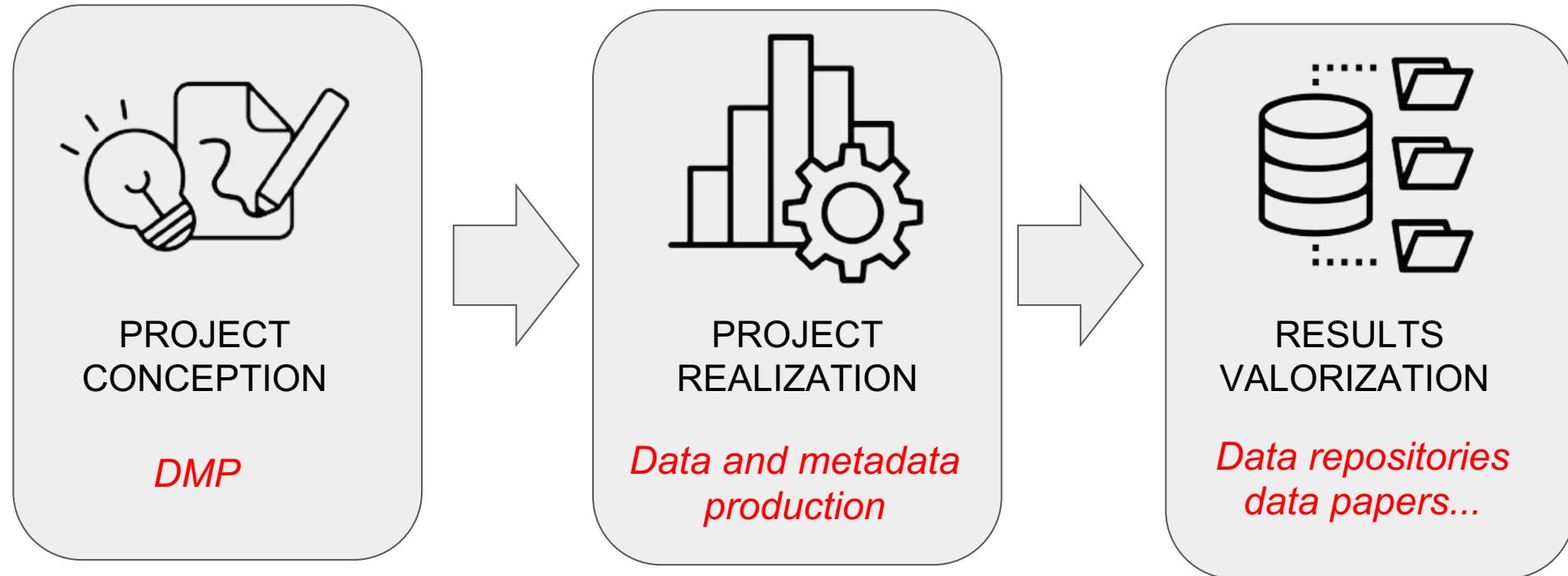
&

Laurent Bouri, IFB, IGBMC Strasbourg
<https://orcid.org/0000-0002-2297-1559>

Hélène Chiapello, IFB, INRAE Jouy-en-Josas
<https://orcid.org/0000-0001-5102-0632>

Thomas Denecker, IFB, CNRS Paris
<https://orcid.org/0000-0003-1421-7641>

Metadata during the project



Metadata concern all steps of a scientific project !

How do I produce metadata?

How do you describe the data?

With a set of metadata

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

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

With a **metadata standard**

Disciplinary standard



General standard



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

Definition of a 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/>

Example of standard in biology : Gene Ontology

The standards concern both data and metadata

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**

Generic and specific standards for metadata

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)

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>

Bioschemas profile specifications for “Gene”



- *Green* properties/types are proposed by Bioschemas, or indicate proposed changes by Bioschemas to Schema.org
- *Red* properties/types exist in the core of Schema.org
- *Blue* properties/types exist in the pending area of Schema.org
- *Black* properties/types are reused from external vocabularies/ontologies

CD = Cardinality

Examples

Property	Expected Type	Description	CD	Controlled Vocabulary	Example
Marginality: Minimum.					
@context	URL	Used to provide the context (namespaces) for the JSON-LD file. Not needed in other serialisations.	ONE		
@type	Text	Schema.org/Bioschemas class for the resource declared using JSON-LD syntax. For other serialisations please use the appropriate mechanism. While it is permissible to provide multiple types, it is preferred to use a single type.	MANY	Schema.org, Bioschemas	
@id	IRI	Used to distinguish the resource being described in JSON-LD. For other serialisations use the appropriate approach.	ONE		
dct:conformsTo	IRI	Used to state the Bioschemas profile that the markup relates to. The versioned URL of the profile must be used. Note that we use a CURIE in the table here but the full URL for Dublin Core terms must be used in the markup (http://purl.org/dc/terms/conformsTo), see example.	ONE	Bioschemas profile versioned URL	
identifier	PropertyValue Text URL	Schema: The identifier property represents any kind of identifier for any kind of Thing, such as ISBNs, GTIN codes, UUIDs etc. Schema.org provides dedicated properties for representing many of these, either as textual strings or as URL (URI) links. See background notes for more details.	ONE		
name	Text	Schema: The name of the item.	ONE		
Marginality: Recommended.					
description	Text	Schema: A description of the item.	ONE		
encodesBioChemEntity	BioChemEntity	Schema: The BioChemEntity to be encoded by this component or associated to.	MANY		

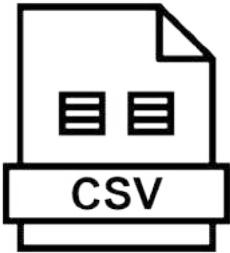
```
//Property: alternateName
{
  "@type": "Gene",
  "alternateName": "AD1"
}

//Property: description
{
  "@type": "Gene",
  "description": "amyloid beta precursor"
}

//Property: encodesBioChemEntity
{
  "@type": "Gene",
  "encodesBioChemEntity": {
    "@type": "Protein",
    "identifier": "uniprotkb:P05067",
    "url": "https://www.uniprot.org/uniprot/P05067"
  }
}
```

Json-LD

Three text formats frequently used for metadata



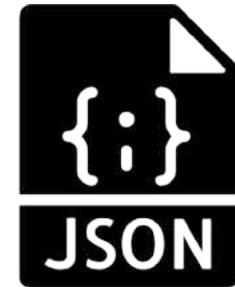
Comma Separated Values

```
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>
```

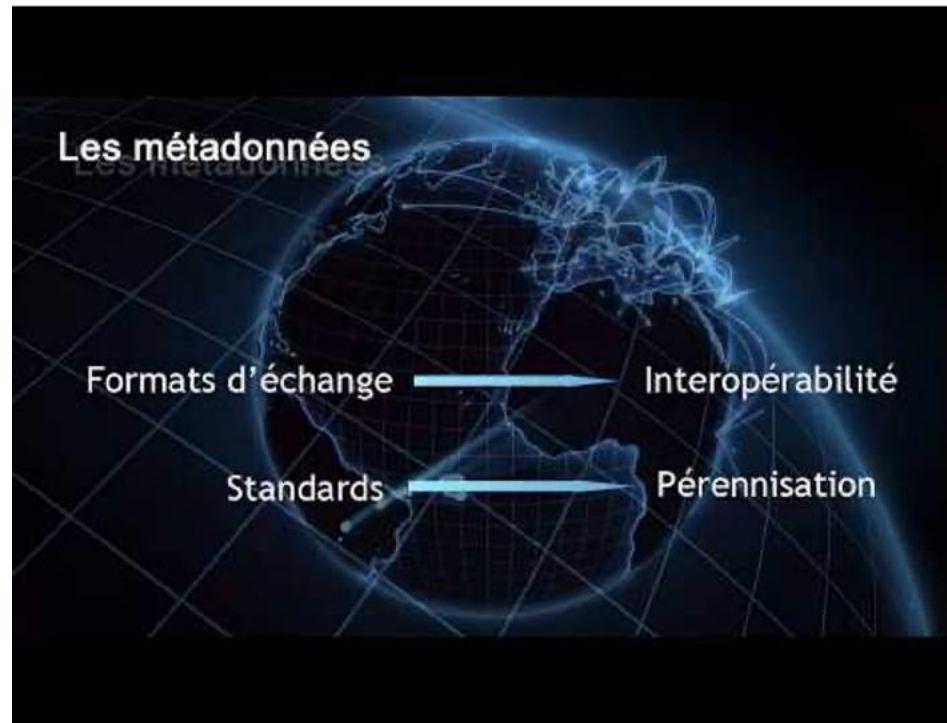


**JavaScript Object
Notation**

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

Vidéo: la minute métadonnées:

<https://doranum.fr/metadonnees-standards-formats/>

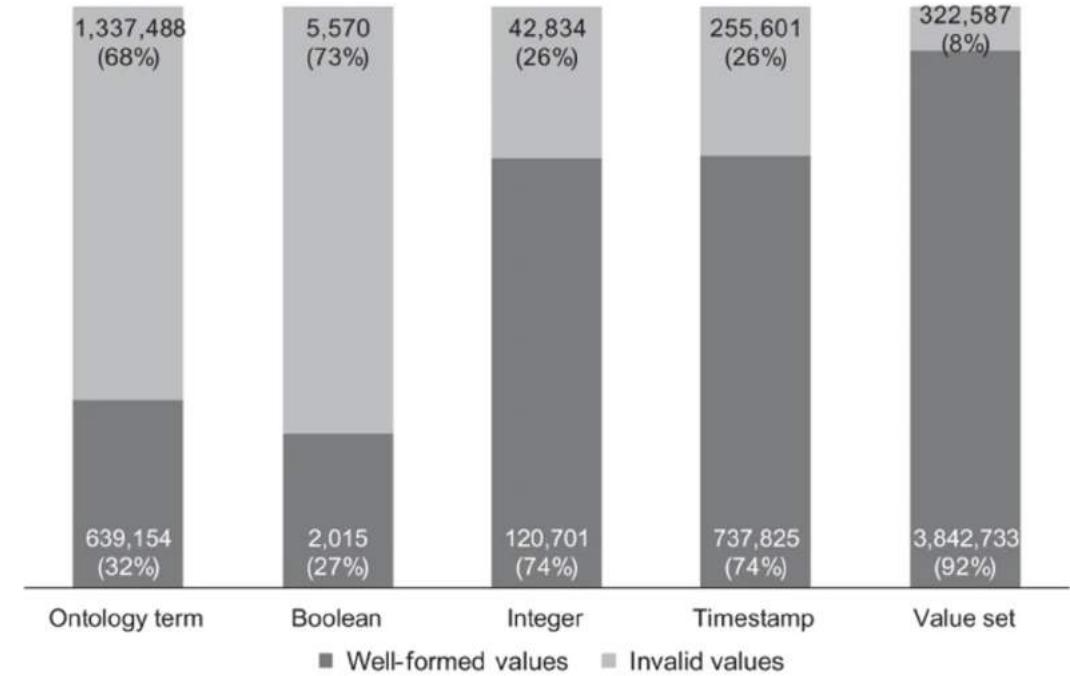


Metadata exhibit questionable quality in biology

Submission in public resources is often a complex task

Submission procedures are heterogeneous

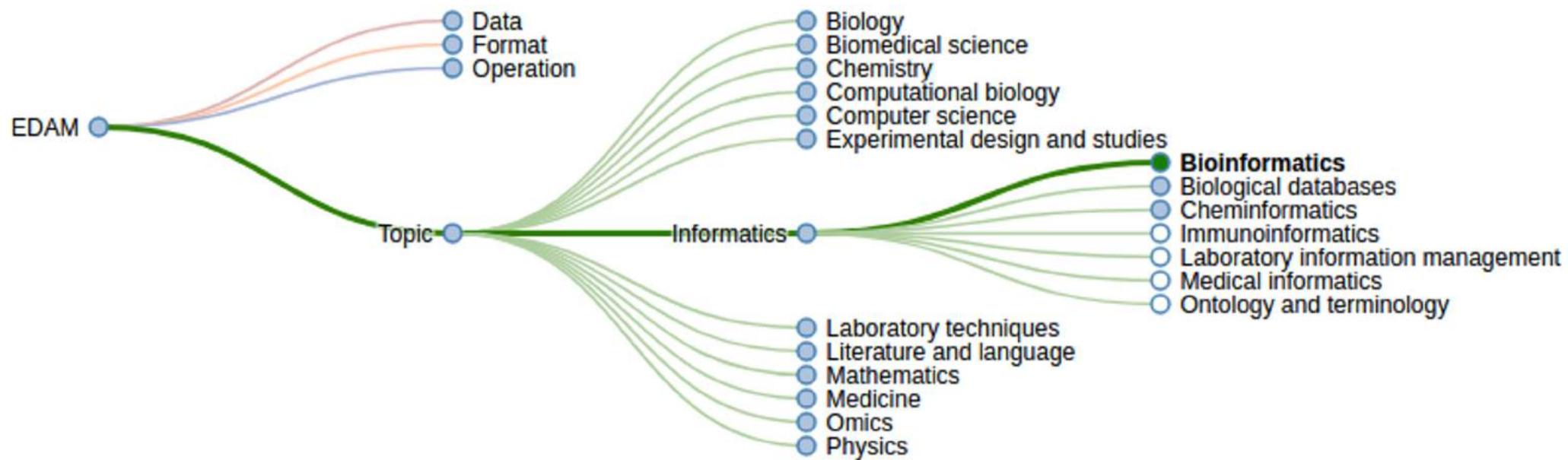
Metadata are often incomplete, inconsistent, redundant or not informative enough



Quality of dictionary attributes in NCBI BioSample according to their type, in Gonçalves et al., 2019

Ontology of bioscientific data analysis and data management

EDAM: a simple ontology of well established, familiar concepts that are prevalent within bioinformatics



