DMP and Open Data training

Session 3: Metadata

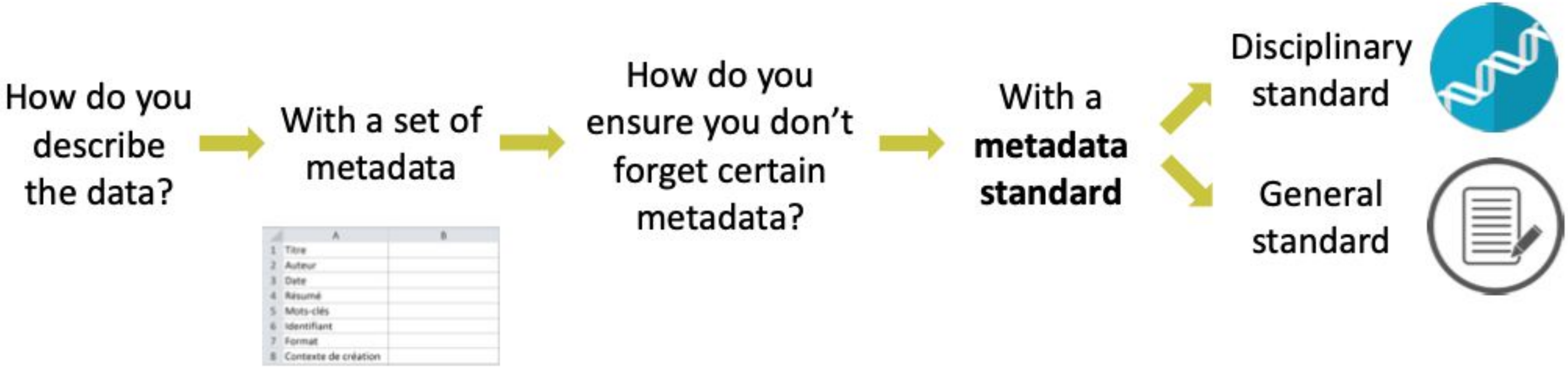
Life science standards and ENA submission

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: Do you know any standard in life sciences ?

