DMP and Open Data training

Session 3: Metadata

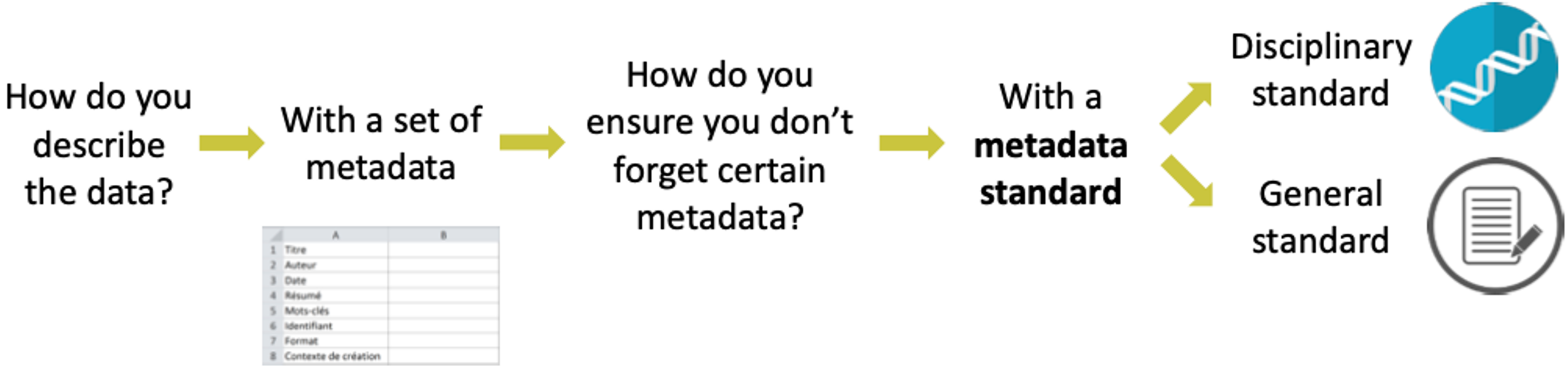
Life science standards and ENA submission

helene.chiapello@inrae.fr & thomas.denecker@france-bioinformatique.fr

Hélène Chiapello : https://orcid.org/0000-0001-5102-0632
Thomas Denecker : https://orcid.org/0000-0003-1421-7641

Metadata & standards in life sciences

Metadata standards help describing data



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

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

Question: What do you know as standard in life sciences ?

10 minutes to find an example (one for data and one for metadata) and write a note in

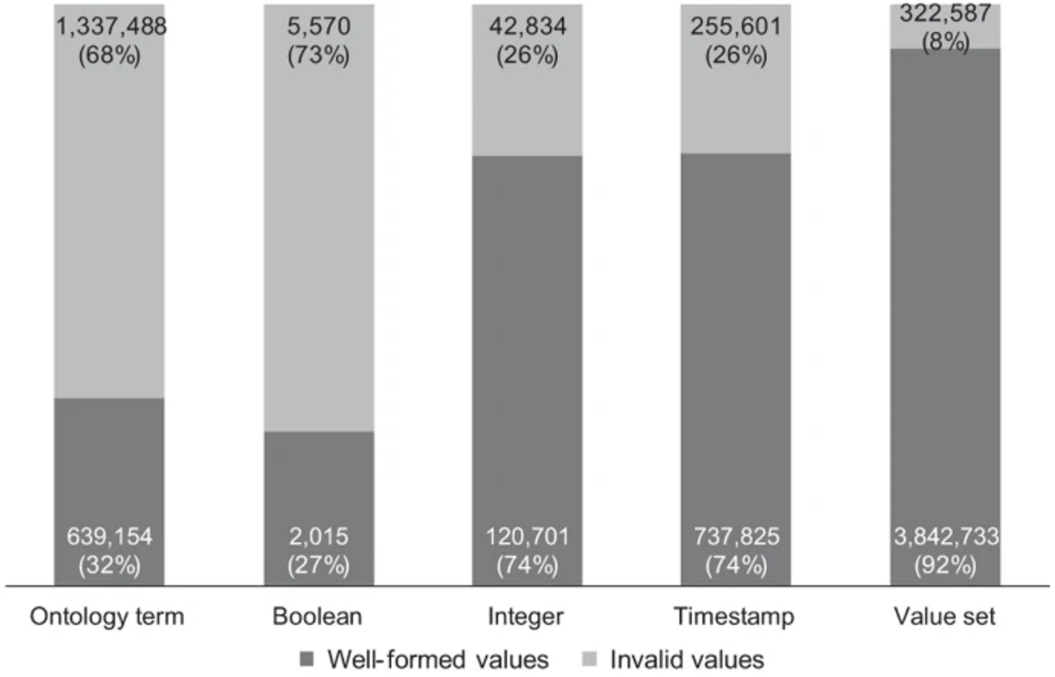
https://scrumblr.ethibox.fr/standard

Metadata exhibit questionable quality in biology

Submission in public resources is often a complex task

Submission procedures are heterogeneous

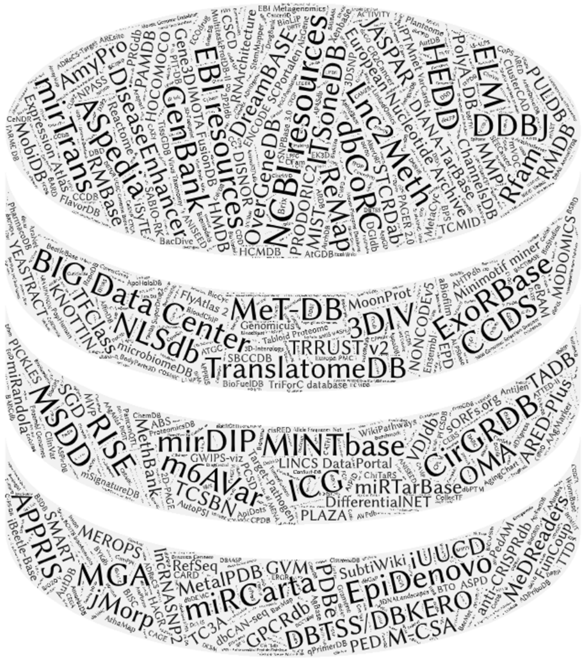
Metadata are often incomplete, inconsistent, redundant or not enough informative



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

Standard adoption and perenity

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



1641 databases in NAR Database 2021

Rigden *et al*, 2021

How do I find the standard I need?

The FAIRsharing portal

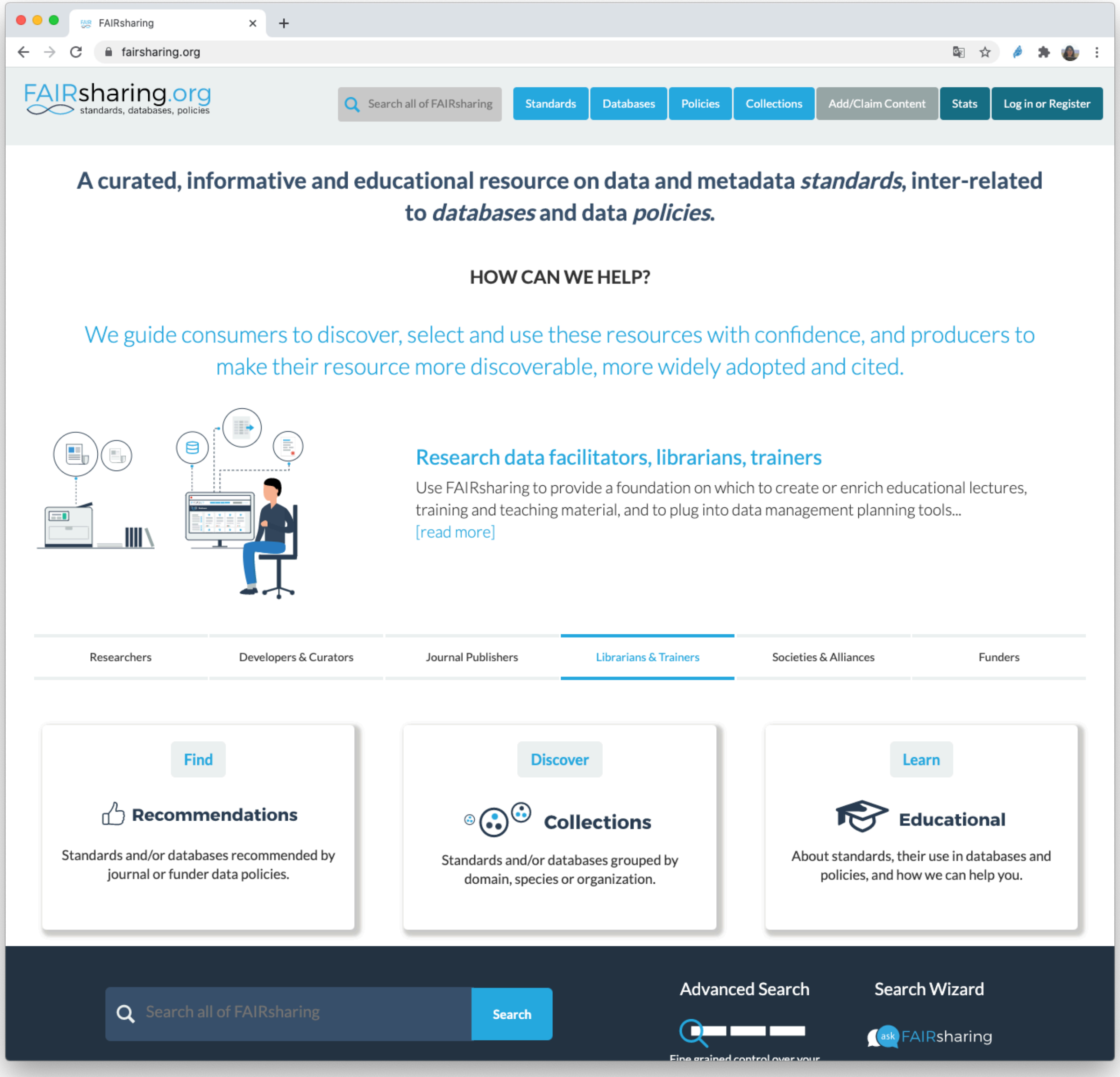
Sansone, *et al.* FAIRsharing as a community approach to standards, repositories and policies.