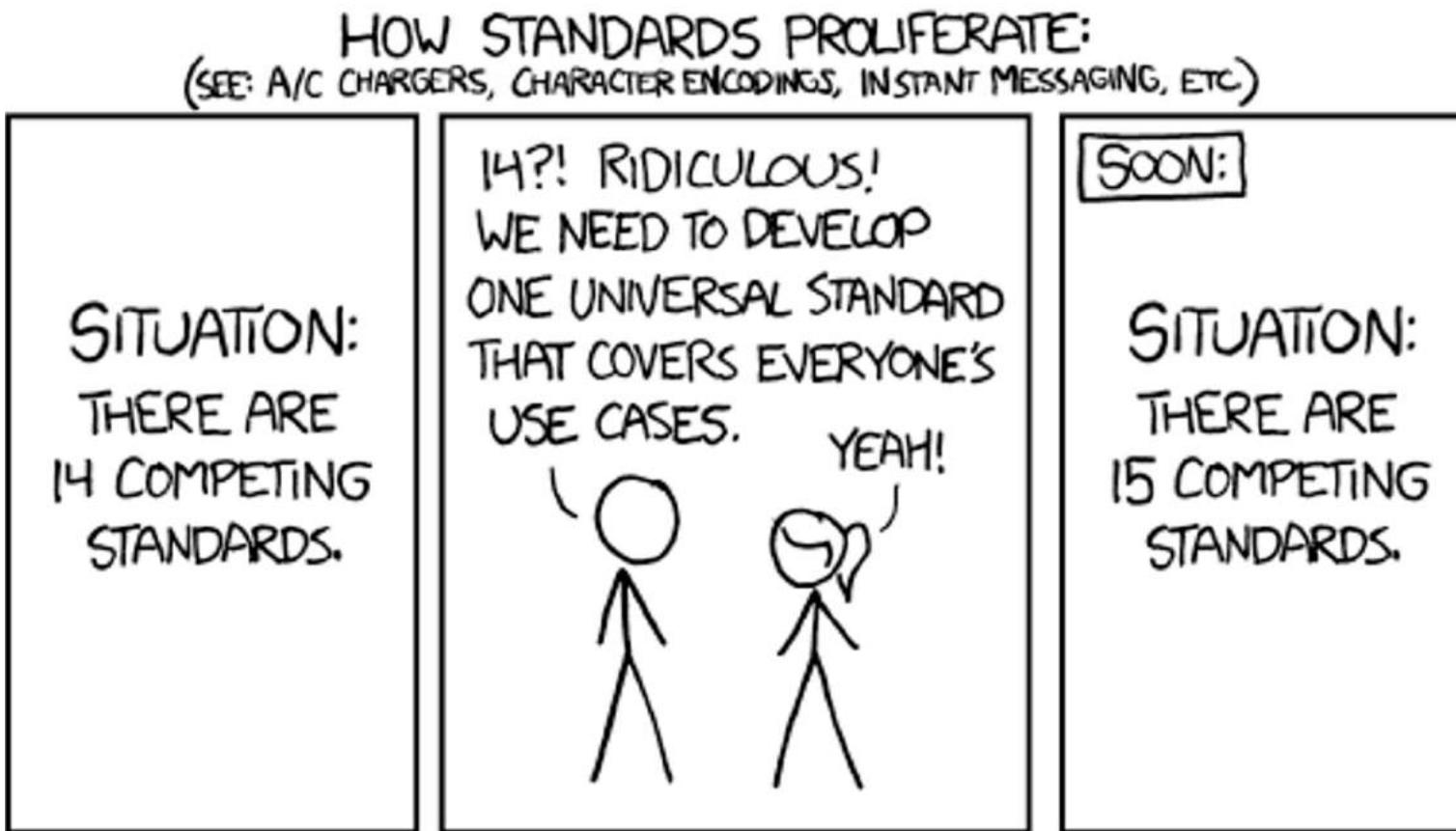
Standard adoption and perennity

- There are thousand of databases, softwares and resources in biology with an **unequal level of standard adoption**
- Is 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](#)

Standard adoption and perennity



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

How do I find the standard I need?

The FAIRsharing portal

Sansone, et al. FAIRsharing as a community approach to standards, repositories and policies.

Nat Biotech. 2019

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



FAIRsharing.org
standards, databases, policies

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 Identify 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... [read more](#)

 1582 Standards

Terminology Artifact	830
Model/Format	504
Reporting Guideline	228
Identifier Schema	20

[VIEW ALL](#)

 1861 Databases

Repositories	956
Knowledgebases	788
Knowledgebase/Repositories	117

[VIEW ALL](#)

 149 Policies

Journal	88
Funder	23
Society	14
Project	13

[VIEW ALL](#)

<https://fairsharing.org>

The FAIRsharing portal

Citable DOI for all records

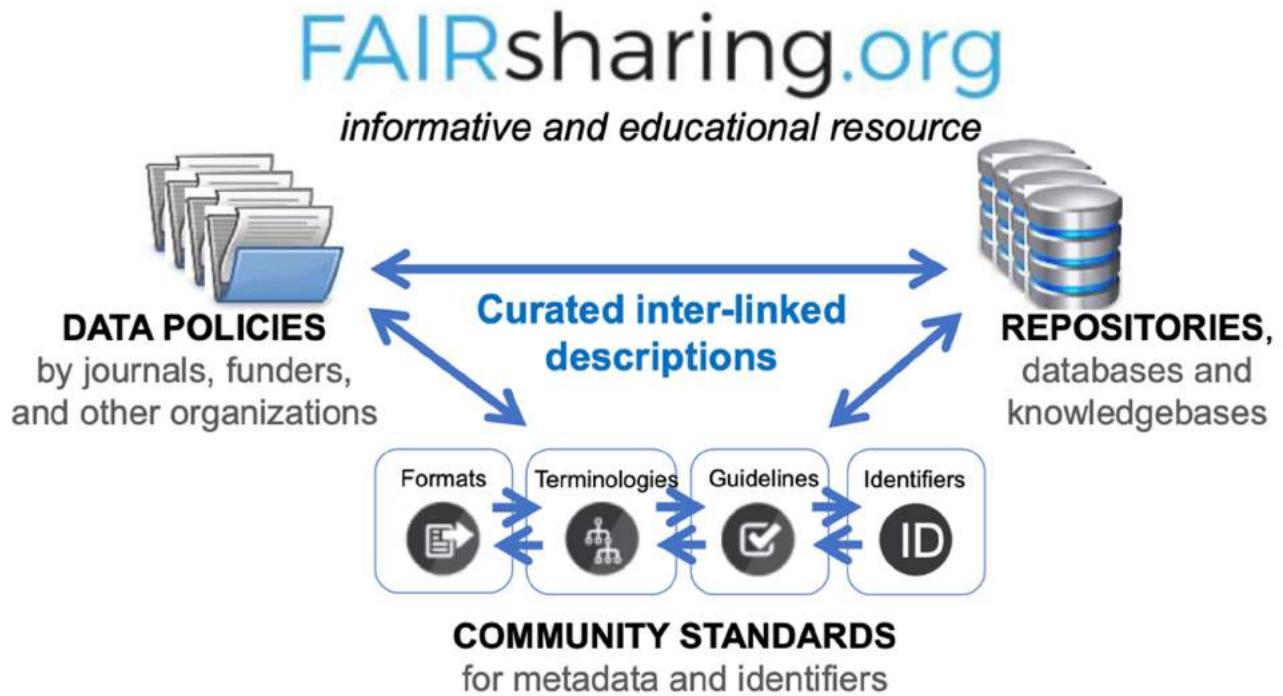
Accessible via API or web interface

Curation

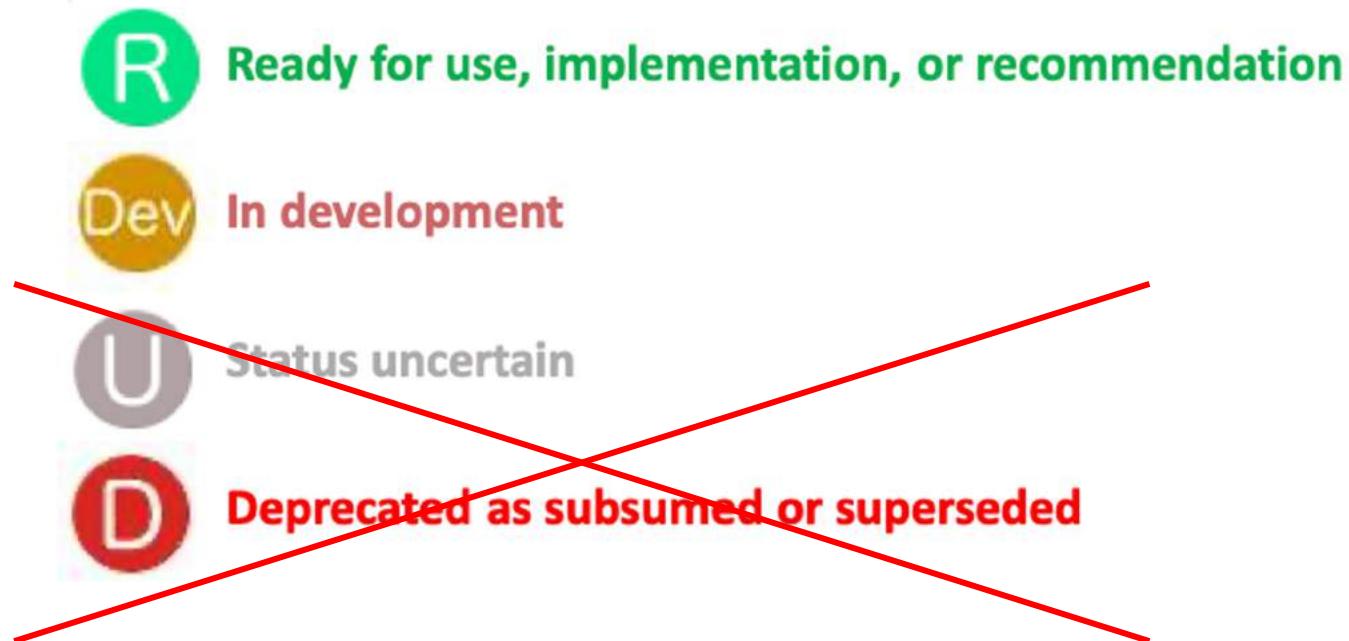
RECORD STATUS



<https://fairsharing.org>



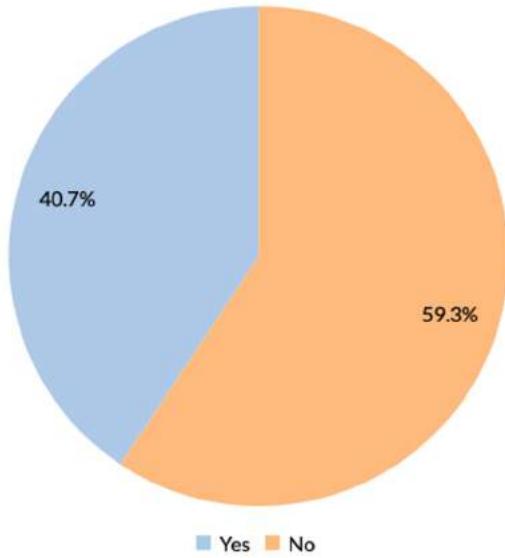
The FAIRsharing portal: record status



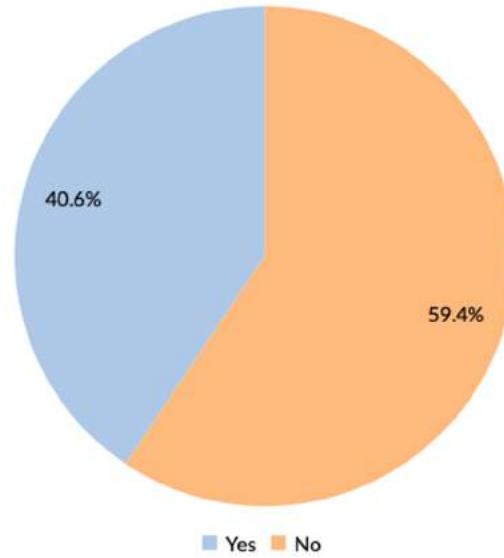
Please don't use “Uncertain” or “Deprecated” standards