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

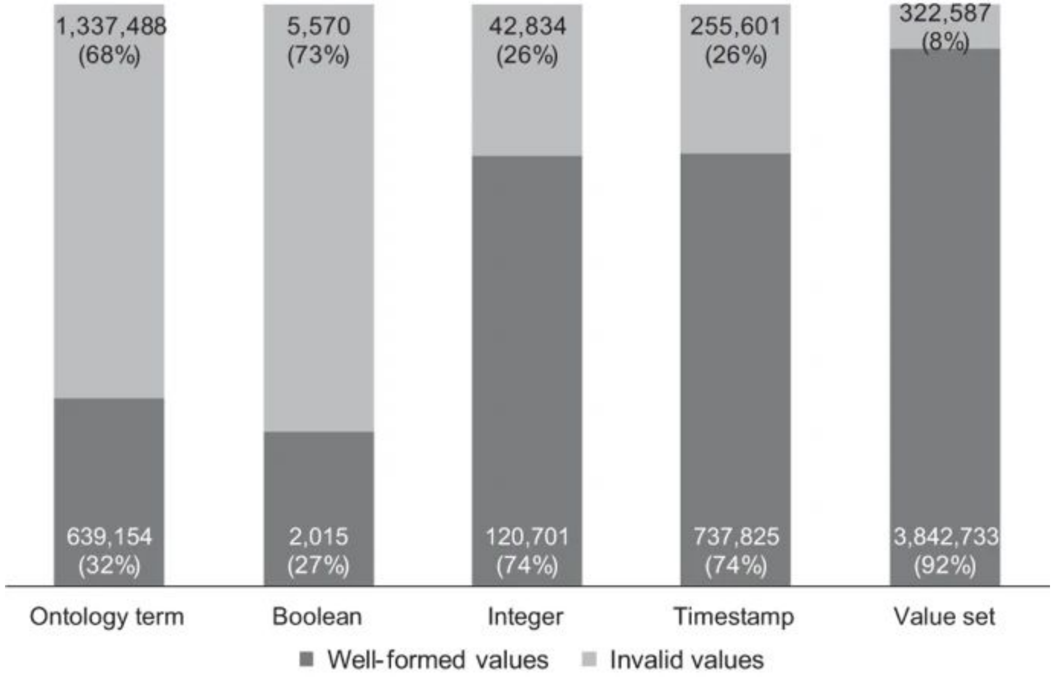
https://scrumblr.ethibox.fr/standard

Metadata exhibit questionable quality in biology

Submission in public resources is often a complex task

Submission procedures are heterogeneous

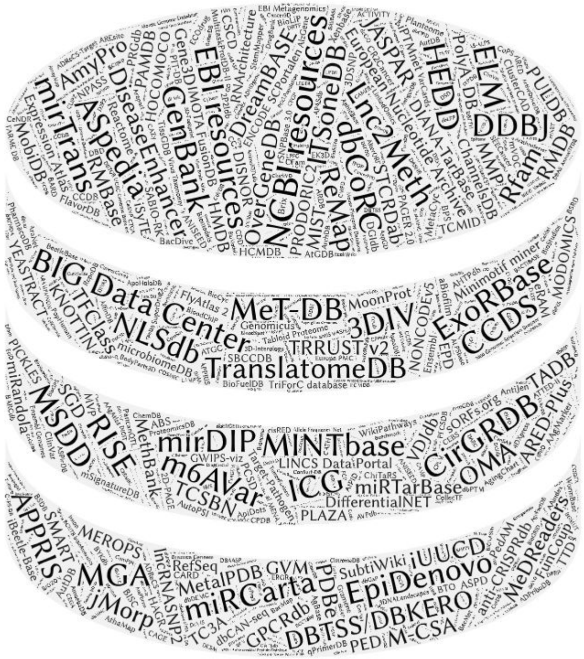
Metadata are often incomplete, inconsistent, redundant or not enough informative



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

Standard adoption and perenity

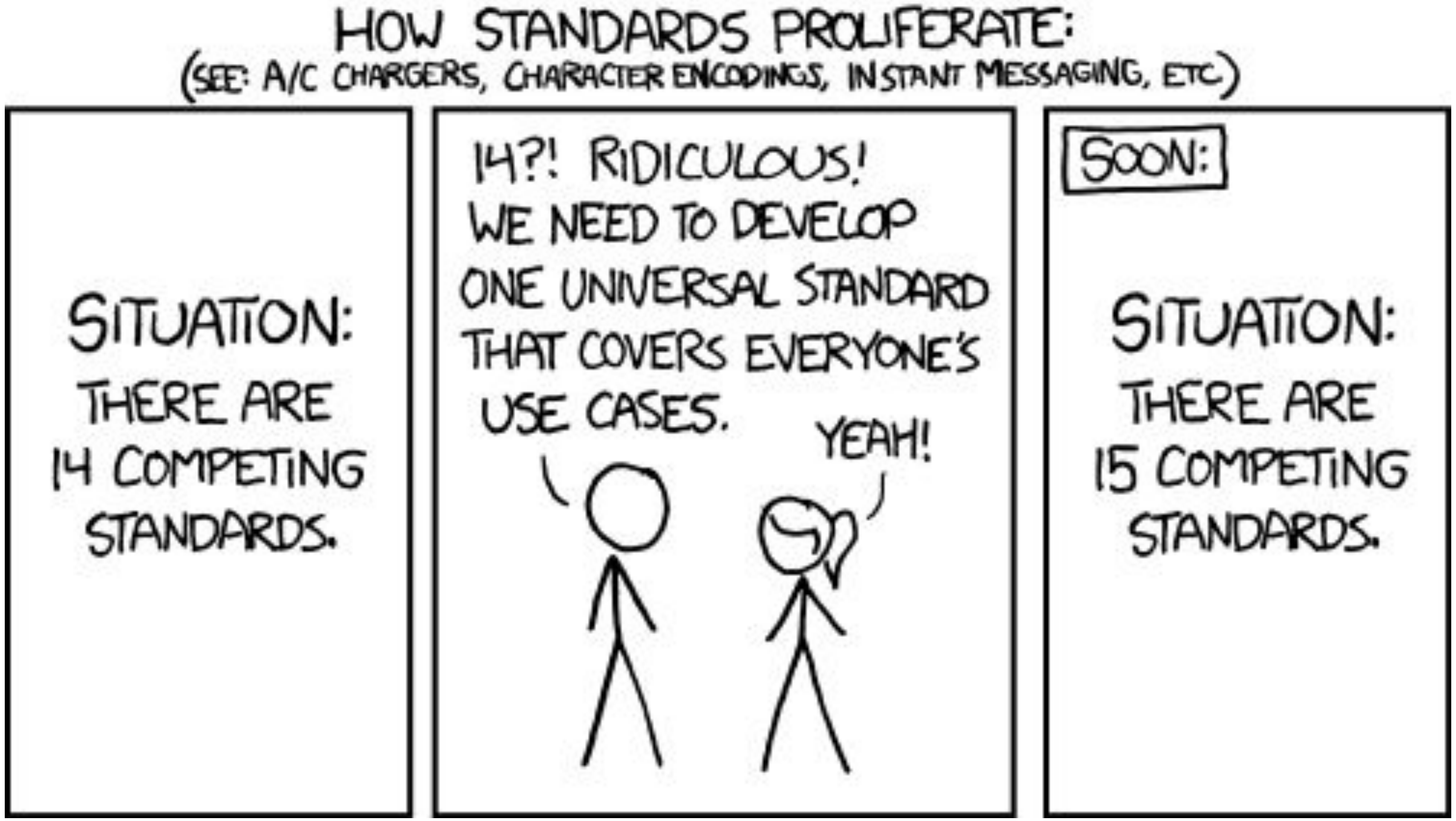
- There are thousand of databases, softwares and resources in biology with **unequal level of standard adoption**
- Is is not always easy for Life scientists and bioinformaticians to identify and use the most appropriate standards