Nat Biotech. 2019

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

UNIVERSITY OF OXFORD

elixir



A curated, informative and educational resource on data and metadata *standards*, inter-related to *databases* and data *policies*.

HOW CAN WE HELP?

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.

Research data facilitators, librarians, trainers

Use FAIRsharing to provide a foundation on which to create or enrich educational lectures, training and teaching material, and to plug into data management planning tools... [read more]

Find

Recommendations

Standards and/or databases recommended by journal or funder data policies.

Discover

Collections

Standards and/or databases grouped by domain, species or organization.

Learn

Educational

About standards, their use in databases and policies, and how we can help you.

Fine grained control over your

https://fairsharing.org

ifb INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

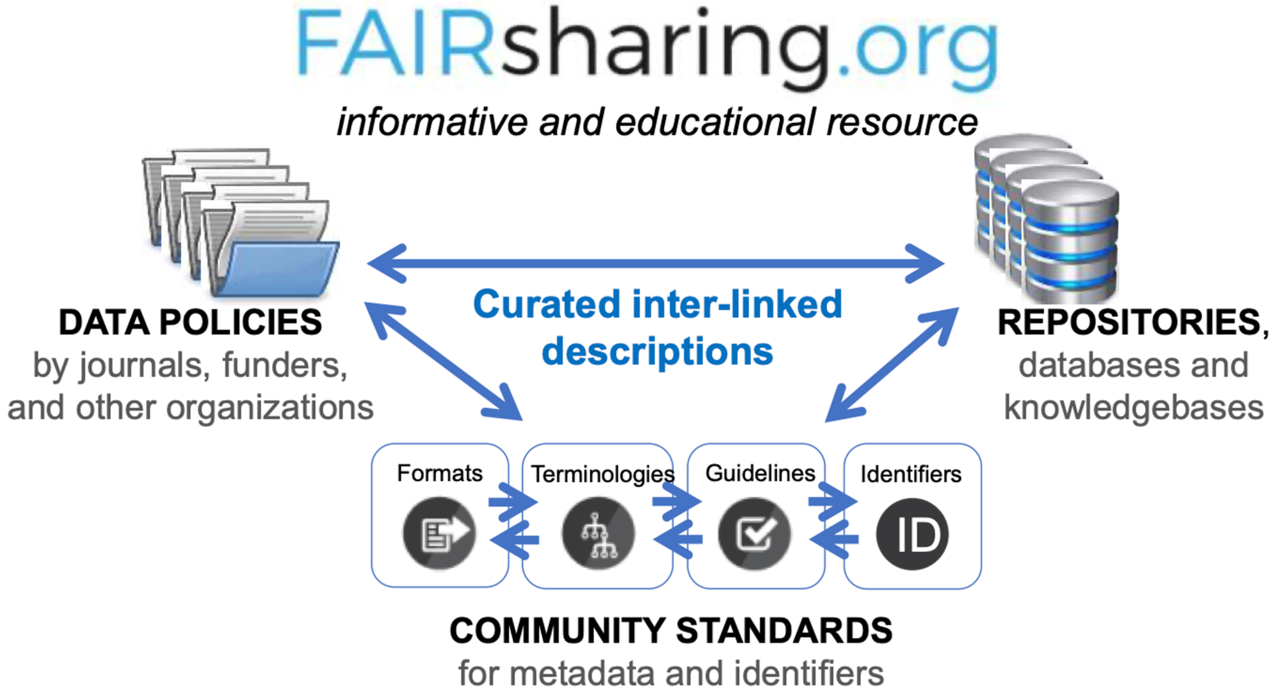
The FAIRsharing portal

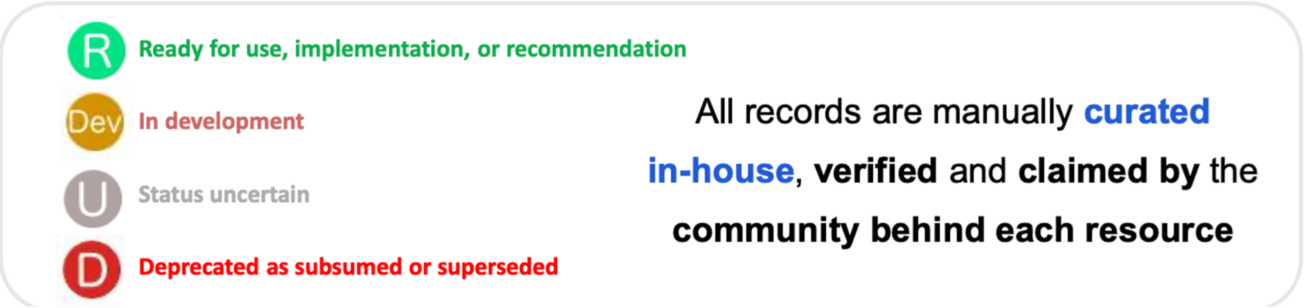
Citable *DOI* for all records

Accessible via *API* or *web interface*

Curation

RECORD STATUS


FAIRsharing.org
informative and educational resource
DATA POLICIES
by journals, funders, and other organizations
Curated inter-linked descriptions
REPOSITORIES,
databases and knowledgebases
Formats
Terminologies
Guidelines
Identifiers
ID
COMMUNITY STANDARDS
for metadata and identifiers



R Ready for use, implementation, or recommendation

Dev In development

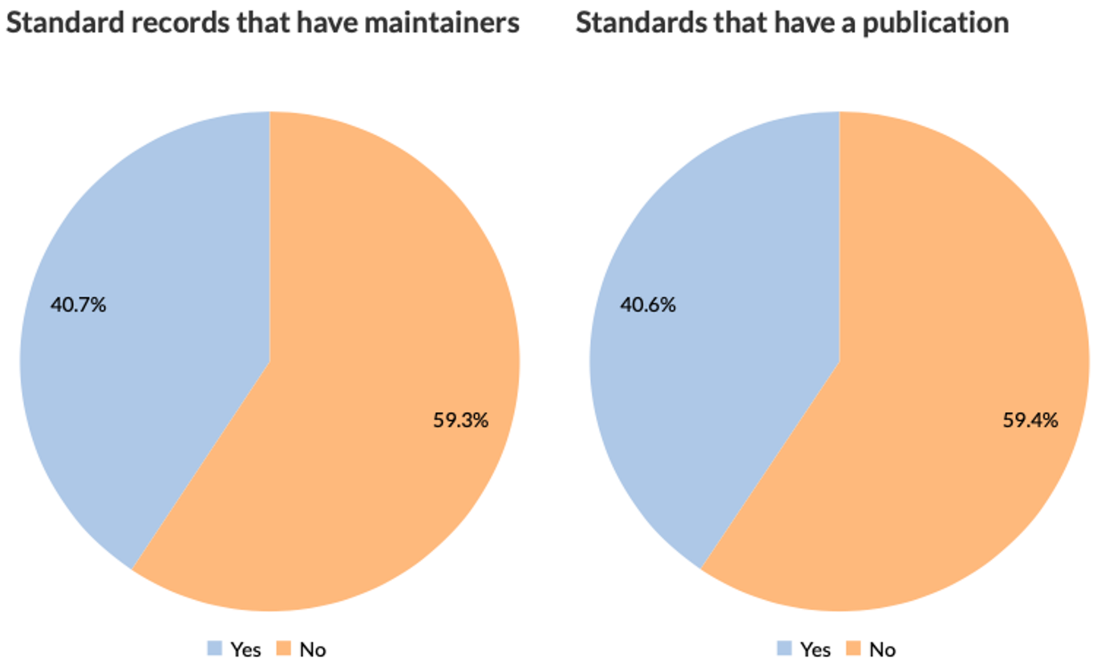
U Status uncertain

D Deprecated as subsumed or superseded

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

https://fairsharing.org

Standard maintenance is a key point



59.3 % of standards have no maintainer

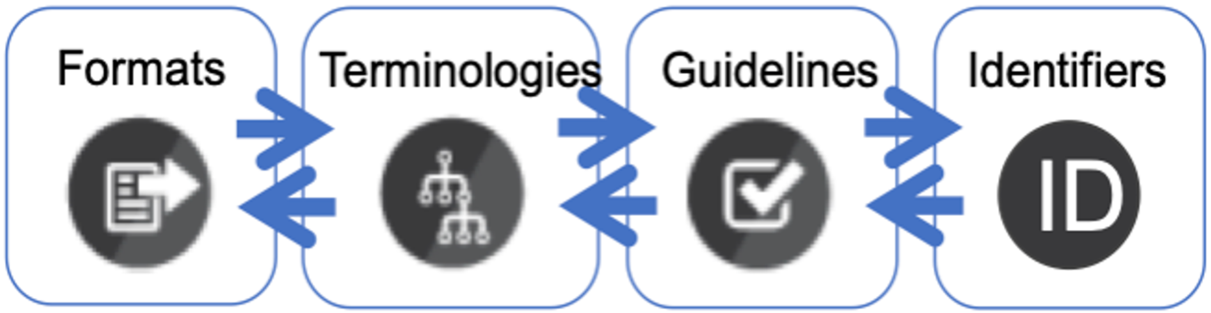
59.4% of standard has no publication

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

Types of data standards

Conceptual model, schema, exchange formats,etc…
e.g. 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