Standard maintenance is a key point

Standard records that have maintainers



Standards that have a publication



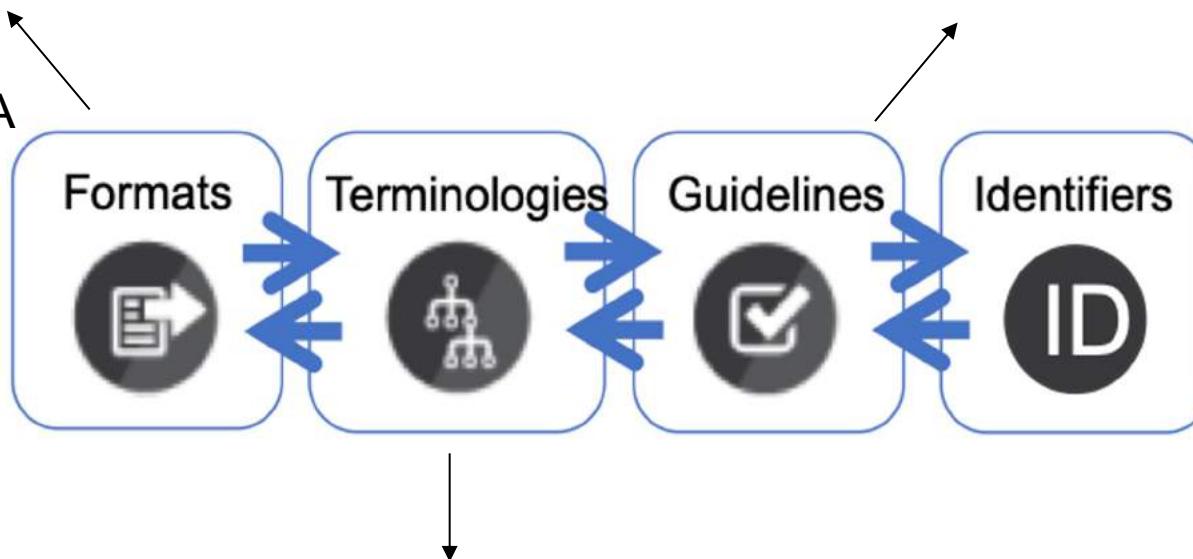
59.3 % of standards have no maintainer

59.4% of standard has no publication

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

Types of data standards

Conceptual model, schema, exchange formats, etc...
e.g. SBML, FASTA

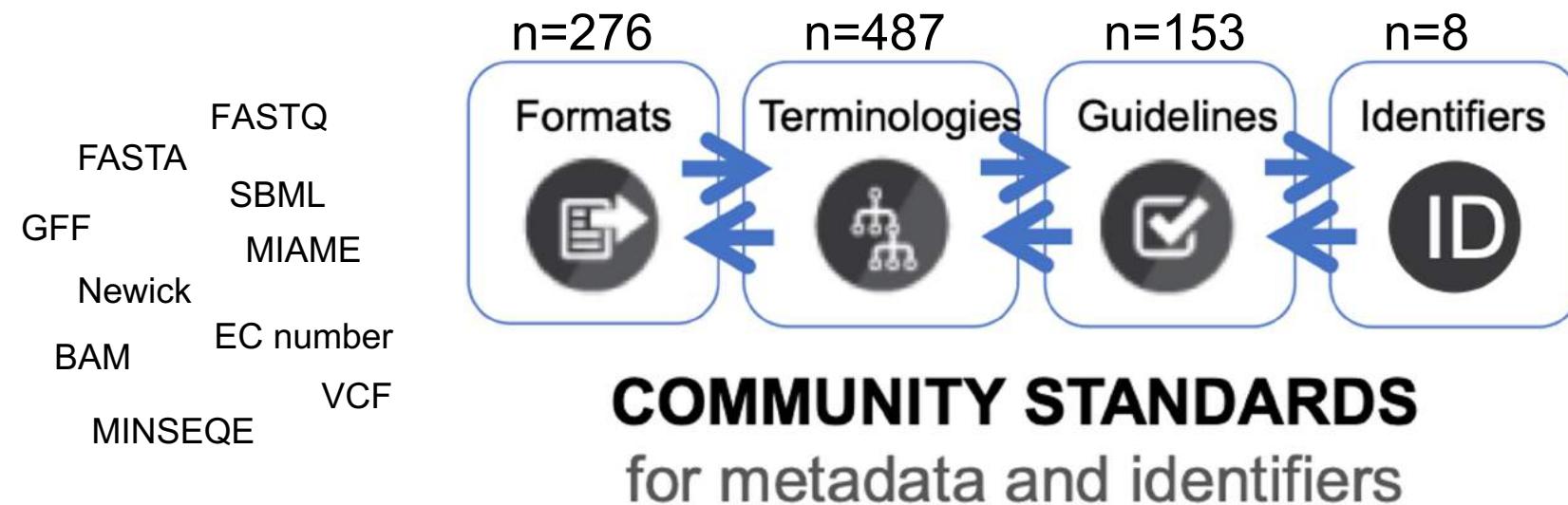


Minimum information reporting requirements, checklists...
e.g. MIAME guidelines

Controlled vocabularies, taxonomies, ontologies...
e.g. Gene Ontology

Formal systems for resources and digital objects that allow their identification
e.g. DOI

The landscape of standards in life sciences



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



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

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

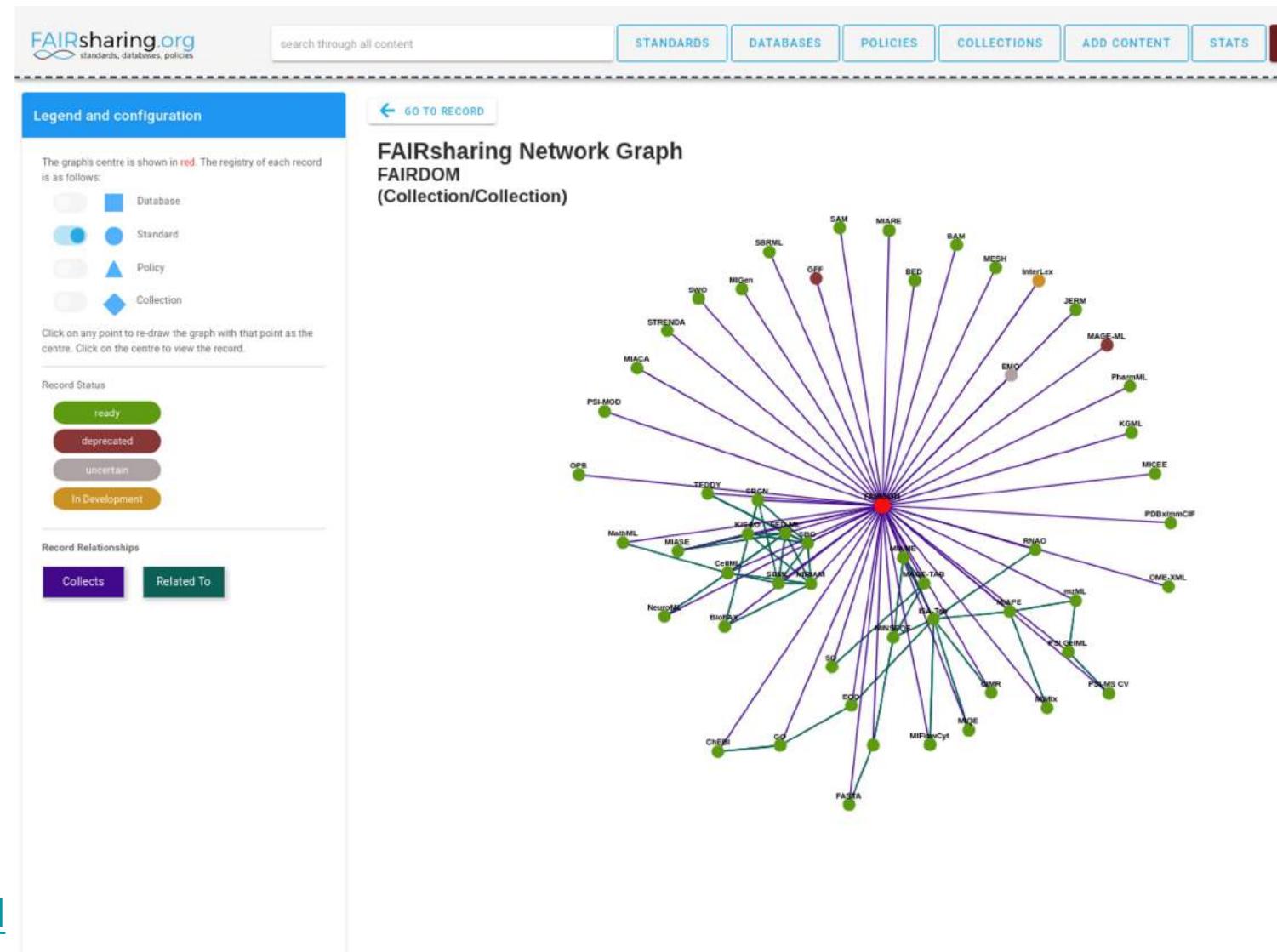
The screenshot shows the FAIRsharing.org interface for collections. At the top, there's a navigation bar with links for STANDARDS, DATABASES, POLICIES, COLLECTIONS (which is highlighted in blue), ADD CONTENT, STATS, and LOGIN. Below the header, a banner titled "Collections" defines what they are: "Collections group together one or more types of resource (standard, database or policy) by domain, project or organisation. A Recommendation is a core-set of resources that are selected or endorsed by data policies from journals, funders or other organisations." The main content area has a search bar at the top left and several filter categories on the left side: Registry, Record Type, Subjects, Domains, and Licence(s). On the right, there are two collection cards. The first card is for "NIH-supported data repositories", which lists NIH-supported data repositories that make data accessible for reuse. It includes a summary, a "Related Standards" section (with 6 items), and a "Related Databases" section (with 21 items). The second card is for the "COVID-19 Rapid Review Initiative" (C19RR), which brings together publishers and scholarly communications organisations to ensure research related to COVID-19 is made available. It also includes a summary, a "Related Standards" section (with 6 items), and a "Related Databases" section (with 14 items). Both cards feature a small circular logo with a gear and the letter 'R'.

Collections in Life Sciences

52 collections related to
Life Science standards in
FAIRsharing

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

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



Some collections are recent

Example 2: The *Covid-19* collection

FAIRsharing.org standards, databases, policies search through all content STANDARDS DATABASES POLICIES COLLECTIONS ADD CONTENT STATS LOGIN

ACTIONS ▾

GENERAL INFORMATION

 R COVID-19 Resources

Type Collection

Registry Collection

Description This is a draft collection that contains databases (including knowledgebases and repositories) and standards that are responding to, or appropriate for, use in the COVID-19 pandemic. These resources may be focused on patient response, clinical trials, virology studies or other related areas and, with respect to the databases in this Collection, contain COVID-19 related datasets. Please contact contact@fairsharing.org if you know of resources that should be added.

Organisations Not applicable

Homepage <https://fairsharing.org/>

Reference URL Not Available

Maintainers #AllSharingTeam

Contacts None

Subjects Clinical Studies, Public Health, Health Sciences, Global Health, Virology, Biomedical Sciences, Epidemiology, Immunological Studies

Domains Infection, Disease, Cardiovascular Disease

Taxonomic Range Coronaviridae, Homo sapiens, Virus

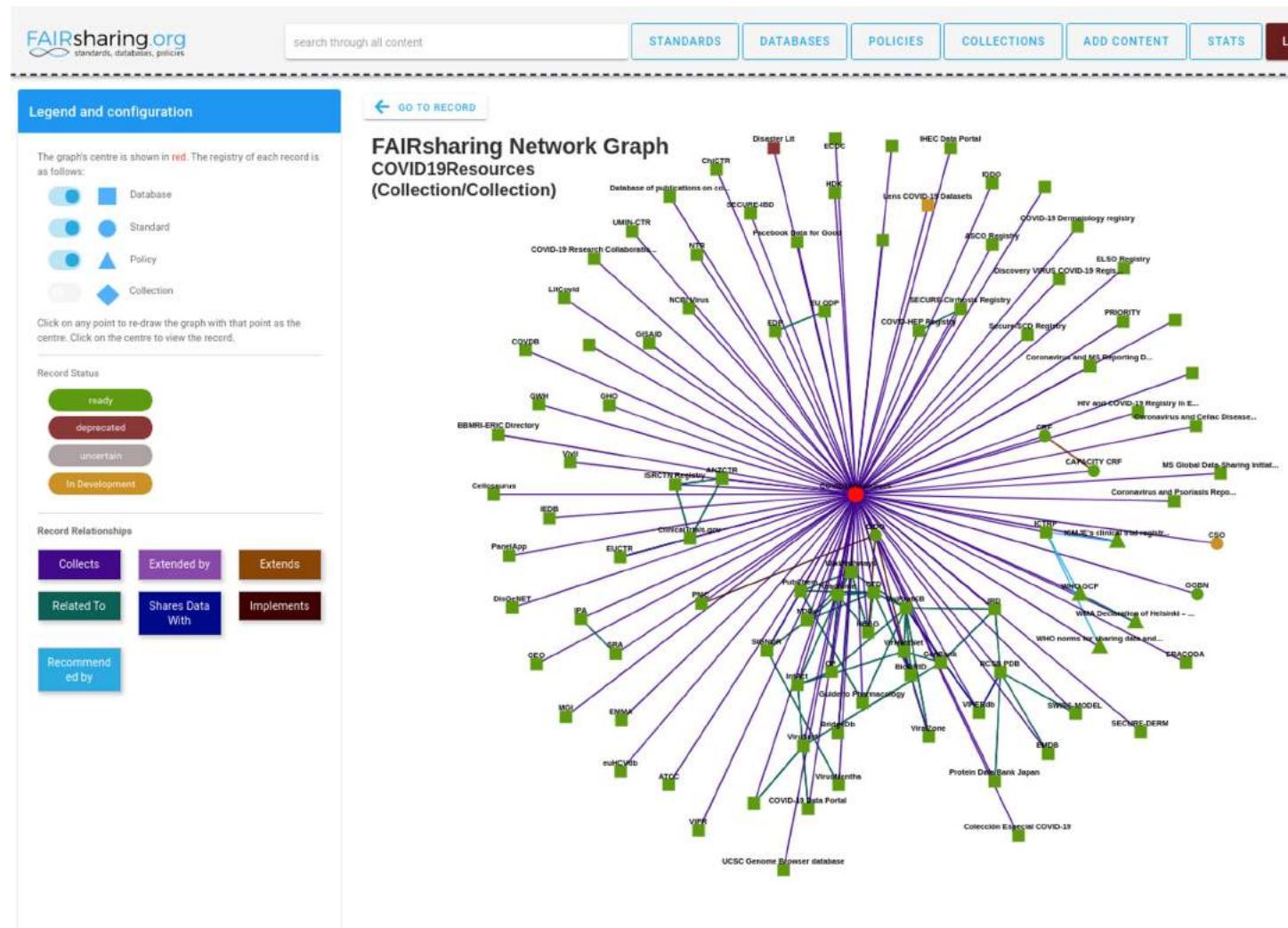
User Defined Tags Respiratory Disease

 [VIEW RELATION GRAPH](#)

 How to cite this record

FAIRsharing.org: COVID19resources; COVID-19 Resources; FAIRsharing ID: <https://fairsharing.org/5530>; Last Edited: Wednesday, January 5th 2022, 12:05; Last Editor: alisonstevie; Last Accessed: Friday, March 11th 2022, 13:38

Its Record created at Friday, April 3rd 2020, 18:01 | Record updated at Wednesday, January 5th 2022, 12:05