1641 databases in NAR Database 2021

Rigden *et al*, 2021

Standard adoption and perenity


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

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

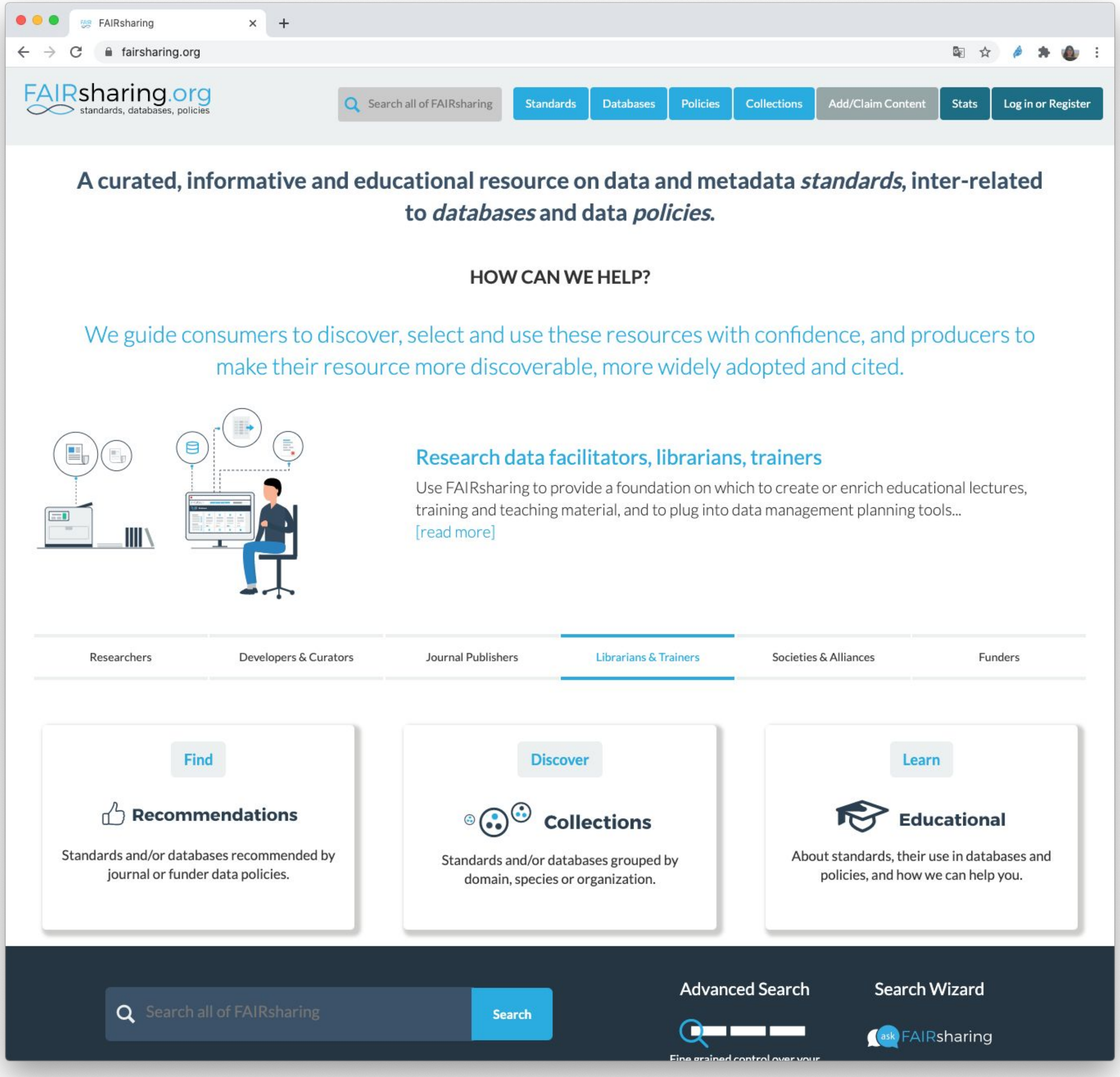
How do I find the standard I need?

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



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.