FASTQ

FASTA

SBML

GFF

MIAME

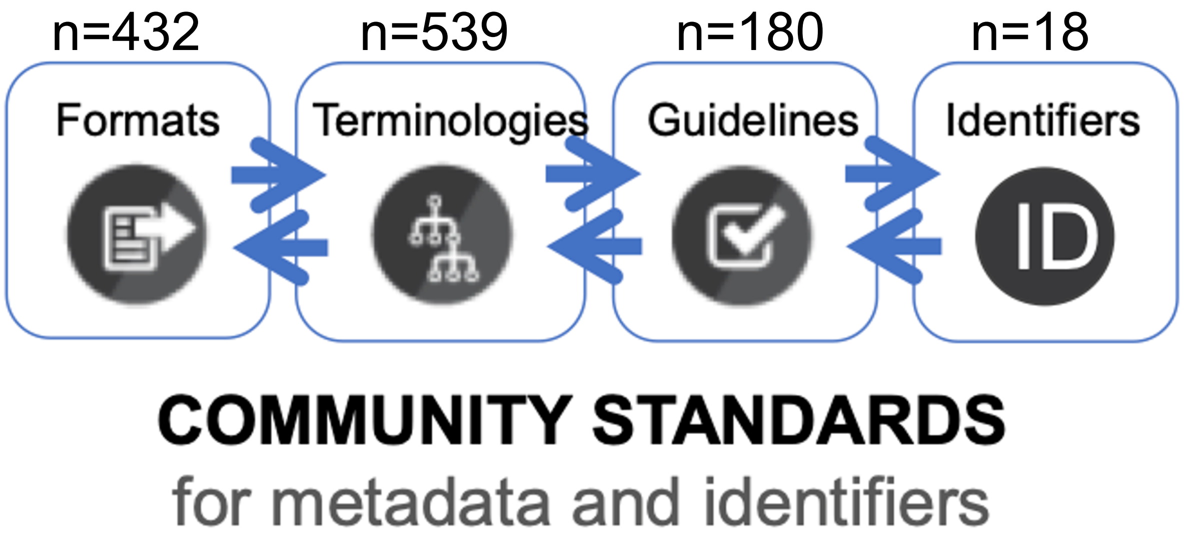
Newick

EC number

BAM

VCF

MINSEQE



Source: https://fairsharing.org/search/?q=Life+science

genomic STANDARDS consortium

isatools

GENEONTOLOGY Unifying Biology

Crop Ontology for agricultural data

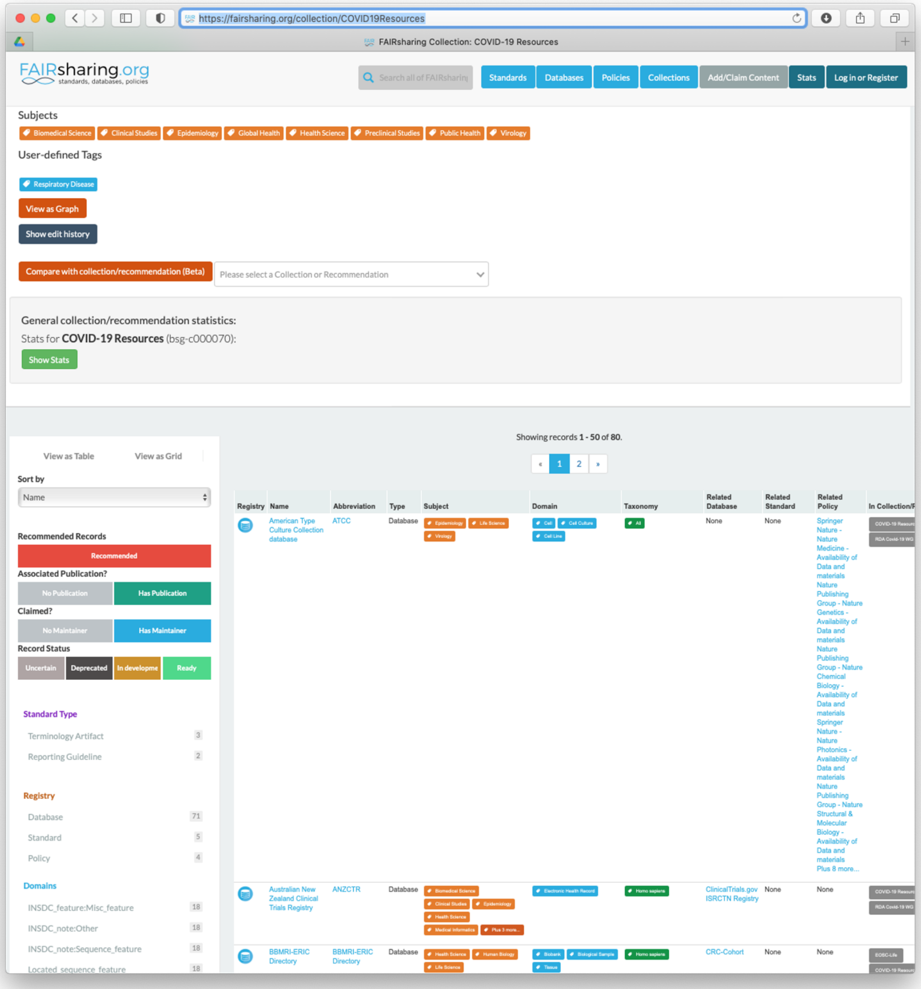
DISEASE ONTOLOGY

Collections in the FAIRsharing portal

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

Graph view to visualize relationship links between resources

https://fairsharing.org/collections/



https://fairsharing.org/collection/COVID19Resources

FAIRsharing Collection: COVID-19 Resources

FAIRsharing.org
standards, databases, policies

Search all of FAIRsharing | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in or Register

Subjects

Biomedical Science | Clinical Studies | Epidemiology | Global Health | Health Science | Preclinical Studies | Public Health | Virology

User-defined Tags

Respiratory Disease

View as Graph

Show edit history

Compare with collection/recommendation (Beta) | Please select a Collection or Recommendation

General collection/recommendation statistics:

Stats for **COVID-19 Resources** (bsg-c000070):

Show Stats

View as Table | View as Grid

Sort by: Name

Recommended Records: Recommended

Associated Publication?: No Publication | Has Publication

Claimed?: No Maintainer | Has Maintainer

Record Status: Uncertain | Deprecated | In developme | Ready

Standard Type

Terminology Artifact	3
Reporting Guideline	2

Registry

Database	71
Standard	5
Policy	4

Domains

INSDC_feature:Misc_feature	18
INSDC_note:Other	18
INSDC_note:Sequence_feature	18
Located_sequence_feature	18

Showing records **1 - 50** of **80**.

« 1 2 »