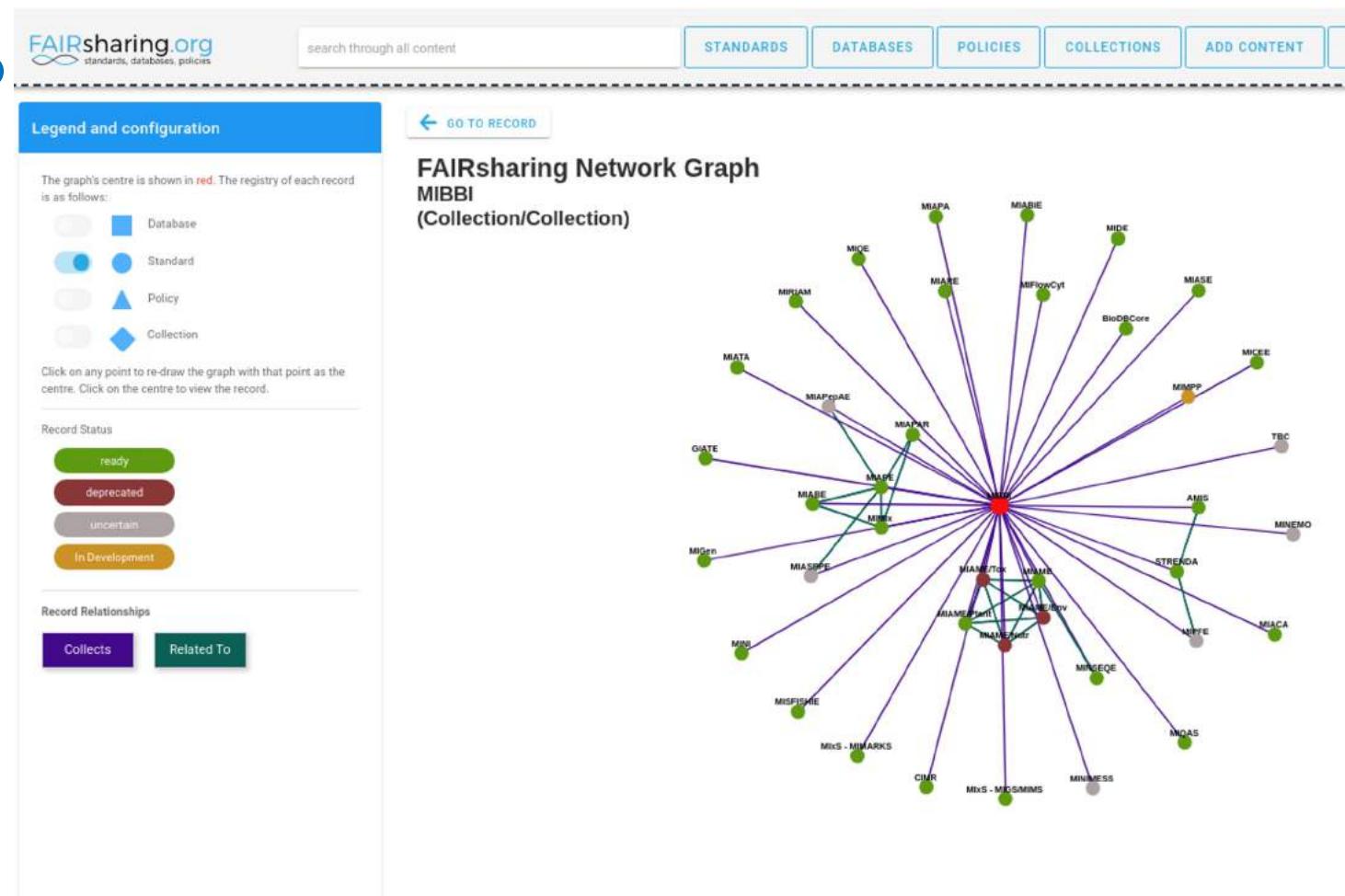
<https://fairsharing.org/collection/COVID19Resources>

<https://fairsharing.org/graph/3538>

What about the minimum required metadata in biology?

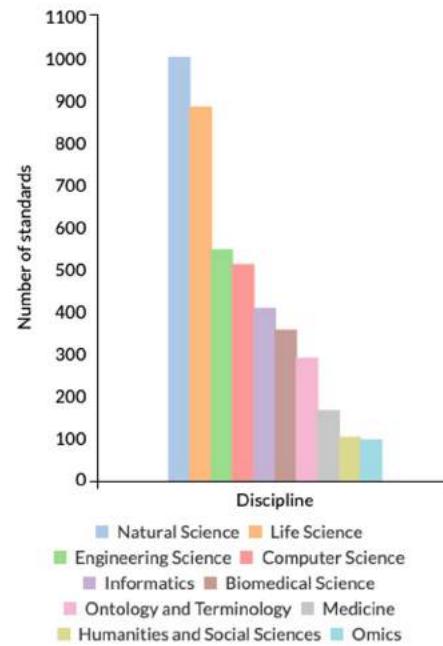
Example 3: the *Minimum Information for Biological and Biomedical Investigations* collection

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

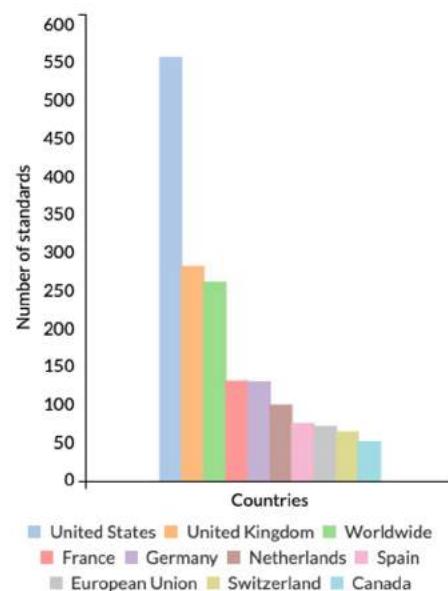


Summary statistics about standards

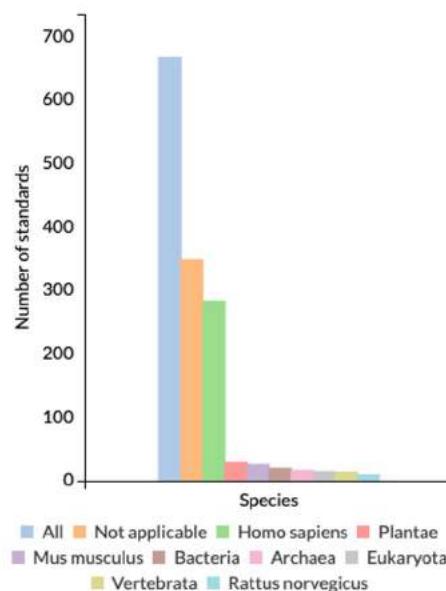
Top 10 disciplines covered by standards



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>

<https://fairsharing.org/browse/subject>

Practice

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 ?
2. What type of standard is *Minimum Information about any (x) Sequence (MiXS)* ?
3. What is the record status of the GAZ record ?

Practice => Answers

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

The Genomic Standards Consortium (GSC)

FAIRsharing.org search through all content

ACTIONS ▾

GENERAL INFORMATION

 Genomic Standards Consortium

Type Collection

Registry Collection

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.

Organisations Not applicable

Homepage <http://genec.org>

Reference URL <http://genec.org/projects/>

Maintainers rwallis (ip)

Contacts None

Subjects Genomics

Domains Genome

Taxonomic Range All

User Defined Tags None

 VIEW RELATION GRAPH

How to cite this record

FAIRsharing.org: GSC; Genomic Standards Consortium, FAIRsharing ID: <https://fairsharing.org/3528>. Last Accessed: Friday, March 11th 2022, 14:02

It's Record created at Tuesday, October 24th 2017, 14:07 | Record updated at Wednesday, November 24th 2021, 14:17

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

FAIRsharing.org search through all content STANDARDS DATABASES POLICIES

Legend and configuration

The graph's centre is shown in red. The registry of each record is as follows:

- Database
- Standard
- Policy
- Collection

Click on any point to re-draw the graph with that point as the centre. Click on the centre to view the record.

GO TO RECORD

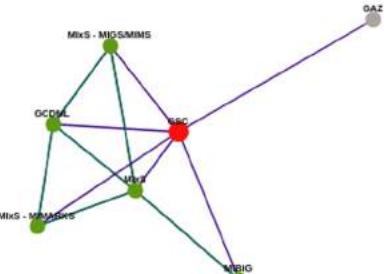
FAIRsharing Network Graph
GSC
(Collection/Collection)

Record Status

- ready
- deprecated
- uncertain
- In Development

Record Relationships

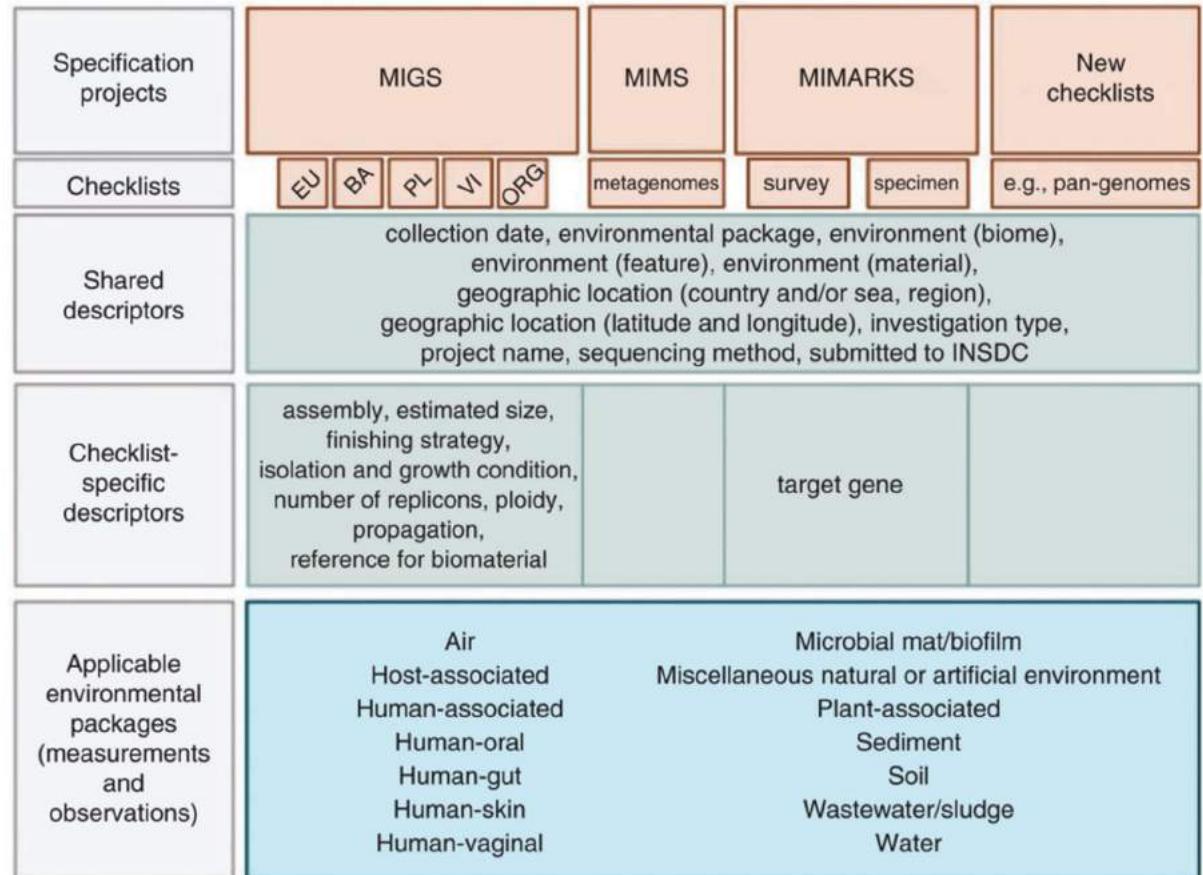
Collects Related To



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

The Genomic Standards Consortium (GSC)

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



Yilmaz et al, 2011

Source: <https://gensc.org>

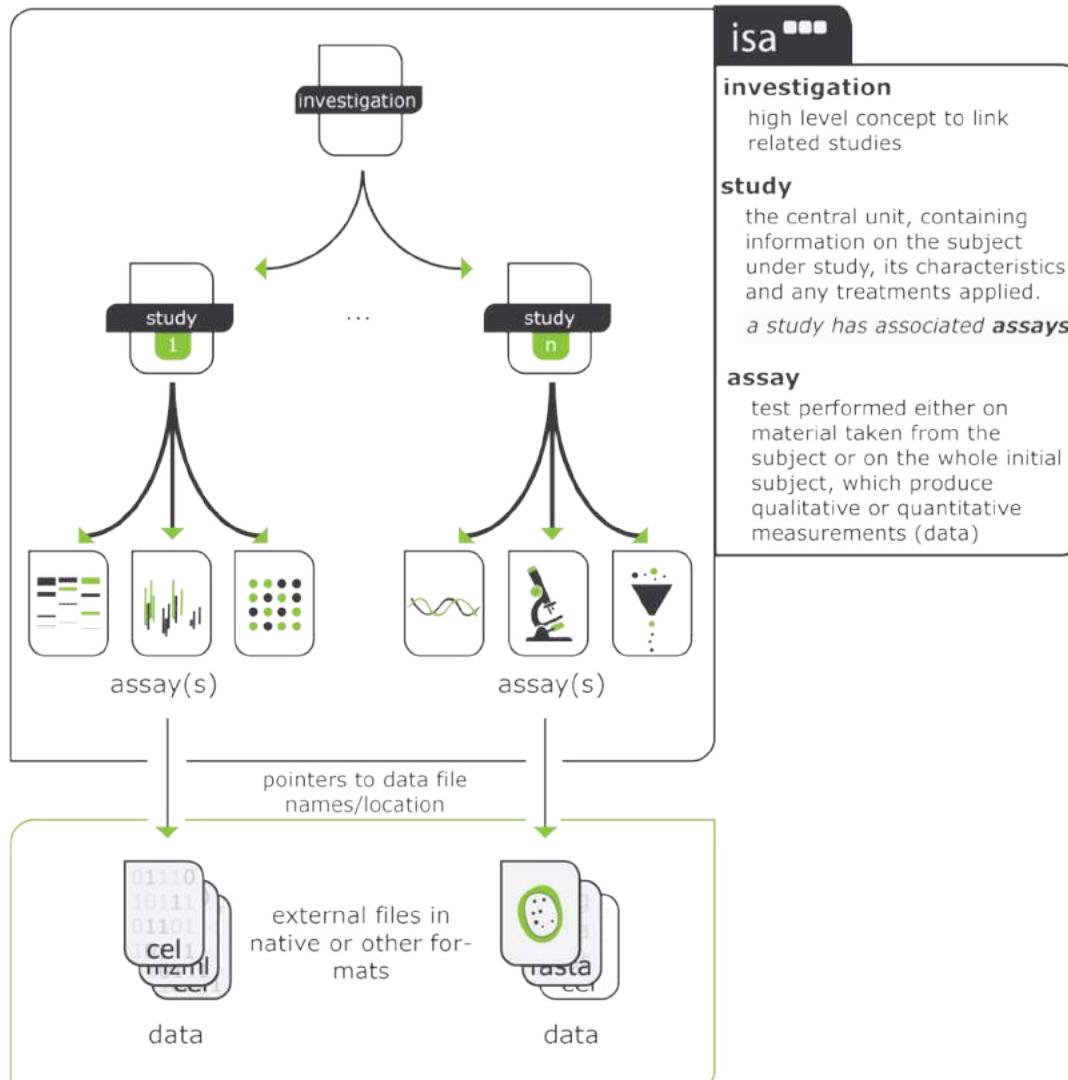
The ISA model

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>

RDMkit

The ELIXIR Research Data Management Kit (RDMkit) is an online guide containing **good data management practices** applicable to research projects from the beginning to the end.