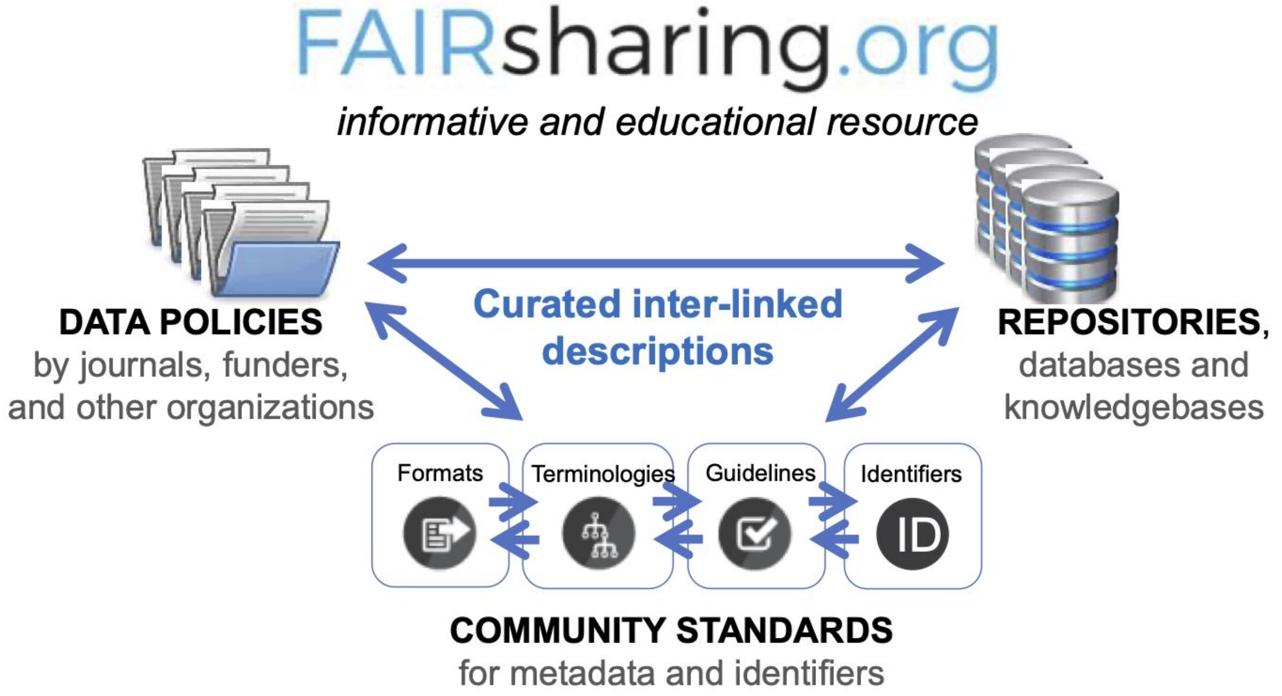
https://fairsharing.org

The FAIRsharing portal

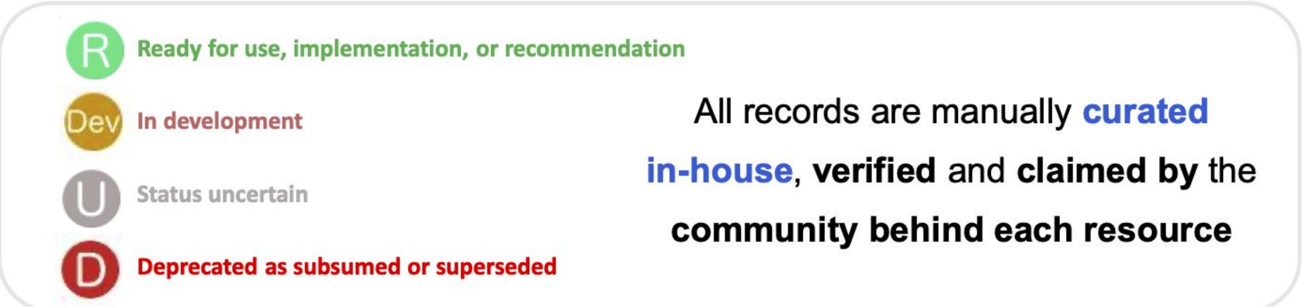
Citable *DOI* for all records

Accessible via *API* or *web interface*

Curation


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

RECORD STATUS



R Ready for use, implementation, or recommendation

Dev In development

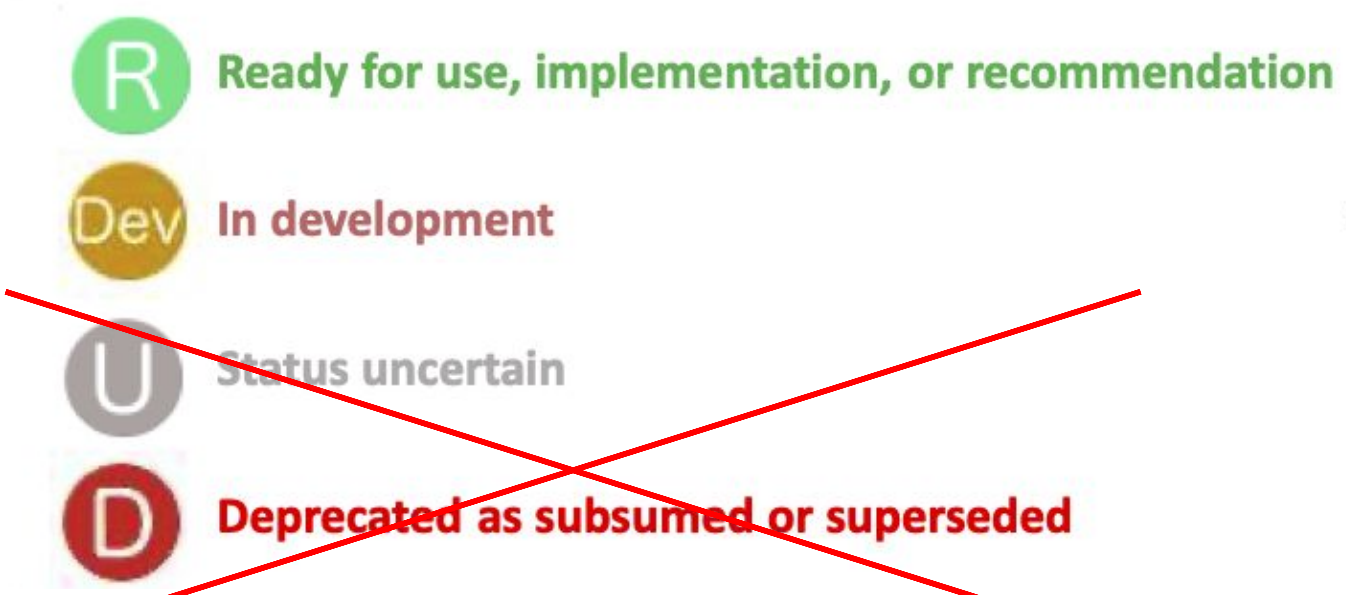
U Status uncertain

D Deprecated as subsumed or superseded

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

https://fairsharing.org