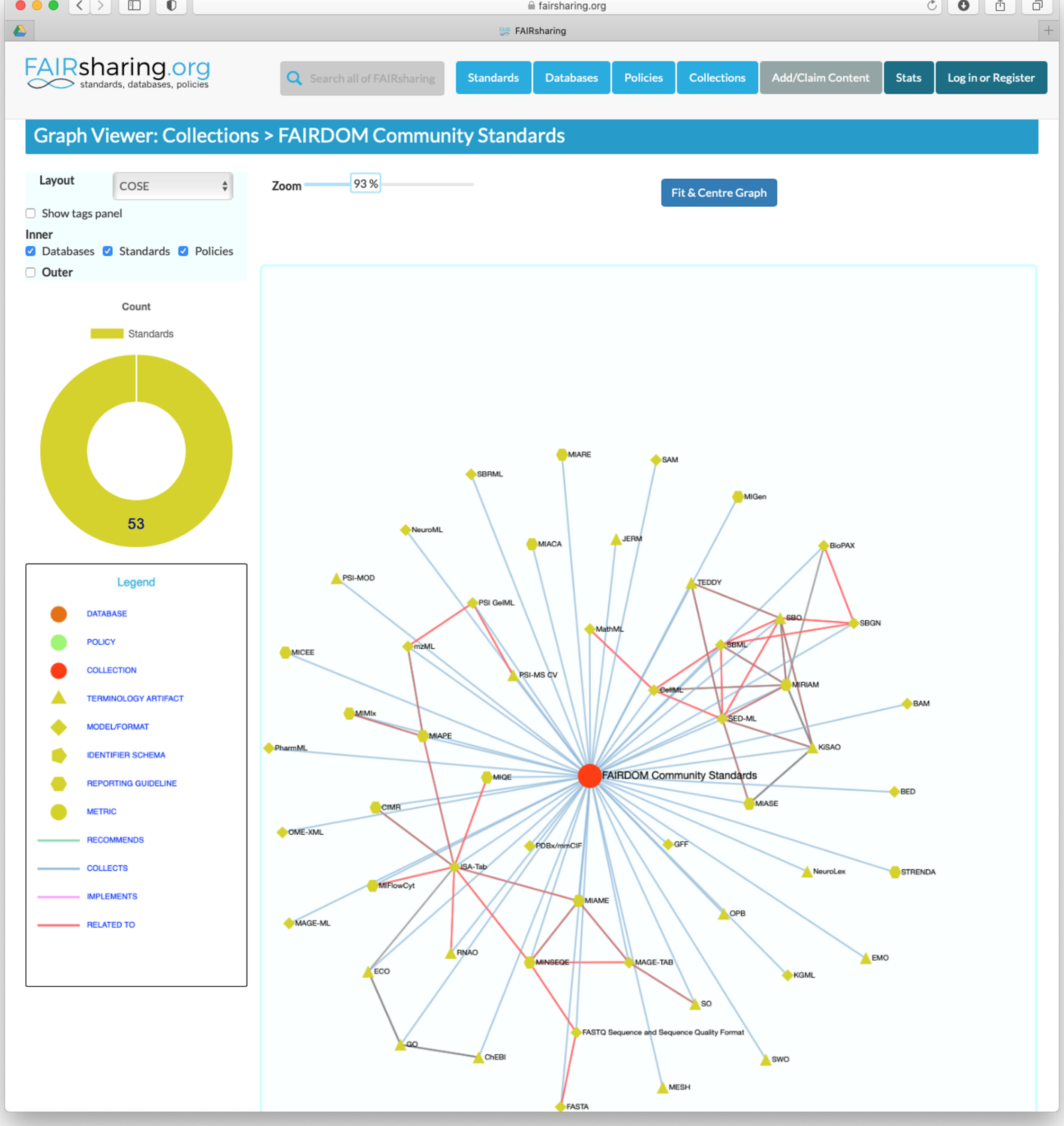
Registry	Name	Abbreviation	Type	Subject	Domain	Taxonomy	Related Database	Related Standard	Related Policy	In Collection/R
	American Type Culture Collection database	ATCC	Database	Epidemiology Life Science Virology	Cell Cell Culture Cell Line	All	None	None	Springer Nature - Nature Medicine - Availability of Data and materials Nature Publishing Group - Nature Genetics - Availability of Data and materials Nature Publishing Group - Nature Chemical Biology - Availability of Data and materials Springer Nature - Nature Photonics - Availability of Data and materials Nature Publishing Group - Nature Structural & Molecular Biology - Availability of Data and materials Plus 8 more...	COVID-19 Resourc RDA Covid-19 WG
	Australian New Zealand Clinical Trials Registry	ANZCTR	Database	Biomedical Science Clinical Studies Epidemiology Health Science Medical Informatics Plus 3 more...	Electronic Health Record	Homo sapiens	ClinicalTrials.gov ISRCTN Registry	None	None	COVID-19 Resourc RDA Covid-19 WG
	BBMRI-ERIC Directory	BBMRI-ERIC Directory	Database	Health Science Human Biology Life Science	Biobank Biological Sample Tissue	Homo sapiens	CRC-Cohort	None	None	EOSC-Life COVID-19 Resourc

Collections in Life Sciences

63 collections related to Life Science standards in FAIRsharing

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

https://fairsharing.org/collection/FAIRDOM



Graph Viewer: Collections > FAIRDOM Community Standards

Layout: COSE

Zoom: 93 %

☐ Show tags panel

Inner

☑ Databases ☑ Standards ☑ Policies

☐ Outer

Legend

● DATABASE

● POLICY

● COLLECTION

▲ TERMINOLOGY ARTIFACT

◆ MODEL/FORMAT

⬢ IDENTIFIER SCHEMA

⬢ REPORTING GUIDELINE

● METRIC

— RECOMMENDS

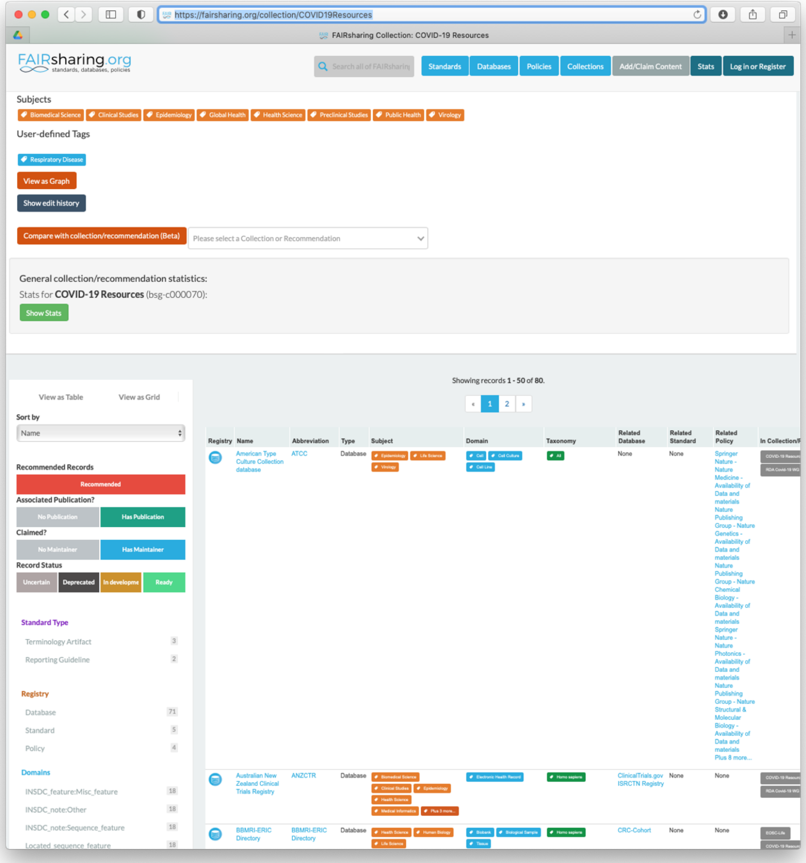
— COLLECTS

— IMPLEMENTS

— RELATED TO

Some collections are recent

Example 2: The *Covid-19* collection



https://fairsharing.org/collection/COVID19Resources

FAIRsharing Collection: COVID-19 Resources

FAIRsharing.org standards, databases, policies

Search all of FAIRsharing | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in or Register

Subjects

Biomedical Science · Clinical Studies · Epidemiology · Global Health · Health Science · Preclinical Studies · Public Health · Virology

User-defined Tags

Respiratory Disease

View as Graph

Show edit history

Compare with collection/recommendation (Beta) | Please select a Collection or Recommendation

General collection/recommendation statistics:

Stats for **COVID-19 Resources** (bsg-c000070):