RDMkit

Data management About Contribute GitHub Search RDMkit

Data management

- Data life cycle
- Your role
- Your domain
- Your tasks
- Tool assembly
- National resources
- All tools and resources
- All training resources

Are you working with data in the Life Sciences? Do you feel overwhelmed when you think about Research Data Management?

The ELIXIR Research Data Management Kit (RDMkit) is an online guide containing good data management practices applicable to research projects from the beginning to the end. Developed and managed by people who work every day with life science data, the RDMkit has guidelines, information, and pointers to help you with problems throughout the data's life cycle. RDMkit supports FAIR data — Findable, Accessible, Interoperable and Reusable — by-design, from the first steps of data management planning to the final steps of depositing data in public archives.

The RDMkit organises information into the six sections displayed below, which are interconnected but can be browsed independently.

Data life cycle

Start here to get an overview of research data management. Click on a section of the diagram below to get an introduction to that stage of the data management life cycle.

Your role

Identify your role in research data management, find data management resources relevant for you, and information to help you progress in your career path.

Show pages ▾

Your domain

Learn about the data management problems that affect your domain or research community, and the solutions adopted to address them.

Show pages ▾

Your tasks

Find guidelines and solutions for tackling common data management problems.

Tool assembly

Find concrete combinations of tools and resources assembled into a workflow for research data

To conclude: sources & useful links

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) ressource: A standard model an 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

Thanks



Paulette Lieby



Jean-François Dufayard



Frédéric de Lamotte



Hélène Chiapello



Thomas Denecker

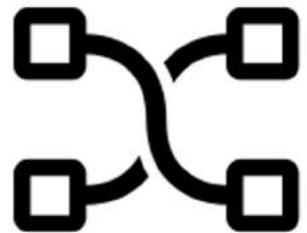
Supplementary slides

Standard for data and metadata



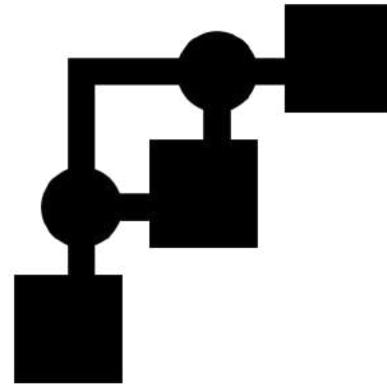
Guidelines or checklists

Ex: the GSC checklist



Models or schemas

Ex: ISA model



**Terminology artefacts,
ontology**

Ex: The Gene Ontology

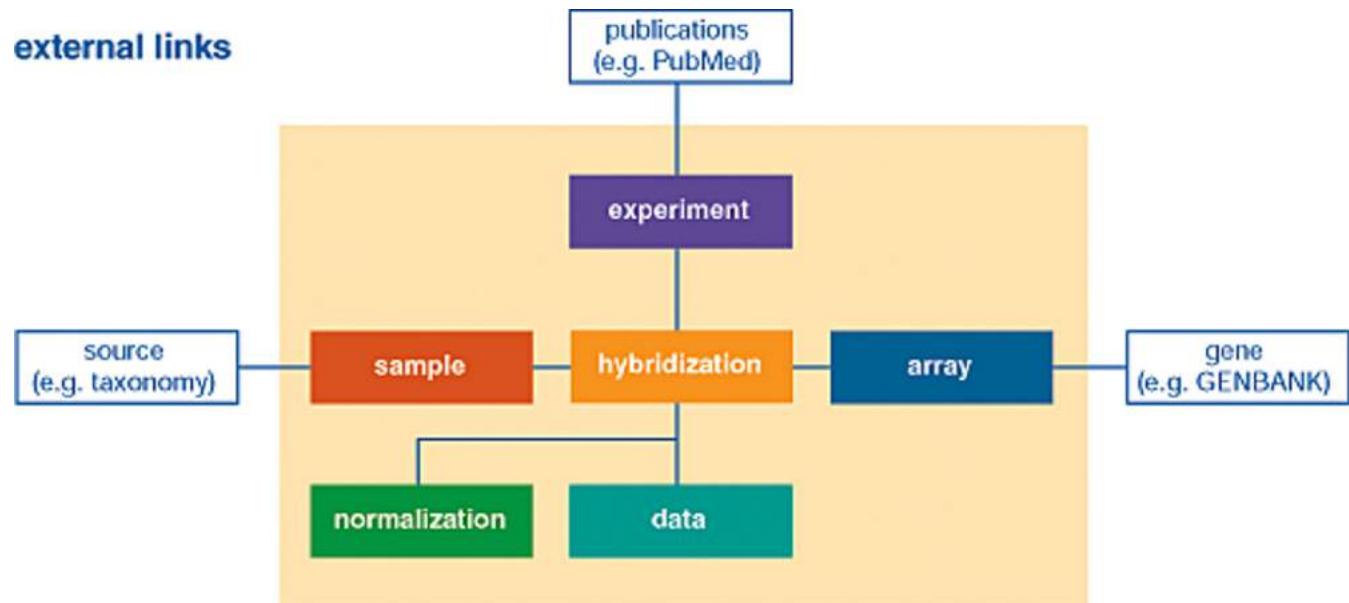


Identifier schemata

<https://fairsharing.org>

The Minimum information standard initiative

- A set of **guidelines** for **reporting data** derived by relevant methods in biosciences.
- Example : the **Minimum Information About a Microarray Experiment (MIAME)**

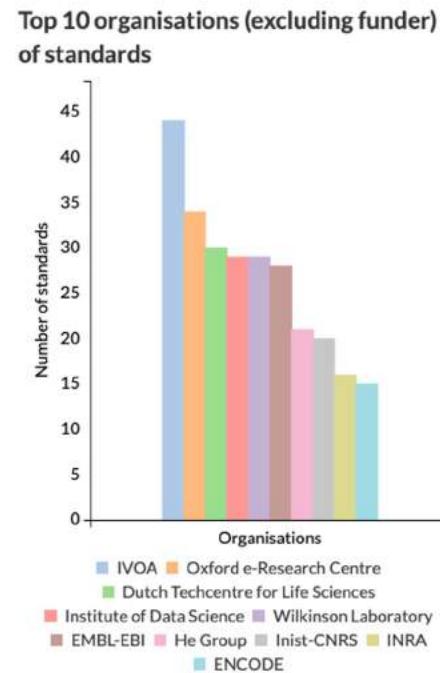
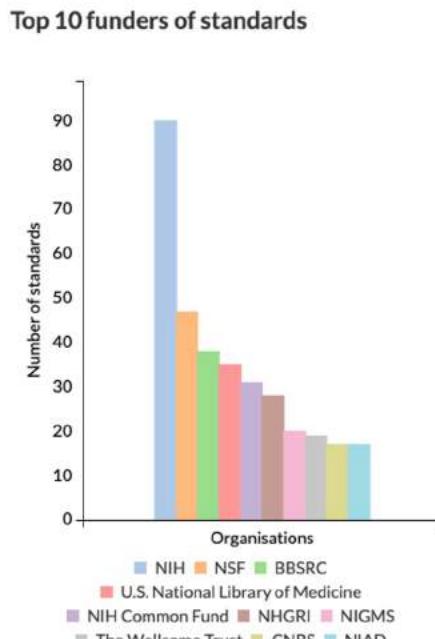
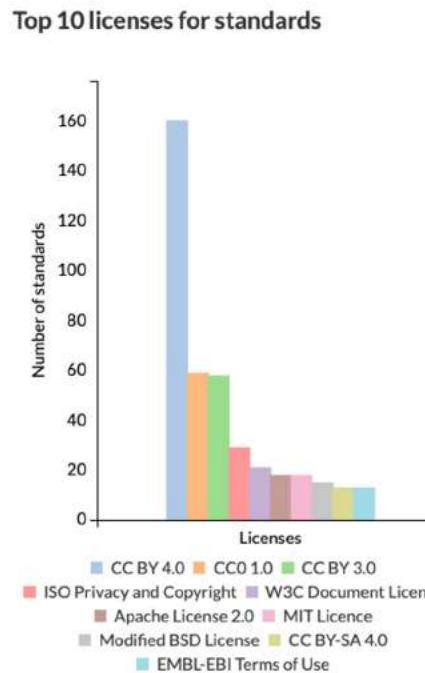


A schematic representation of six components of a microarray experiment.

https://en.wikipedia.org/wiki/Minimum_information_standard

[10.1038/ng1201-365](https://doi.org/10.1038/ng1201-365)

Summary statistics about standards



The CC by 4.0 licence is the most adopted

US and UK National institutes are the most important funders

Worldwide Research Organisations produce standards

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

Retour d'expérience de soumission en banque de données internationales

valerie.cognat@ibmp-cnrs.unistra.fr

&

laurent.bouri@igbmc.fr

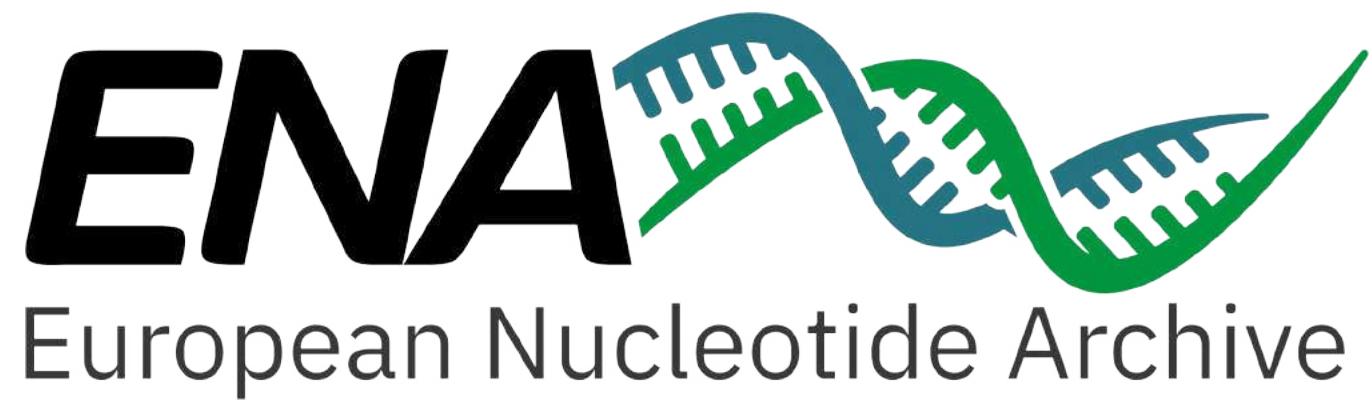


Pourquoi soumettre mes données ?

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

3 bases de données





**Qui a déjà
soumis à
l'ENA ?**

C'était facile ?



La base de données

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

La documentation

The screenshot shows a documentation page for the European Nucleotide Archive (ENA). The top navigation bar includes 'ENATraining Modules' (latest), a search bar, and a 'Docs' link. On the right, there's a 'Edit on GitHub' button. The main content area is titled 'ENA: Guidelines and Tutorials'. It features a welcome message and links to various submission and retrieval guides. Below this, sections include 'ENA Data Submission', 'ENA Data Discovery & Retrieval', 'ENA Data Updates', and 'Tips and FAQs', each with a list of related topics.