The FAIRsharing portal: record status



R Ready for use, implementation, or recommendation

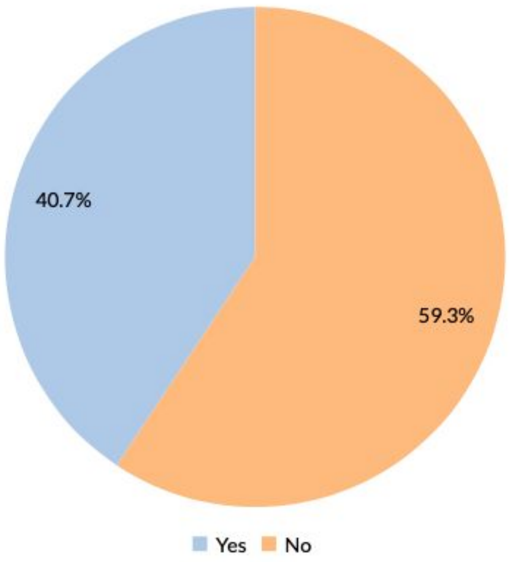
Dev In development

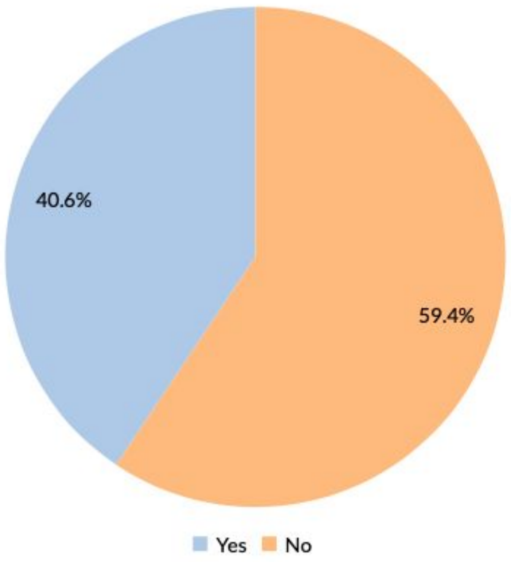
~~U Status uncertain~~

~~D Deprecated as subsumed or superseded~~

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

Standard maintenance is a key point


Standard records that have maintainers


Standards that have a publication

59.3 % of standards have no maintainer

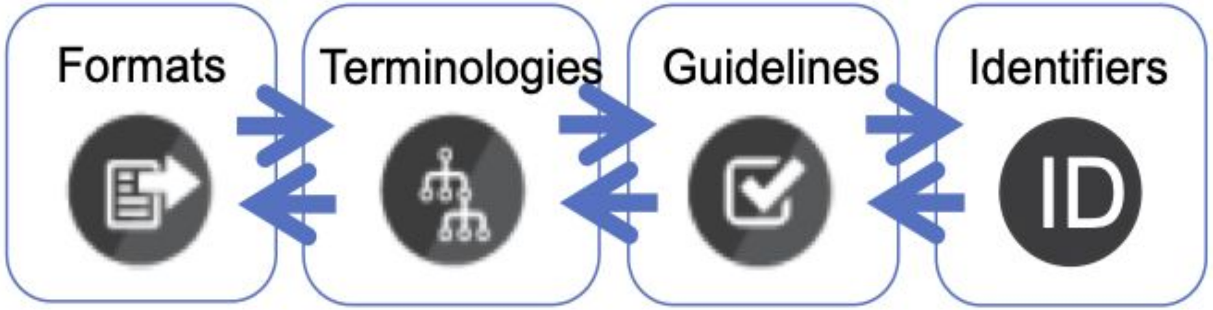
59.4% of standard has no publication