Show Stats

View as Table | View as Grid

Sort by: Name

Recommended Records: Recommended

Associated Publication?: No Publication | Has Publication

Claimed?: No Maintainer | Has Maintainer

Record Status: Uncertain | Deprecated | In developme | Ready

Standard Type

Terminology Artifact	3
Reporting Guideline	2

Registry

Database	71
Standard	5
Policy	4

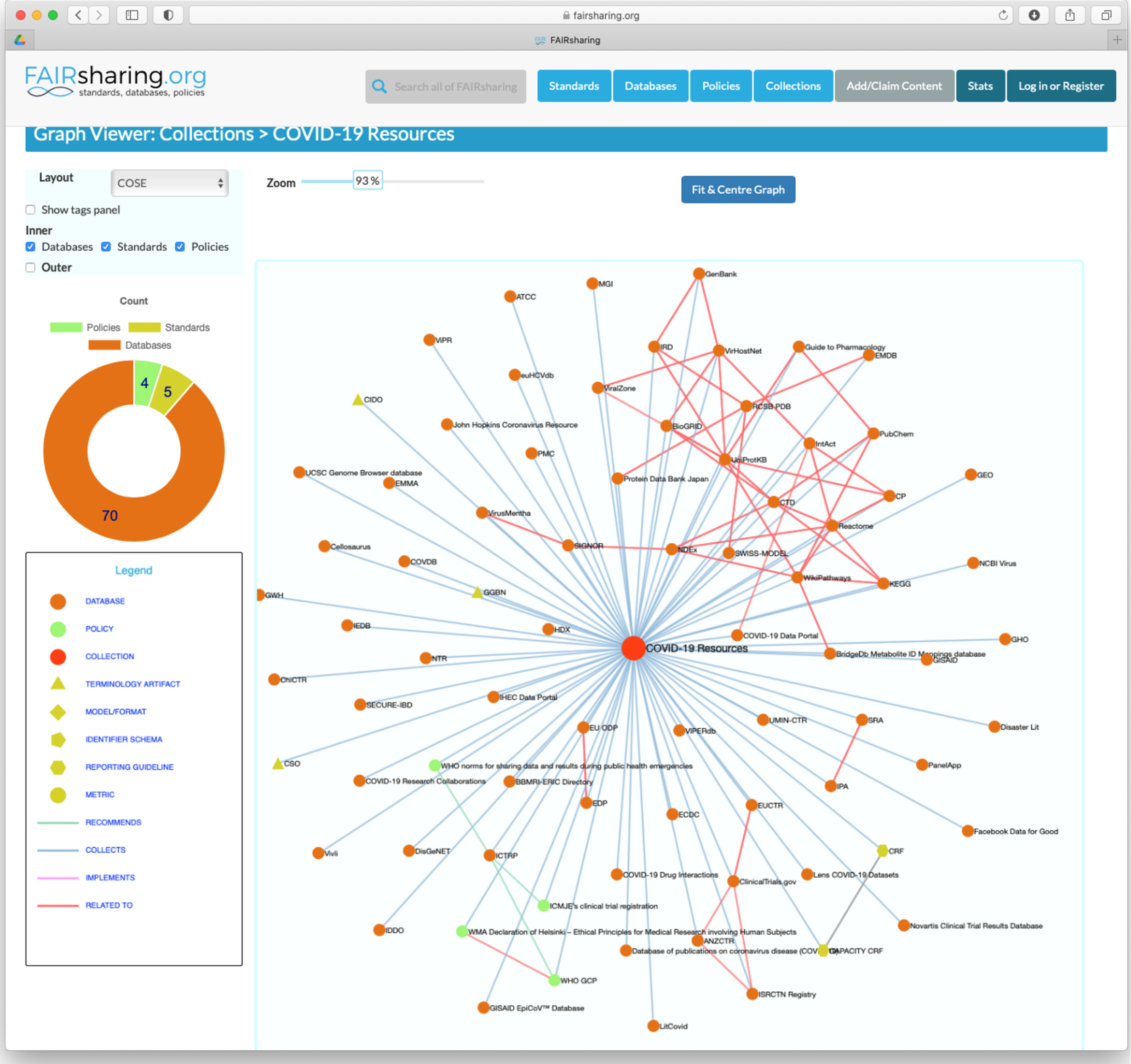
Domains

INSDC_feature:Misc_feature	18
INSDC_note:Other	18
INSDC_note:Sequence_feature	18
Located_sequence_feature	18

Showing records **1 - 50** of **80**.

« 1 2 »

Registry	Name	Abbreviation	Type	Subject	Domain	Taxonomy	Related Database	Related Standard	Related Policy	In Collection/R
	American Type Culture Collection database	ATCC	Database	Epidemiology Life Science Virology	Cell Cell Culture Cell Line	All	None	None	Springer Nature - Nature Medicine - Availability of Data and materials Nature Publishing Group - Nature Genetics - Availability of Data and materials Nature Publishing Group - Nature Chemical Biology - Availability of Data and materials Springer Nature - Nature Photonics - Availability of Data and materials Nature Publishing Group - Nature Structural & Molecular Biology - Availability of Data and materials Plus 8 more...	COVID-19 Resourc RDA Covid-19 WG
	Australian New Zealand Clinical Trials Registry	ANZCTR	Database	Biomedical Science Clinical Studies Epidemiology Health Science Medical Informatics Plus 3 more...	Electronic Health Record	Homo sapiens	ClinicalTrials.gov ISRCTN Registry	None	None	COVID-19 Resourc RDA Covid-19 WG
	BBMRI-ERIC Directory	BBMRI-ERIC Directory	Database	Health Science Human Biology Life Science	Biobank Biological Sample Tissue	Homo sapiens	CRC-Cohort	None	None	EOSC-Life COVID-19 Resourc



fairsharing.org

FAIRsharing

FAIRsharing.org standards, databases, policies

Search all of FAIRsharing | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in or Register

Graph Viewer: Collections > COVID-19 Resources

Layout: COSE

Zoom: 93 %

Fit & Centre Graph

□ Show tags panel

Inner

☑ Databases ☑ Standards ☑ Policies

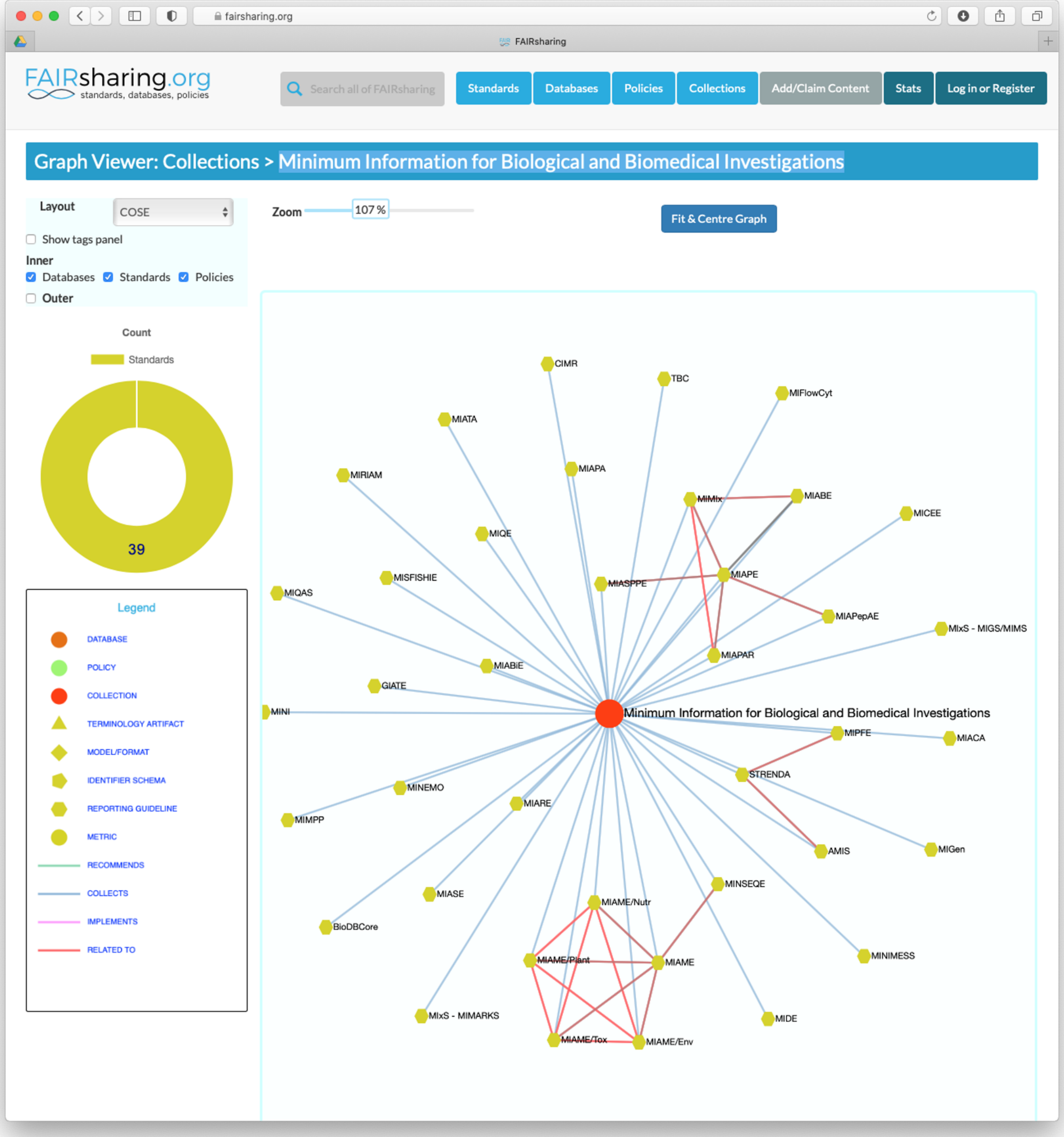
□ Outer

https://fairsharing.org/graph/#/collection/bsg-c000070

What about the minimum required metadata in biology?

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

https://fairsharing.org/collection/MIBBI



fairsharing.org

FAIRsharing

FAIRsharing.org standards, databases, policies

Search all of FAIRsharing | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in or Register

Graph Viewer: Collections > Minimum Information for Biological and Biomedical Investigations

Layout: COSE

Zoom: 107 %

Fit & Centre Graph

□ Show tags panel

Inner

☑ Databases ☑ Standards ☑ Policies

□ Outer

Count

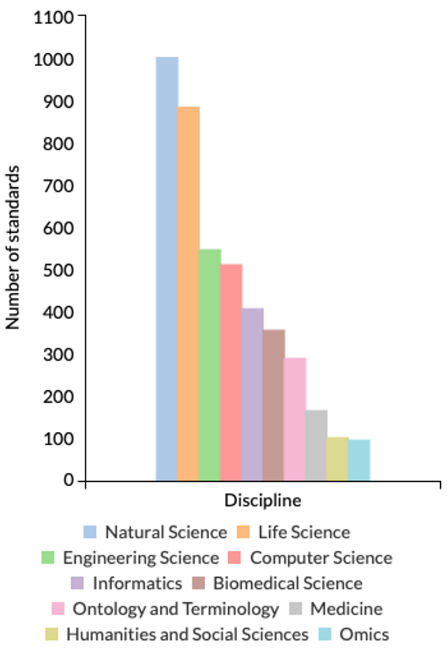
Standards: 39

Legend

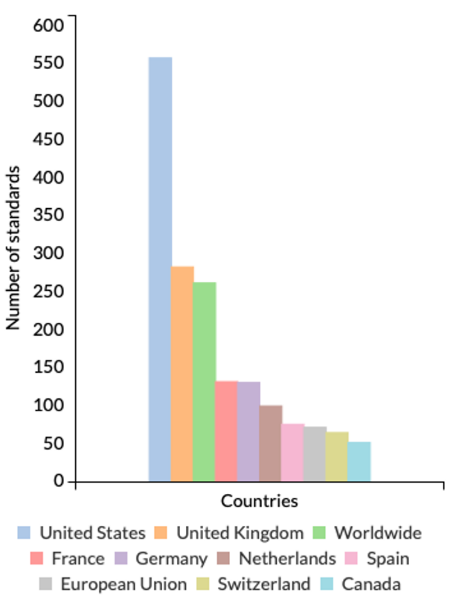
- ● DATABASE
- ● POLICY
- ● COLLECTION
- ▲ TERMINOLOGY ARTIFACT
- ◆ MODEL/FORMAT
- ⬟ IDENTIFIER SCHEMA
- ⬢ REPORTING GUIDELINE
- ● METRIC
- — RECOMMENDS
- — COLLECTS
- — IMPLEMENTS
- — RELATED TO