ENATraining Modules latest

Search docs

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

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

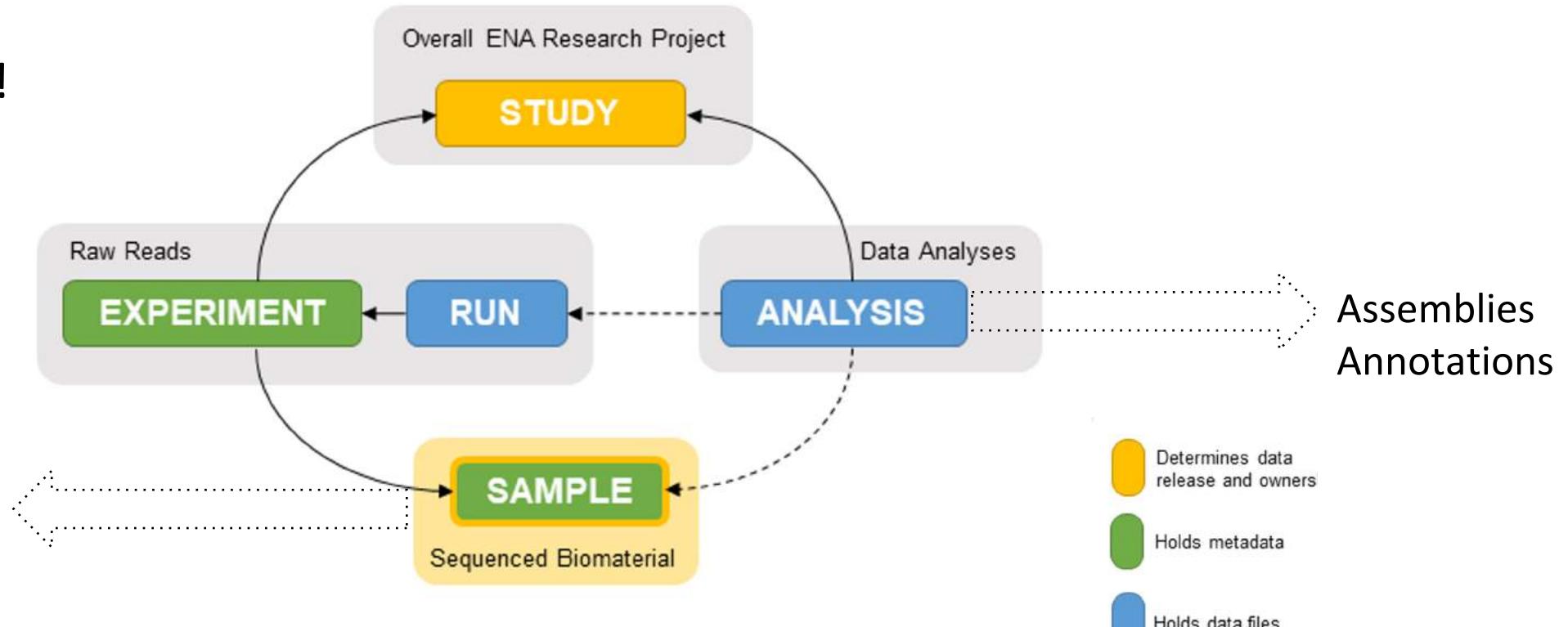
Tips and FAQs

- 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/>

Modèle des métadonnées

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>

Description des expériences et validation

Metadata validation

Permitted values for platform

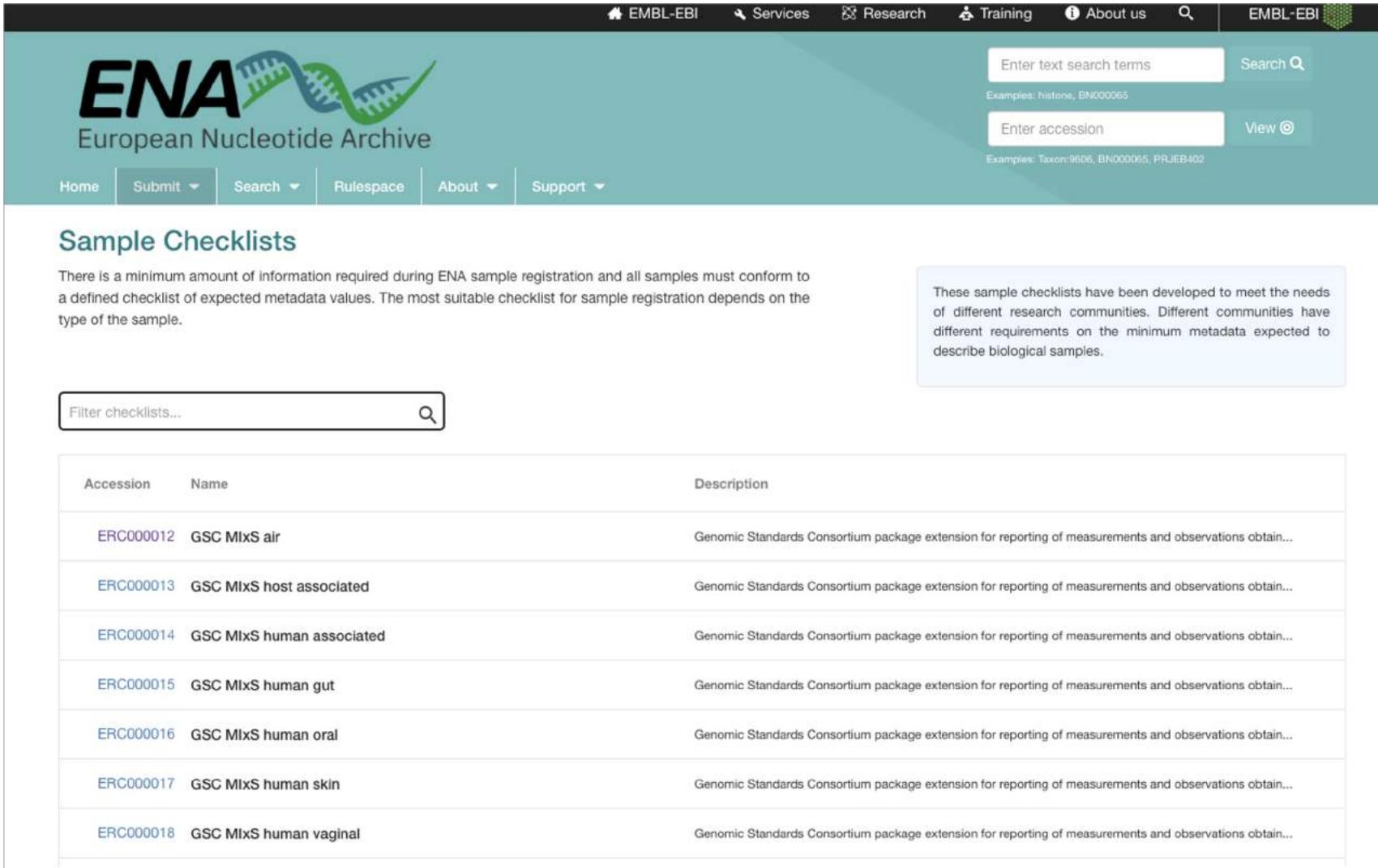
- LS454: 454 technology use 1-color sequential flows
- ILLUMINA: Illumina is 4-channel flowgram with 1-to-1 mapping between basecalls and flows
- PACBIO_SMRT: PacificBiosciences platform type for the single molecule real time (SMRT) technology.
- ION_TORRENT: Ion Torrent Personal Genome Machine (PGM) 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>

Les checklists de l'ENA pour les “samples”

- 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>

Listes des checklists pour les “Sample”



The screenshot shows the ENA (European Nucleotide Archive) website with a focus on 'Sample Checklists'. The top navigation bar includes links for EMBL-EBI, Services, Research, Training, About us, and a search function. The main header features the ENA logo and 'European Nucleotide Archive'. Below the header, there are two search boxes: one for 'Enter text search terms' and another for 'Enter accession'. The main content area is titled 'Sample Checklists' and contains a sub-section about minimum metadata requirements. A table lists 18 sample checklists, each with an accession number, name, and a truncated description.

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...

Méthodes de soumission

	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

Interactive

≡ 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](#).

```
graph TD; A[Overall ENA Research Project] --> B(STUDY); B --> C[EXPERIMENT]; B --> D[RUN]; B --> E[ANALYSIS]; C --> F[Raw Reads]; F --> D; E --> G[SAMPLE<br/>Sequenced Biomaterial];
```

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

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

Web-Cli

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)	