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

FASTQ

FASTA

SBML

GFF

MIAME

Newick

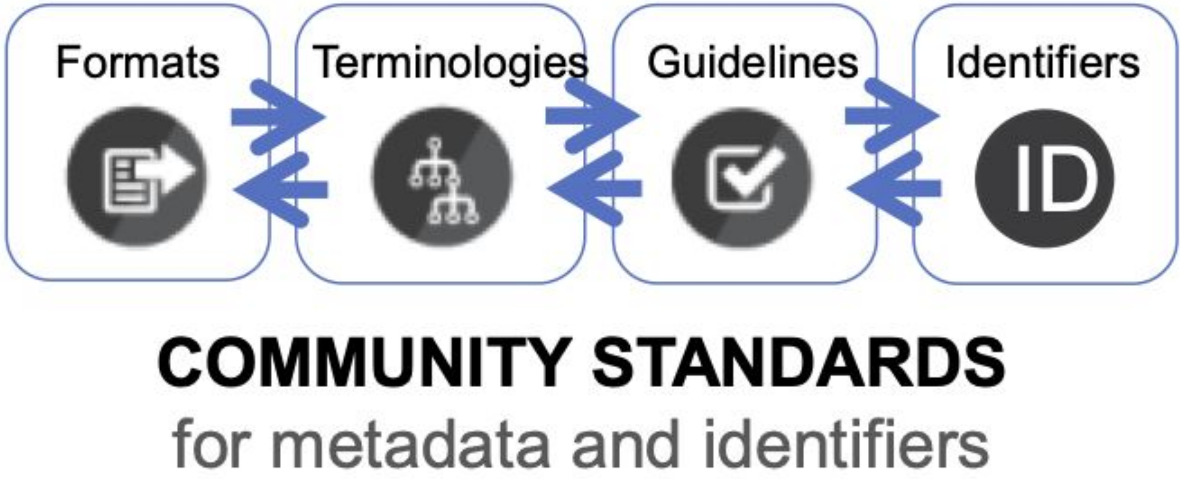
EC number

BAM

VCF

MINSEQE

n=432 n=539 n=180 n=18



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

genomic STANDARDS consortium gsc

isatools

GENEONTOLOGY Unifying Biology

Crop Ontology for agricultural data

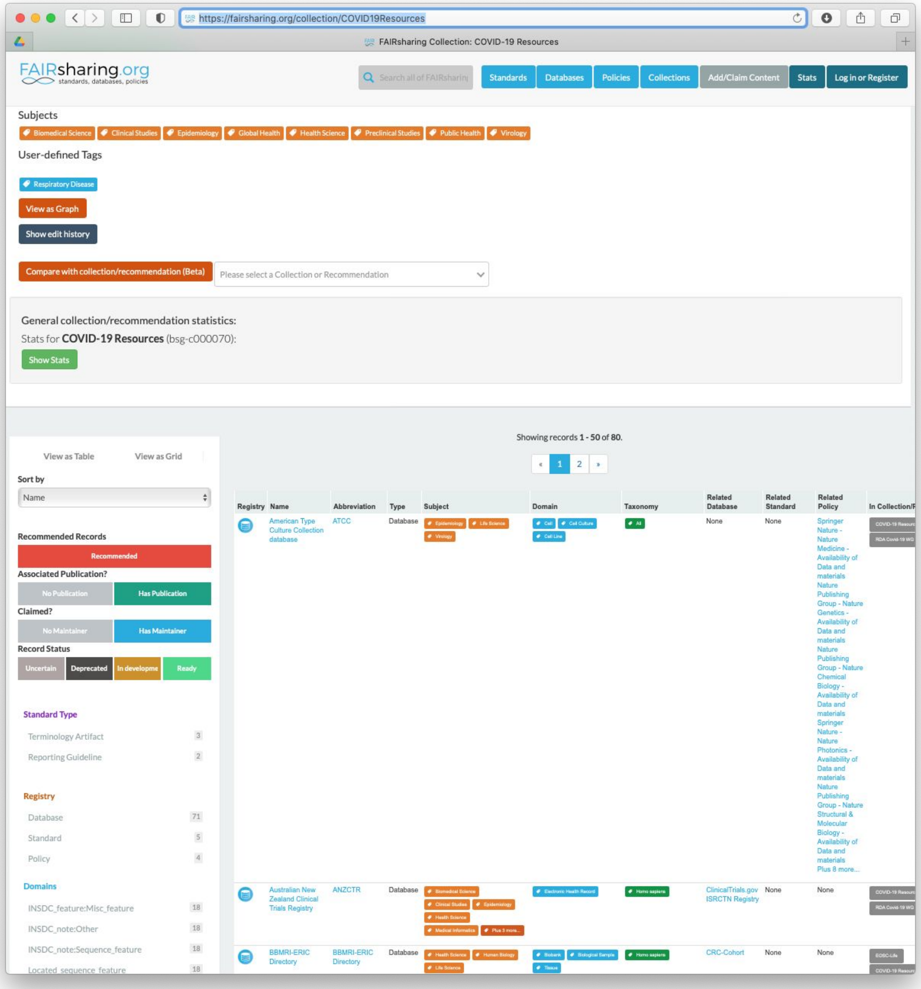
DISEASE ONTOLOGY

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):

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 »