Summary statistics about standards

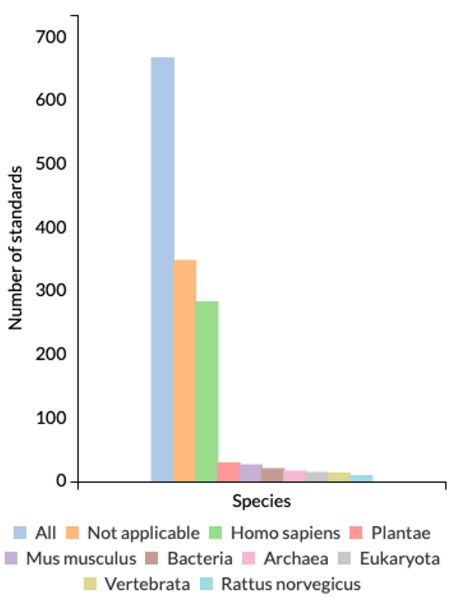
Top 10 disciplines covered by standards



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

Practice

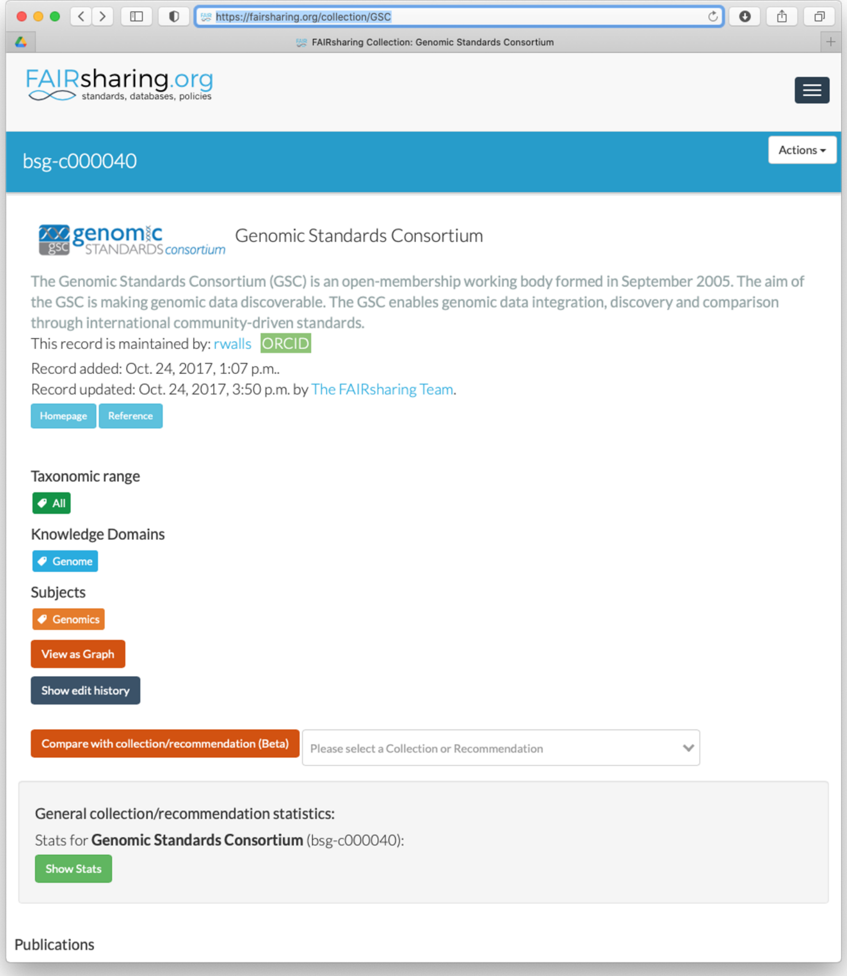
Find the *Genomics Standard Consortium (GSC) used by both ENA and SRA databases* in the FAIRsharing resource

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 ?

Source: https://gensc.org

The Genomics Standard Consortium (GSC)



https://fairsharing.org/collection/GSC

FAIRsharing Collection: Genomic Standards Consortium

FAIRsharing.org standards, databases, policies

bsg-c000040

Actions ▾

Genomic Standards Consortium

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.

This record is maintained by: rwalls ORCID

Record added: Oct. 24, 2017, 1:07 p.m..

Record updated: Oct. 24, 2017, 3:50 p.m. by The FAIRsharing Team.

Homepage Reference

Taxonomic range

All

Knowledge Domains

Genome

Subjects

Genomics

View as Graph

Show edit history

Compare with collection/recommendation (Beta) | Please select a Collection or Recommendation

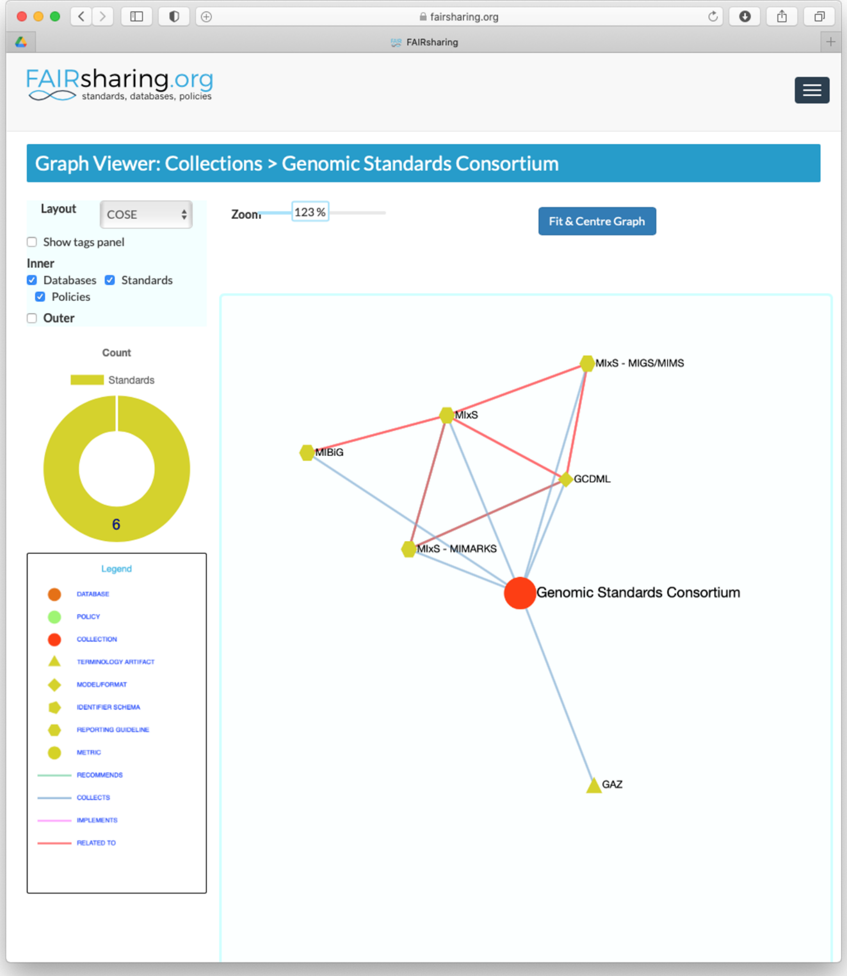
General collection/recommendation statistics:

Stats for **Genomic Standards Consortium** (bsg-c000040):