Programmatic

- **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>
```

Les outils complémentaires

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 EMBL) 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.

ChEMBL



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

[Web API](#)

[See all data resources](#)

[EMBL-EBI](#) [Services](#) [Research](#) [Training](#) [About us](#) [EMBL-EBI](#)

Search Examples: MGYS0000410, Tara Oceans, Human Gut



Getting started

Search by

Name, biome, or keyword

[Text search](#)

Sequence similarity

[Sequence search](#)

Or by data type

XXX

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

Or by selected biomes



[Browse all biomes](#)

Request analysis of

Your data

[Submit and/or Request](#)

A public dataset

[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_cyan

Micrometla aeruginosa PMC 728.11 cyan metagenome sequencing

[View all studies](#)

<https://www.ebi.ac.uk/services/all>



**Qui a déjà
soumis à
GEO ?**

C'était facile ?



La base de données

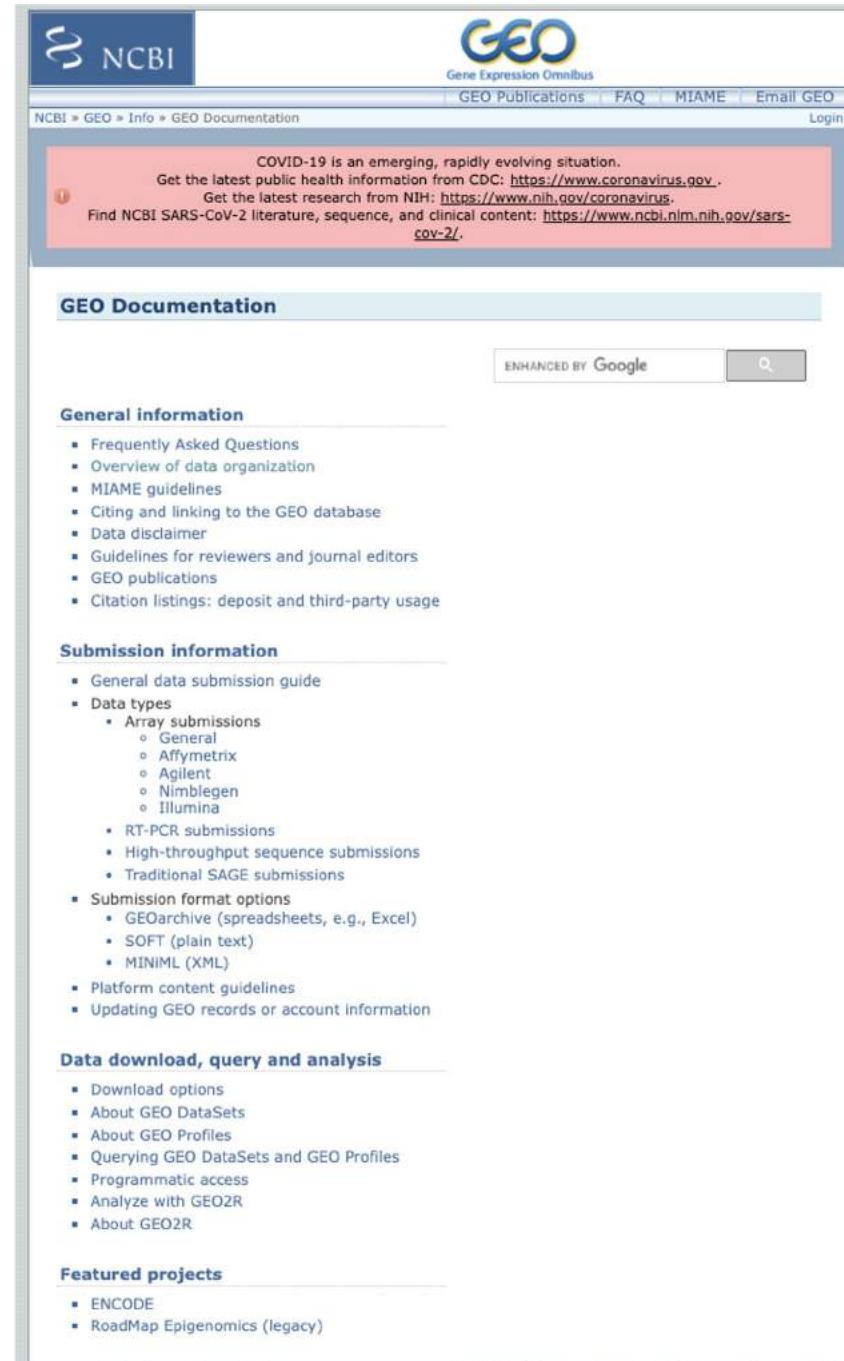
GEO est un dépôt public international qui archive et distribue librement des données de:

- 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/>



The screenshot shows the GEO Documentation page. At the top, there are links for GEO Publications, FAQ, MIAME, Email GEO, and Login. A red banner at the top of the content area provides information about COVID-19, CDC, NIH, and NCBI SARS-CoV-2 resources. Below the banner, the page title is "GEO Documentation". There is an "ENHANCED BY Google" button and a search icon. The page is divided into several sections: "General information", "Submission information", "Data download, query and analysis", and "Featured projects". Each section contains a bulleted list of topics.

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

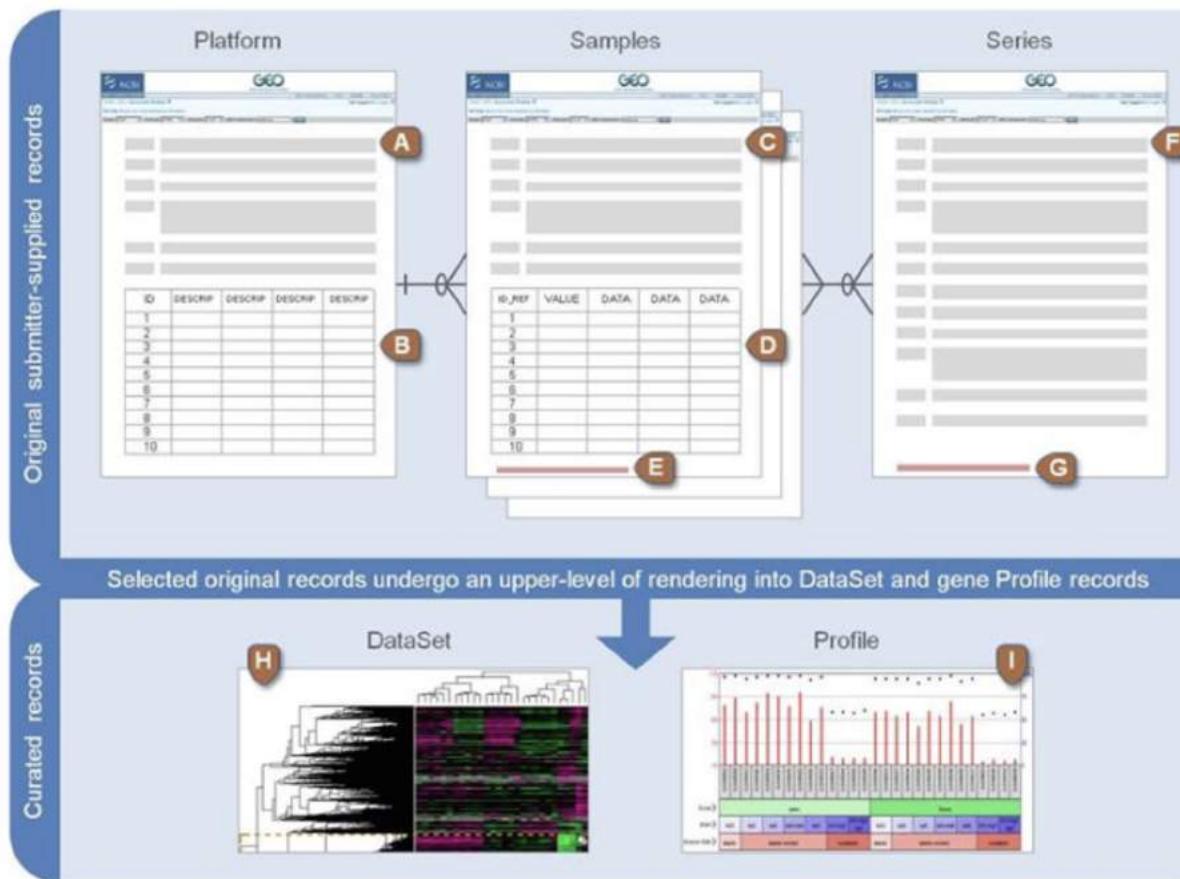
Data download, query and analysis

- 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 records are supplied by submitters

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 »](#)

A Text description of the array or sequencer

B Text tab-delimited table of the array template

C Text description of the biological sample and protocols to which it was subjected

D Text tab-delimited table of processed hybridization result (may optionally include raw data columns)

E Original raw data file, or processed sequence data file

F Text description of the overall experiment

G Tar archive of original raw data files, or processed sequence data files

Sample records are supplied by submitters

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 »](#)

Series records are supplied by submitters

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 »](#)

Fichiers

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 SAMPLES section of the metadata spreadsheet. If your array does not already exist in GEO, please include a PLATFORM 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](#)

Exemple Excel Illumina

GA_illumina_expression.xls [Mode de compatibilité]

Accueil Insertion Dessin Mise en page Formules Données Révision Affichage

Coller G I S Mise en forme conditionnelle Mettre sous forme de tableau Styles de cellule Insérer Supprimer Mise en forme Trier et filtrer Rechercher et sélectionner

F7 fx

	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 <i>in vitro</i> cyclic tensile strain compared to unstrained control fascicles.										
5	contributor	Jane.Doe										
6	contributor	John.A.Smith										
7	SAMPLES											
8	# The corresponding example matrix table is included in the next worksheet.											
9	Sample name	title	source name	organism	idat file	characteristics: Strain	characteristics: age	characteristics: tiss molecule	label	description	platform	
10	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
11	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
12	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
13	PROTOCOLS											
14	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 Bioanalyzer.										
15	label protocol	Biotinylated cRNA were prepared with the Ambion MessageAmp kit for Illumina arrays										
16	hyb protocol	Standard Illumina hybridization protocol										
17	scan protocol	Standard Illumina scanning protocol										
18	data processing	The data were normalised using quantile normalisation with IlluminaGUI in R										
19	value definition	quantile normalized										
20												
21												
22												
23												
24												
25												

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

Prêt

Les outils complémentaires : GeoToR

exemple : GSE25724

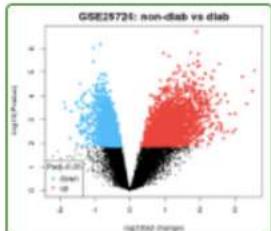
GEO accession GSE25724 Set Expression data from type 2 diabetic and non-diabetic isolated human islets

Samples												Define groups		Selected 13 out of 13 samples	
Group	Accession	Title	Source name	Tissue	Disease state	Age	Gender	Characteristics	Columns	Set					
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							

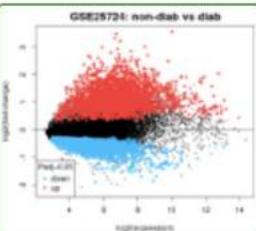
Visualization



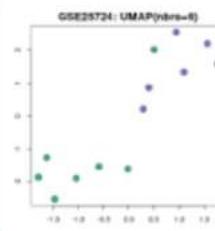
GSE25724: non-diab vs diab



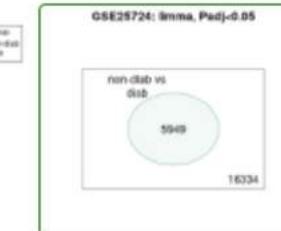
GSE25724: non-diab vs diab



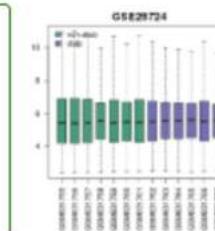
GSE25724: UMAP(nbr=8)



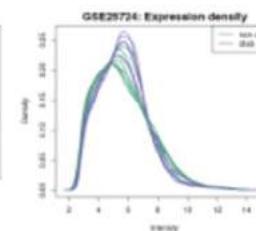
GSE25724: limma, Padj<0.05



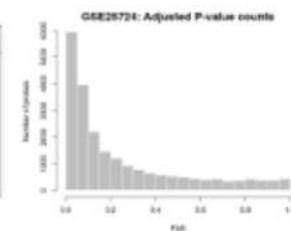
GSE25724



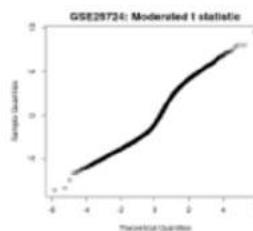
GSE25724: Expression density



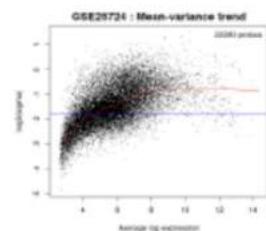
GSE25724: Adjusted P-value counts



GSE25724: Moderated t statistic



GSE25724 : Mean-variance trend



<https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html>



**Qui a déjà
soumis à
GISAID ?**

C'était facile ?



Présentation de la base

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**.

Le fichier de métadonnées

Fichier excel

The screenshot shows a Microsoft Excel spreadsheet titled "20210222_EpiCoV_BulkUpload_Template.xls". The "Instructions" tab is selected. The content includes:

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"

Column information

Column	Type	Description
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, Viral
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, <i>Manis javanica</i> , <i>Rhinolophus affinis</i> , etc
Additional host information		e.g. Patient infected while traveling in ...
Sampling Strategy		e.g. Sentinel surveillance (IL), 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, Cloakai 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	mandatory	Where sequence data have been generated and submitted to GISAID
Submitting lab	mandatory	
Address	mandatory	
Sample ID given by the submitting laboratory	mandatory	a comma separated list of Authors with complete First followed by Last Name
Authors	mandatory	
Comment	leave empty	do not use this column
Comment icon	leave empty	do not use this column

Instructions Submissions +

GISAID

WEB - Single

© 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* hCoV-19/Country/Identifier/2021

Accession ID

Type betacoronavirus

Passage details/history* Example: Original, Vero

Sample information

Collection date* Example: 2021-03-27, 2021-03 (collection in March, specific day unknown), 2021 (collection in 2021, month and day unknown)

Location* Continent / Country or Territory / Region

Additional location information Travel history, Residence, Cruise ship; ...

Host* Human, Environment, Canis lupus

Additional host information Example: Underlying health conditions; other host relevant characteristics

Outbreak Detail Example: Date, Place, Family cluster

Sampling strategy Baseline surveillance, Active surveillance, Clinical trial; ...

Gender* Male, Female, or unknown

Patient age* Example: 65, 7 months, or unknown

Patient status* Hospitalized, Released, Live, Deceased

Specimen source Sewage, Sputum, Alveolar lavage fluid, Oropharyngeal swab, Mid-Turbinate swab, Nasopharyngeal swab, Blood, Tracheal swab, Urine, Stool, Other

Last vaccinated provide details if applicable

Treatment Example: Include drug name, dosage

Sequencing technology* Example: Illumina MiSeq, Sanger, Nanopore MinION, Ion Torrent, etc.

Assembly method Example: CLC Genomics Workbench 12, Geneious 10.2.4, SPAdes/MEGAHIT v1.2.9, UGENE v. 33, etc.

Coverage Example: 70x, 1,000x, 10,000x (average)

Institute information

Originating lab* Where the clinical specimen or virus isolate was first obtained

Web - Batch upload

The screenshot shows the GISAID EpiCoV™ web interface with the following details:

- Header:** GISAID logo, navigation links for Registered Users, EpiFlu™, EpiCoV™, My profile, and a user login message.
- Main Content:** A large green header bar with the text "GISAID hCoV-19 Batch Upload". Below it, a sub-header reads: "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."
- Form Fields:**
 - Metadata as Excel or CSV***: A file input field with a size limit of 5M. It displays the message "max size: 5M Choisir le fichier aucun fichier sélectionné".
 - Sequences as FASTA***: A file input field with a size limit of 32M. It displays the message "max size: 32M Choisir le fichier aucun fichier sélectionné".
 - Confirmation options**: A dropdown menu set to "(Default) Notify me about ALL DETECTED FRAMESHIFTS in this submission for reconfirmation of affected sequences".
 - Report**: A link to "Upload XLS/CSV and FASTA...".
- Footer:** Buttons for "Download Instructions and Template", "Contact Curator", and "Verify and Submit". A note at the bottom states: "Important note: In the GISAID EpiCoV™ 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 EpiCoV™ Database Access Agreement in respect of such data in the same manner as if they were data relating to influenza viruses."

GISAID CLI2

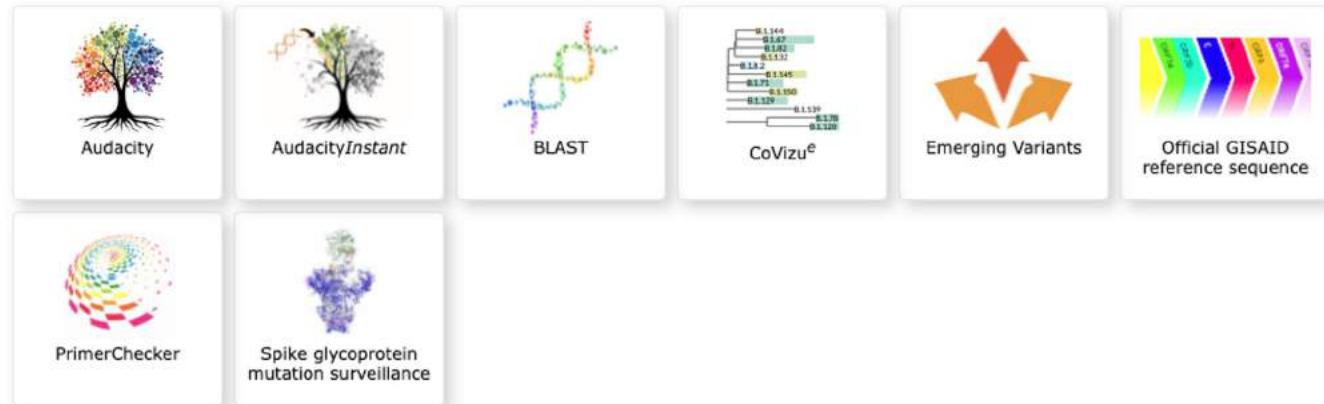
Version 2 Command Line Interface (CLI) for batch uploading

```
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]

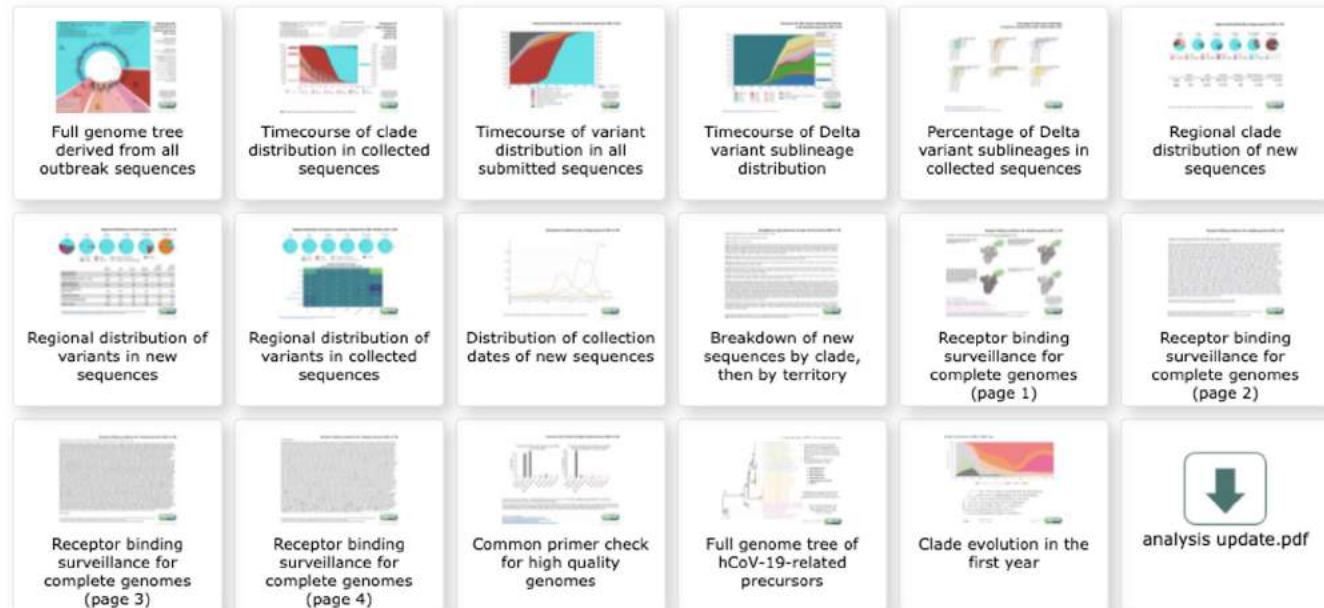
Perform upload of sequences and metadata to GISAID's curation zone.