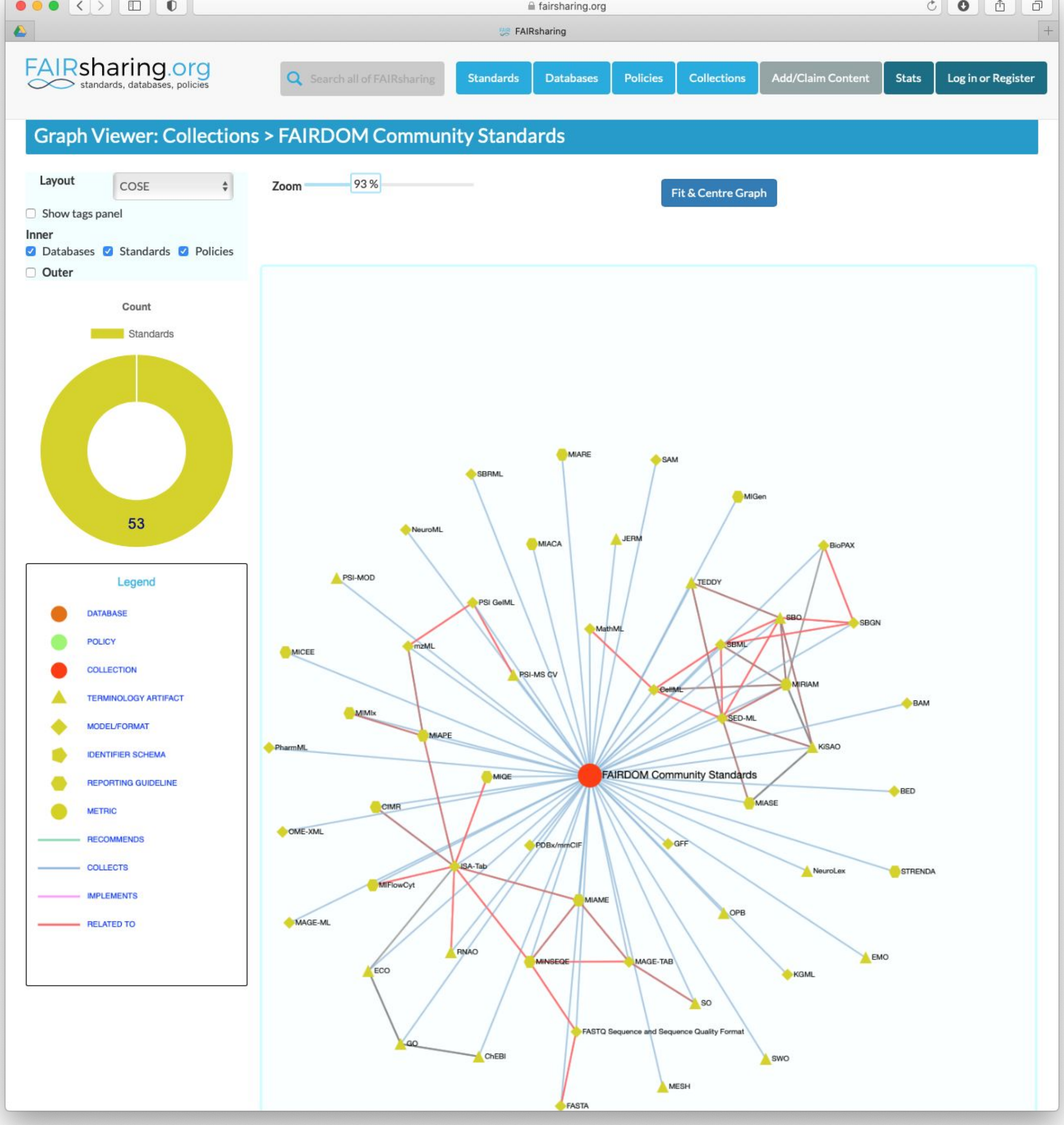
Registry	Name	Abbreviation	Type	Subject	Domain	Taxonomy	Related Database	Related Standard	Related Policy	In Collection/R
	American Type Culture Collection database	ATCC	Database	Epidemiology Life Science Virology	Cell Cell Culture Cell Line	All	None	None	Springer Nature - Nature Medicine - Availability 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



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 > FAIRDOM Community Standards