Show Stats

Publications

https://fairsharing.org/collection/GSC



fairsharing.org

FAIRsharing

FAIRsharing.org standards, databases, policies

Graph Viewer: Collections > Genomic Standards Consortium

Layout: COSE

Zoom: 123 %

Fit & Centre Graph

□ Show tags panel

Inner

☑ Databases ☑ Standards

☑ Policies

□ Outer

Count

Standards

6

Legend

● DATABASE

● POLICY

● COLLECTION

▲ TERMINOLOGY ARTIFACT

◆ MODEL/FORMAT

⬢ IDENTIFIER SCHEMA

⬢ REPORTING GUIDELINE

● METRIC

— RECOMMENDS

— COLLECTS

— IMPLEMENTS

— RELATED TO

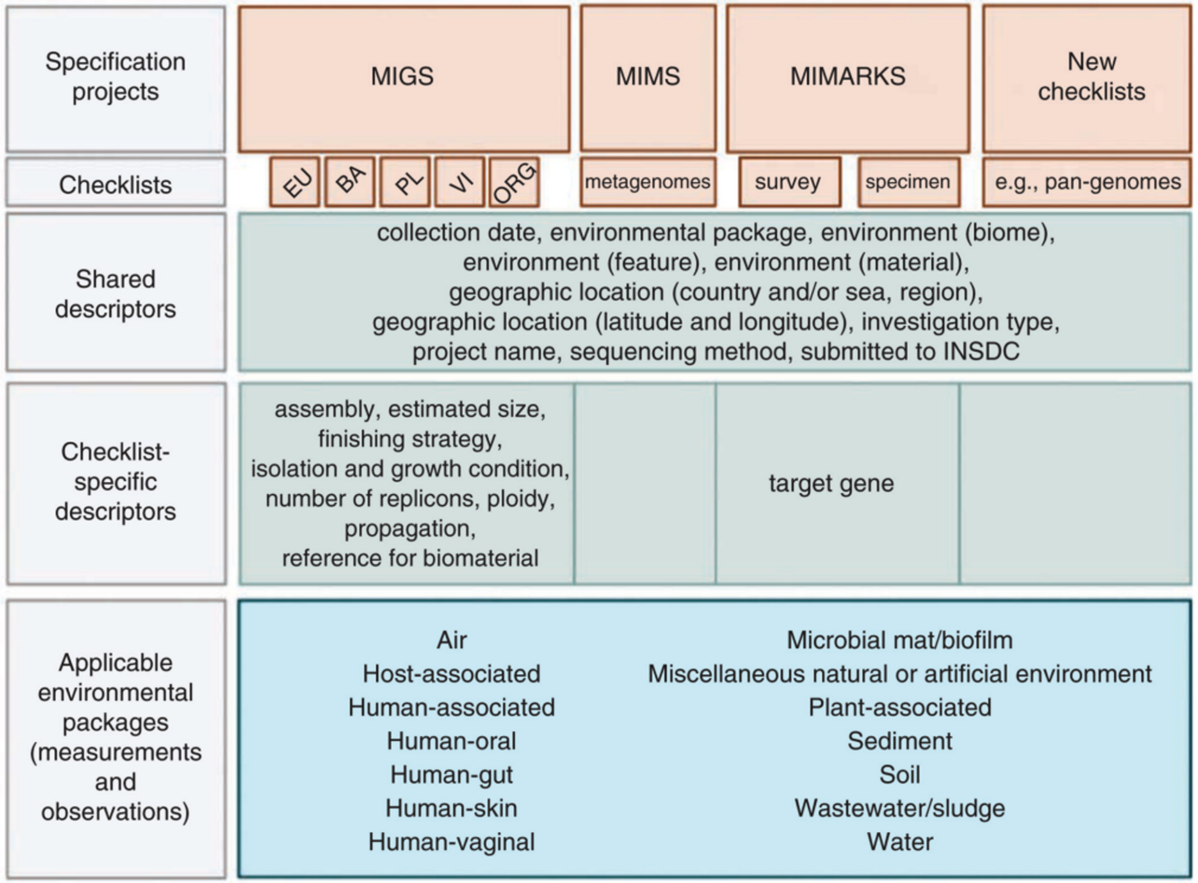
https://fairsharing.org/graph/#/collection/bsg-c000040

The Genomics Standard Consortium (GSC)

Genomic data integration, and comparison through international community-driven standards

Producer of the *Minimum Information Standards* (Checklists) used by ENA (EBI) & SRA (NCBI)

Ex: MIxS : Minimum Information about any (x) Sequence



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

Yilmaz et *al*, 2011

Source: https://gensc.org

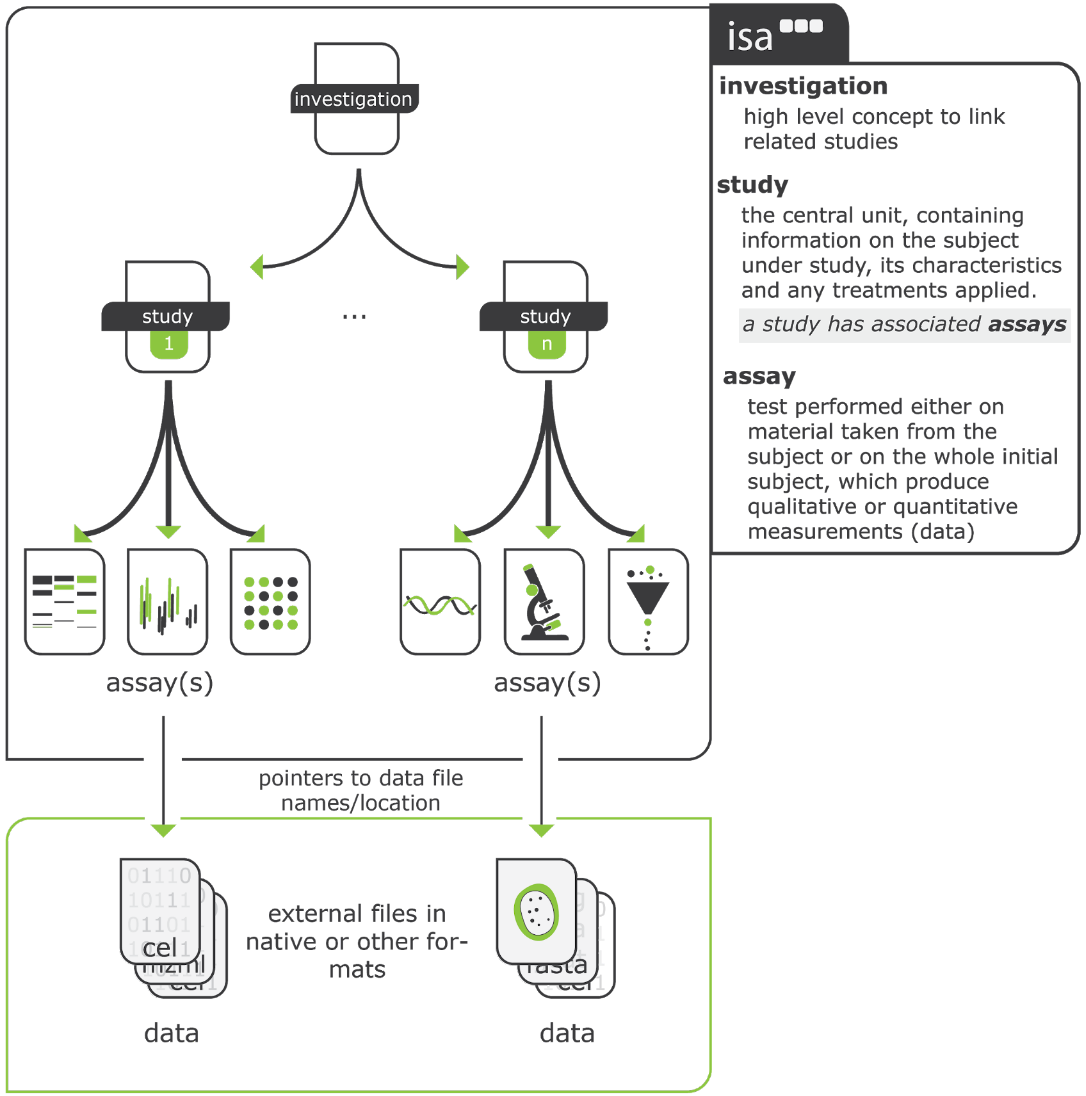
The ISA model

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


isa
investigation
study 1
...
study n
assay(s)
assay(s)
pointers to data file names/location
cel
mzml
cel
fasta
cel
external files in native or other formats
data
data

investigation
high level concept to link related studies

study
the central unit, containing information on the subject under study, its characteristics and any treatments applied.
*a study has associated **assays***

assay
test performed either on material taken from the subject or on the whole initial subject, which produce qualitative or quantitative measurements (data)

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

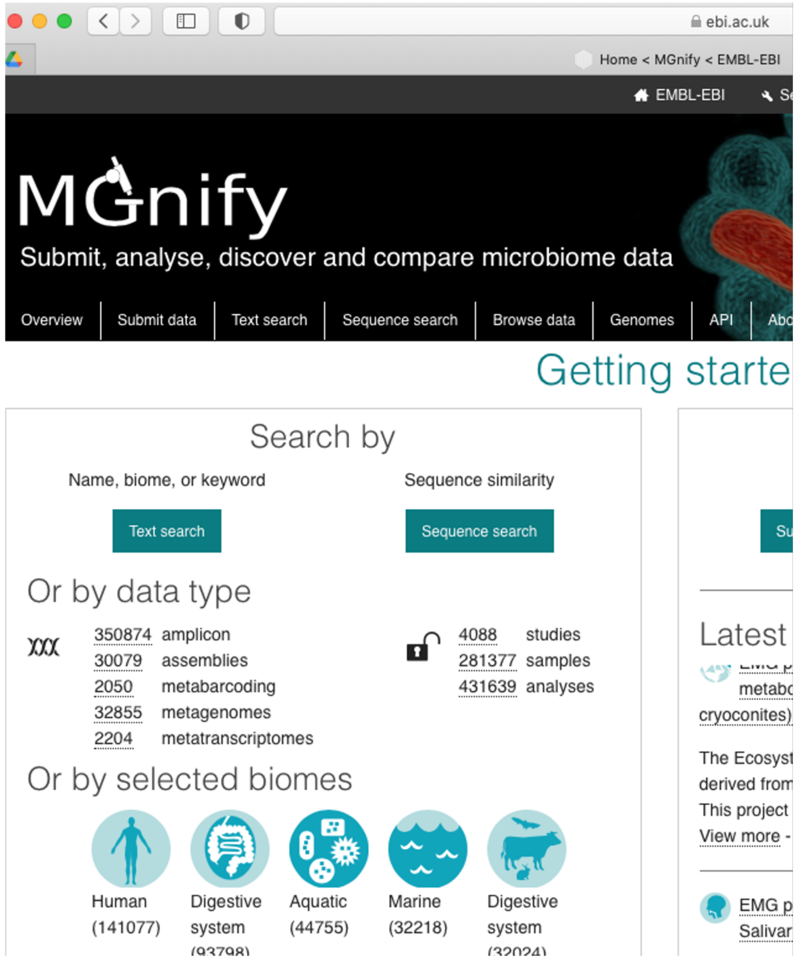
European Nucleotide Archive (ENA) submission

ENA
European Nucleotide Archive

EMBL-EBI

Why do I need to submit my data and metadata to ENA ?