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)
```

Les outils complémentaires



Analysis Update (2021-11-05)



Data brokering à l'IFB

Pourquoi le développer 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.

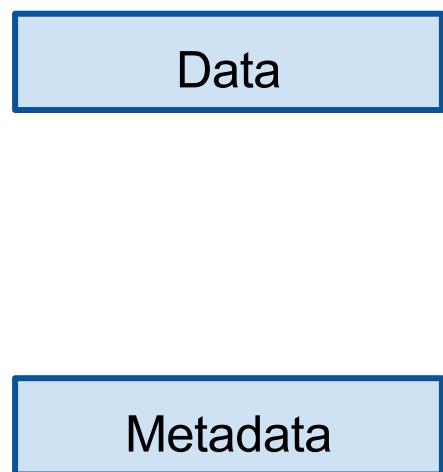
L'ENA a demandé à l'IFB de devenir le data broker français

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.

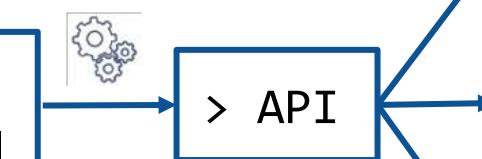
3 types d'activités : le développement d'outils, la formation et le support aux utilisateurs.

Data Brokering service developed by IFB

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



...

Des outils de data brokering déjà disponibles

The screenshot shows the homepage of the gfbio website. At the top, there is a navigation bar with links for 'About', 'Services', 'Infothek', 'Events', and 'GFBio e.V.'. The main title 'FAIR • Research • Data' is prominently displayed, followed by the subtitle 'Biodiversity, Ecology & Environmental Science'. Below the title is a search bar with the placeholder 'Enter a search term...' and a 'FIND DATA' button. Two thumbnail images are shown: one of a landscape with hills and another of a butterfly on a flower. Below these are four icons with labels: 'Plan' (pencil and ruler), 'Submit' (cloud with arrow), 'Visualize' (map), and 'Find Data' (magnifying glass). The background features a faint radial pattern.

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

The screenshot shows the homepage of the METAGENOTE website. At the top, there is a navigation bar with links for 'NIH NIAID', 'METAGENOTE', 'BROWSE', 'USER GUIDE', 'ABOUT', and 'FAQS'. A 'Contact Us' button is located in the top right corner. A red banner at the bottom of the header area states 'COVID-19 is an emerging, rapidly evolving situation' and provides links to CDC and NIH websites. To the right of the banner is a button labeled 'Learn to Publish COVID-19 Data to SRA'. The main content area features a heading 'METAGENOTE is a quick and intuitive way to annotate data from genomics studies including microbiome.' and a 'Start Here!' button. Below this are sections for 'Why use METAGENOTE?' and four functional icons: 'Annotate' (document icon), 'Use Standards' (blue ribbon icon), 'Store & Search' (network icon), and 'Publish' (cloud with arrow icon).

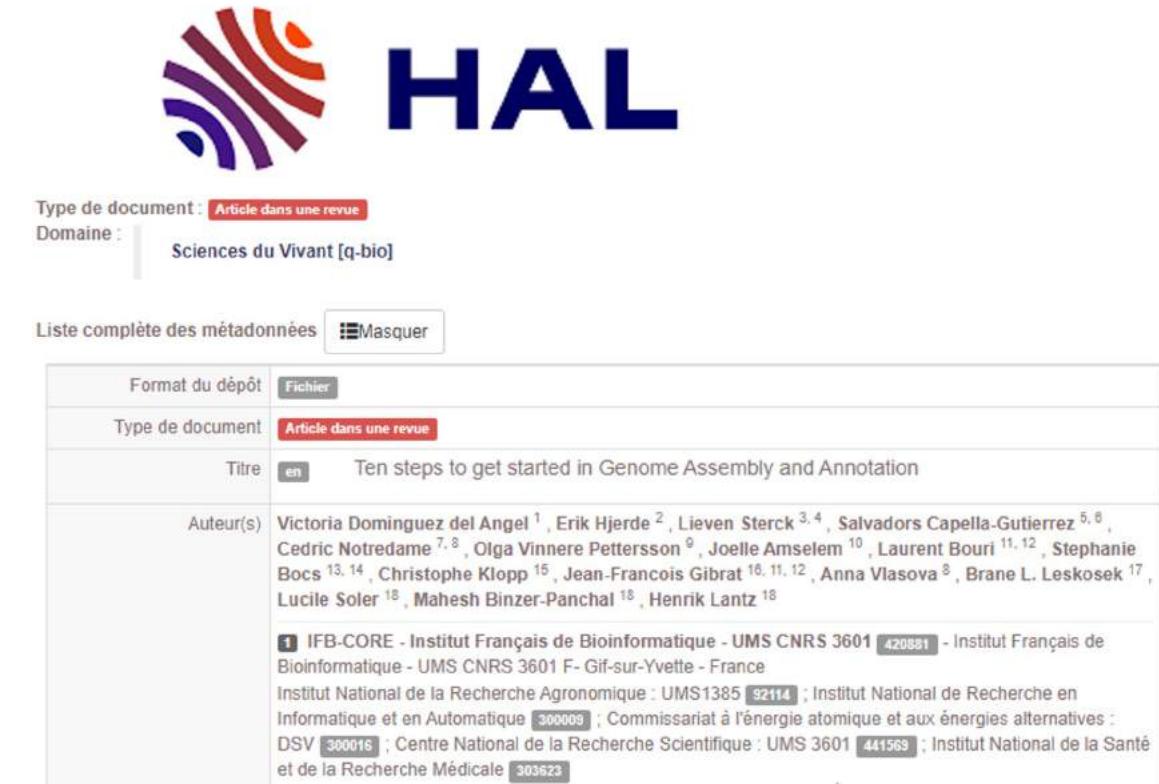
<https://metagenote.niaid.nih.gov/>

Publication d'articles/ **OpenLink**

Les entrepôts de données



- Applications Web open source pour partager, conserver, citer, explorer et analyser les données de recherche.
- Facilite la mise à disposition des données aux autres et vous permet de reproduire leurs travaux.



A screenshot of the HAL API interface showing a search result for an article. The article title is "Ten steps to get started in Genome Assembly and Annotation". It lists authors from various institutions, including IFB-CORE, Institut Français de Bioinformatique, UMS CNRS 3601, Institut National de la Recherche Agronomique, Institut National de Recherche en Informatique et en Automatique, Commissariat à l'énergie atomique et aux énergies alternatives, DSV, Centre National de la Recherche Scientifique, UMS 3601, Institut National de la Santé et de la Recherche Médicale, and Institut Français de Bioinformatique. The document is in English and has a DOI of 10.5281/zenodo.420581.

Type de document: Article dans une revue
Domaine : Sciences du Vivant [q-bio]

Liste complète des métadonnées

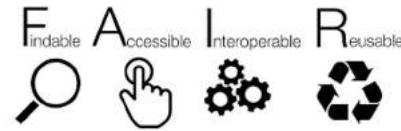
Format du dépôt	Fichier
Type de document	Article dans une revue
Titre	en Ten steps to get started in Genome Assembly and Annotation
Auteur(s)	Victoria Dominguez del Angel ¹ , Erik Hjerde ² , Lieven Sterck ^{3,4} , Salvadors Capella-Gutierrez ^{5,6} , Cedric Notredame ^{7,8} , Olga Vinnere Pettersson ⁹ , Joelle Amselem ¹⁰ , Laurent Bouri ^{11,12} , Stephanie Bocs ^{13,14} , Christophe Klopp ¹⁵ , Jean-Francois Gibrat ^{16,11,12} , Anna Vlasova ⁸ , Brane L. Leskosek ¹⁷ , Lucile Soler ¹⁸ , Mahesh Binzer-Panchal ¹⁸ , Henrik Lantz ¹⁸
1 IFB-CORE - Institut Français de Bioinformatique - UMS CNRS 3601 420581 - Institut Français de Bioinformatique - UMS CNRS 3601 F- Gif-sur-Yvette - France	
Institut National de la Recherche Agronomique : UMS1385 32114 ; Institut National de Recherche en Informatique et en Automatique 300009 ; Commissariat à l'énergie atomique et aux énergies alternatives : DSV 300016 ; Centre National de la Recherche Scientifique : UMS 3601 441563 ; Institut National de la Santé et de la Recherche Médicale 303623	

Organisation des données avec OpenLink

Buts



Une **vision claire** des données associées à chaque projet de recherche



Réduire les obstacles à l'adoption des principes FAIR



Limiter l'impact de FAIR sur le temps de gestion des données



Assister les chercheurs dans la publication de leurs données

Organisation des données avec OpenLink

Solution

django

Une application web open-source basée sur le framework Django (langage Python)



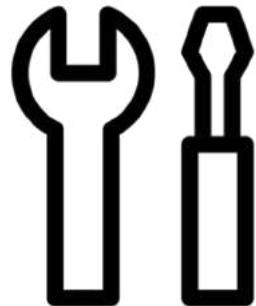
Une base de données pour créer des liens entre la structure d'un projet de recherche (modèle ISA) et de multiples sources de données

Source code: <https://gitlab.com/igbmc/openlink>

Documentation: <https://openlink.readthedocs.io/en/latest>

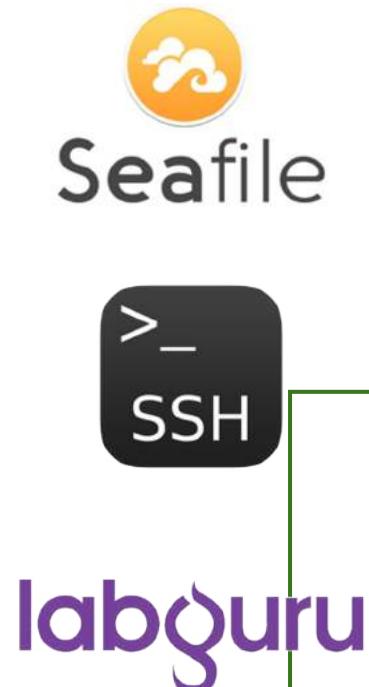


Une collection évolutive de connecteurs aux outils couramment utilisés par les chercheurs: LabGuru, Omero, Seafile, mass storage, etc.



Des outils intégrés pour faciliter la manipulation des données

Openlink: De la recherche à la publication de données



OpenLink



Projet complet avec plus de données liées

OpenLink

Micro irradiation UV

Seafile: 4.8 MB / Omero: 265.67 KB / Zenodo: 0 B / SSH: 47.05 KB / total: 5.11 MB

DNA damage

Microlrradiation UV

Test Micro irradiation UV / DNA damage

Folder_DNA_UV

paper103_article_rev5213_20191025_09160...

img cells data

cells

FRAP

FRAP experience

protocole.txt

Add study Add assay Add dataset

Expand All Collapse All

Version 0.9

Publication sur Zenodo

Micro irradiation UV [Edit](#) [Manage Users](#) [Manage Tools](#) [Add investigation](#)

Select Tool
Show 25 entries Search:

Tool name	Url host	Date created
Zenodo <small>zenodo</small>	https://sandbox.zenodo.org/	Sept. 23, 2021, 11:11 a.m.

Previous [1](#) Next

[Back](#) [Next](#)



Publication sur Zenodo

Micro irradiation UV Edit Manage Users Manage Tools Add investigation

- Microirradiation UV
 - Test Micro irradiation UV / DNA damage
 - Folder_DNA_UV
 - paper103_article_rev5213_20191025_09160...
 - img cells data
 - cells
 - FRAP
 - + FRAP experience

back Publish Data



Publication sur Zenodo

The screenshot shows the OpenLink interface for publishing data to Zenodo. The left sidebar includes links for Home, Projects, Contact, Logout, Igbmc, and Admin, along with logos for IFB and IGBMC. The main content area displays a dataset titled "Micro irradiation UV".

Dataset Table:

Dataset	Assay	Study	Link
cell4	Test Micro irradiation UV / DNA damage	MicroIrradiation UV	
Folder_test_lenght_name	Test Micro irradiation UV / DNA damage	MicroIrradiation UV	View on Seafie
img cells data	Test Micro irradiation UV / DNA damage	MicroIrradiation UV	View on Omero
paper103_article_rev5213_20191025_091609.pdf	Test Micro irradiation UV / DNA damage	MicroIrradiation UV	View on Seafie
test.txt	FRAP experience	FRAP	

Form Fields:

- Title: DNA damage
- Author: bouri (IGBMC, ORCID: 0000-0002-2297-1559)
- Author: jloup (IGBMC, ORCID: blank)
- Tags: DNA damage, Micro irradiation UV
- Description: DNA damage
Imaging of dynamics of proteins involved in the DNA damage (DNA double-strand breaks (DSBs) response :
- micro irradiation to create DNA damage and observation of the recruitment of reparation protein,
- FRAP
- Buttons: back, Valid Data and Publish

A large green arrow points to the "Valid Data and Publish" button at the bottom of the form.

Publication sur Zenodo



Publication sur Zenodo

zenodo

Search Dataset Open Access

Upload Communities laurent.bouri2@gmail.com

January 26, 2022

DNA damage

bouril; jloup

DNA damage

Imaging of dynamics of proteins involved in the DNA damage (DNA double-strand breaks (DSBs)) response :

- micro irradiation to create DNA damage and observation of the recruitment of reparation protein,
- FRAP

Articles de référence:

<https://www.nature.com/articles/ncb945#Sec2>

<http://jcb.rupress.org/content/170/2/201>

Preview

DNA damage.zip

- DNA damage
 - ■ Microirradiation UV
 - FRAP
 - FRAP experience
 - test.txt
 - Test Micro irradiation UV
 - DNA damage
 - Folder_test_lenght_name
 - cahier_lab.pdf
 - cell4
 - img cells data
 - cell0001.lsm.tif
 - cell0003.lsm.tif
 - cell0004_AOt42.lsm.tif

0 views 0 downloads

See more details...

Indexed in

OpenAIRE

Publication date: January 26, 2022

DOI: DOI 10.5072/zenodo.1003314

Keyword(s): Micro irradiation UV DNA damage

License (for files): Creative Commons Zero v1.0 Universal

Versions

Version 1 Jan 26, 2022 10.5072/zenodo.1003314

Supplementary data



The omicsBroker tool

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,
- ...

Exemple du prototype

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-cl.html#permitted-values-for-platform>

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

Harmonized Name
PLATFORM

* Mandatory