- Open Science and reproducibility of experiments
- 3rd party access
- Archival
- Publication
- Analyses, example: [MGinfy]



ebi.ac.uk

Home < MGnify < EMBL-EBI

EMBL-EBI

MGnify

Submit, analyse, discover and compare microbiome data

Overview | Submit data | Text search | Sequence search | Browse data | Genomes | API | Abo

Getting starte

Search by

Name, biome, or keyword

Text search

Sequence similarity

Sequence search

Or by data type

350874	amplicon
30079	assemblies
2050	metabarcoding
32855	metagenomes
2204	metatranscriptomes

4088	studies
281377	samples
431639	analyses

Or by selected biomes

Human (141077)

Digestive system (93798)

Aquatic (44755)

Marine (32218)

Digestive system (32024)

Su

Latest

metabo

cryoconites)

The Ecosyst

derived from

This project

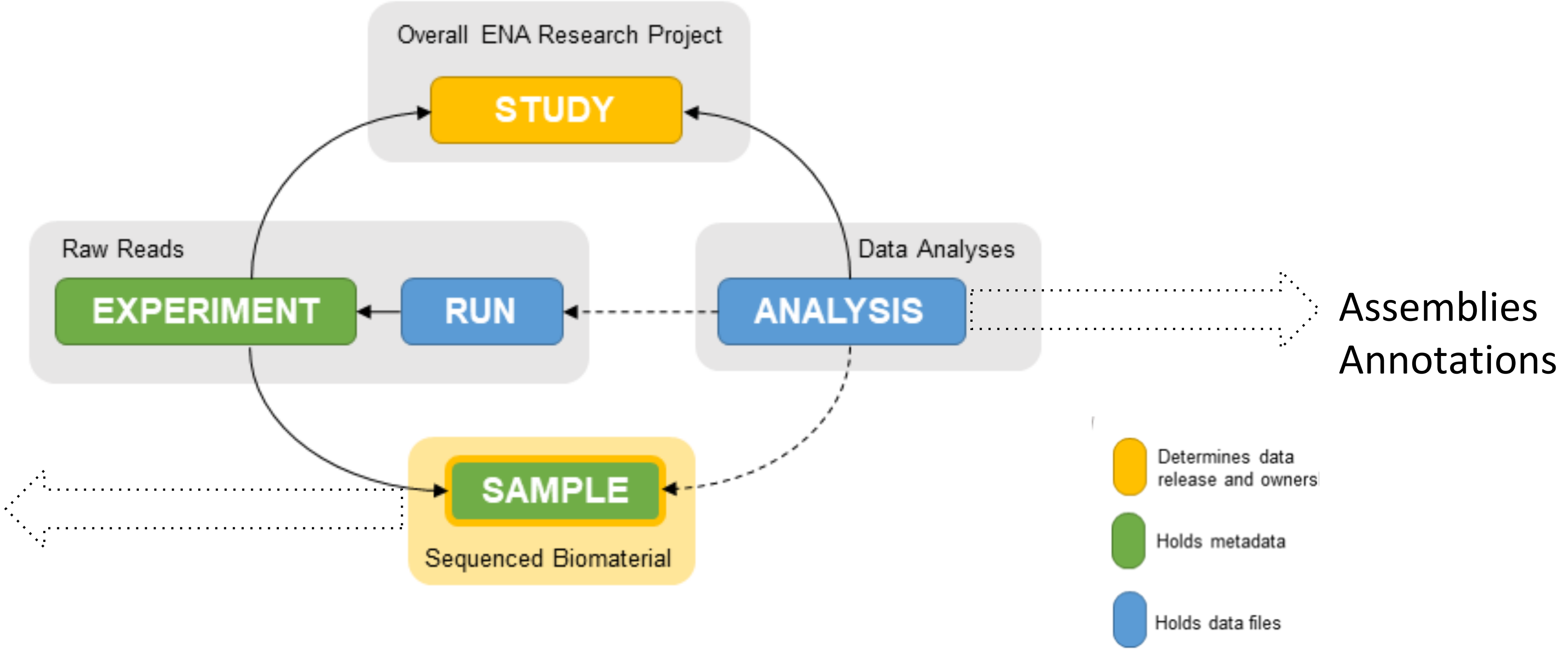
View more -

EMG p

Salivar

The ENA metadata model

ISA compliant !



Assemblies
Annotations

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

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

THE ENA Checklists

- 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

Data brokering at IFB

Why developing data brokering at IFB?

Observations:

- Submissions are often complex and difficult to perform by individual teams
- Metadata are often poorly understood resulting in incomplete, redundant and inconsistent submissions
- ENA asks that IFB becomes the French national broker

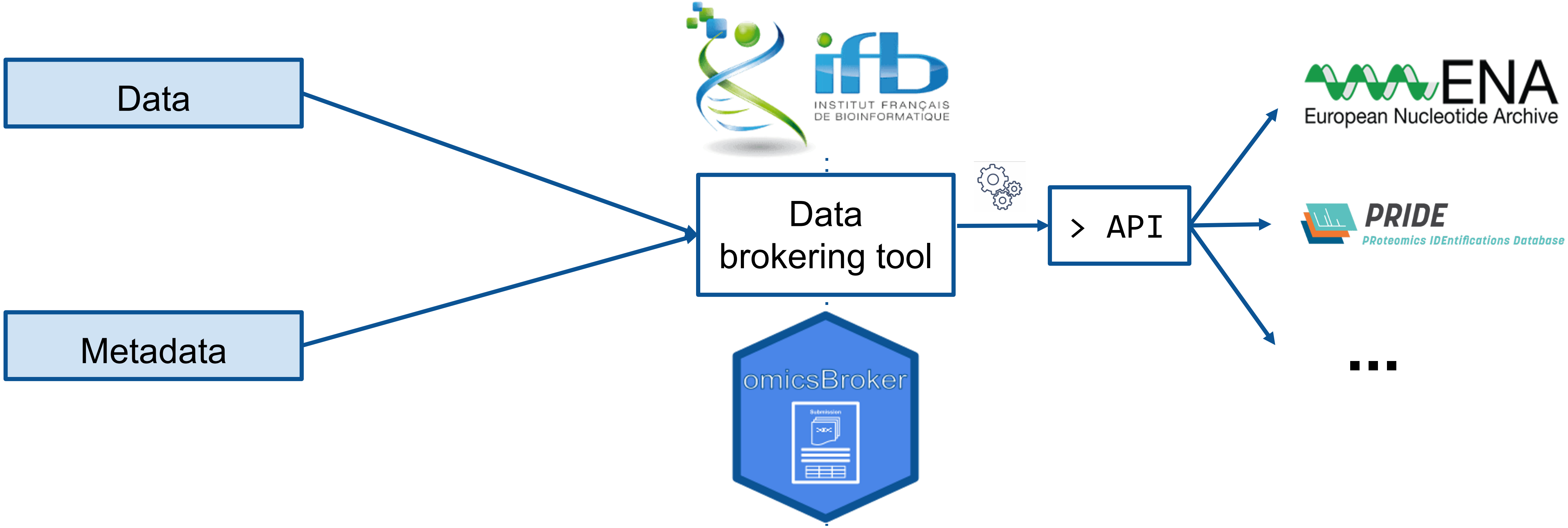
Main idea: offer a national service of **data brokering at IFB** to simplify and rationalize data exchange between international resources and the french Elixir node IFB.

Brokering include 3 types of activities: tools development, training and support to users

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



INSTITUT FRANÇAIS DE BIOINFORMATIQUE
Data
Metadata
Data brokering tool
omicsBroker
Submission
> API
ENA
European Nucleotide Archive
PRIDE
PRoteomics IDEntifications Database
...

The omicsBroker tool

- **omicsBroker** is a tool to easily annotate and submit **omics data** to **international repositories**
- For now, only available as a **PROTOTYPE**
  - based on **R Shiny** technology
  - allowing to test submission of genomic and transcriptomic samples and reads to **ENA test instance**
- The final tool will be developed using Django technology and will **manage data and metadata from different sources** to make submission to international resources easier

https://github.com/IFB-ElixirFr/omicsBroker

Practice

Use omicsBroker prototype (134.158.247.213:443) to test submission of samples to ENA

Use information of the corresponding DMP to associate relevant metadata to data https://dmp.opidor.fr

3 groups

- bacterial genome (IFB_Training_salivarius)
- plant transcriptome (IFB Training : Sars-CoV-2)
- SARS-Cov2 genome (IFB_Training_plant)

https://ifb-elixirfr.github.io/IFB-FAIR-data-training/sequences/module3_sequence3_tp.html

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
European National Archive Checklists	ENA Checklists	https://www.ebi.ac.uk/ena/browser/checklists
European National Archive submission documentation	ENA submission guide	https://ena-docs.readthedocs.io/en/latest/submit/general-guide.html
A prototype to test submission of samples and DNAseq or RNAseq reads to ENA	omicsBroker	https://github.com/IFB-ElixirFr/omicsBroker

Thanks

Paulette Lieby

Jean-François Dufayard

Frédéric de Lamotte