Layout: COSE

Zoom: 93 %

Fit & Centre Graph

☐ Show tags panel

Inner

☑ Databases ☑ Standards ☑ Policies

☐ Outer

Count

▬ Standards

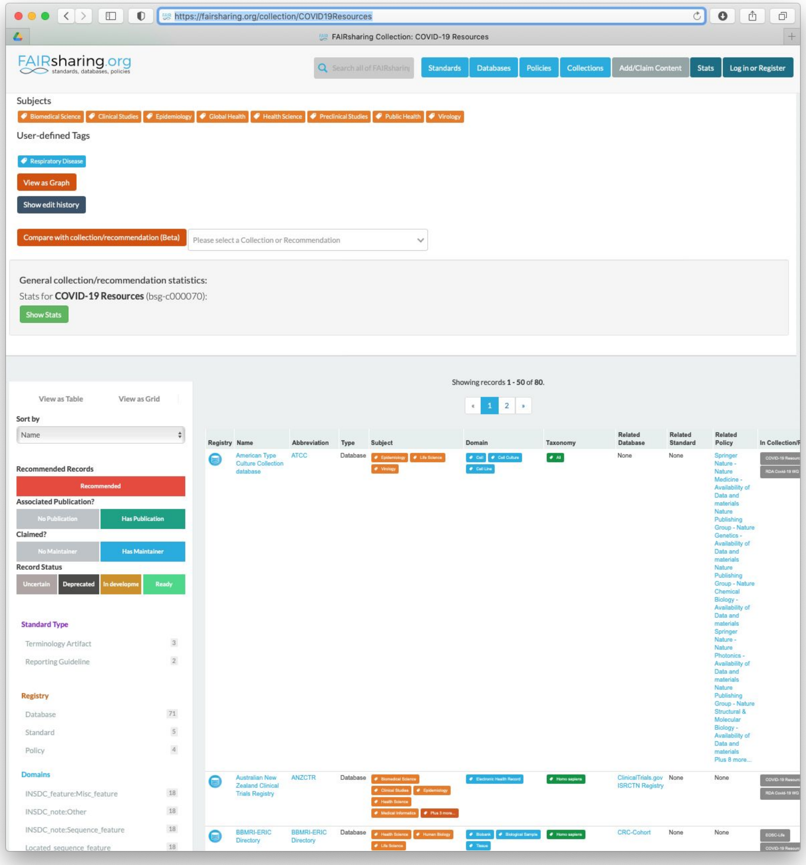
53

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

Subjects

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):

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

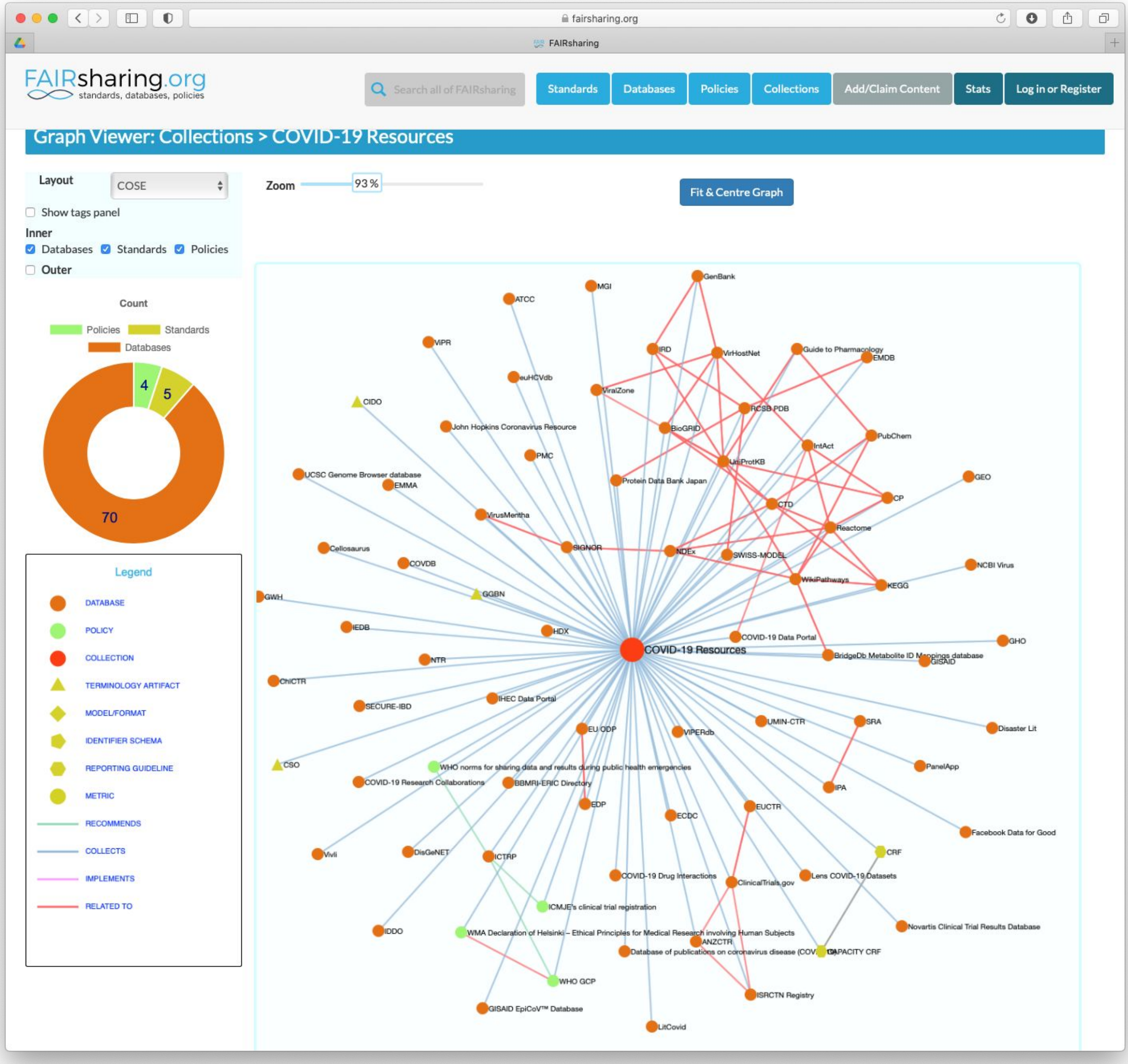
Domains

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

Graph Viewer: Collections > COVID-19 Resources

Layout: COSE

Zoom: 93 %

□ Show tags panel

Inner

☑ Databases ☑ Standards ☑ Policies

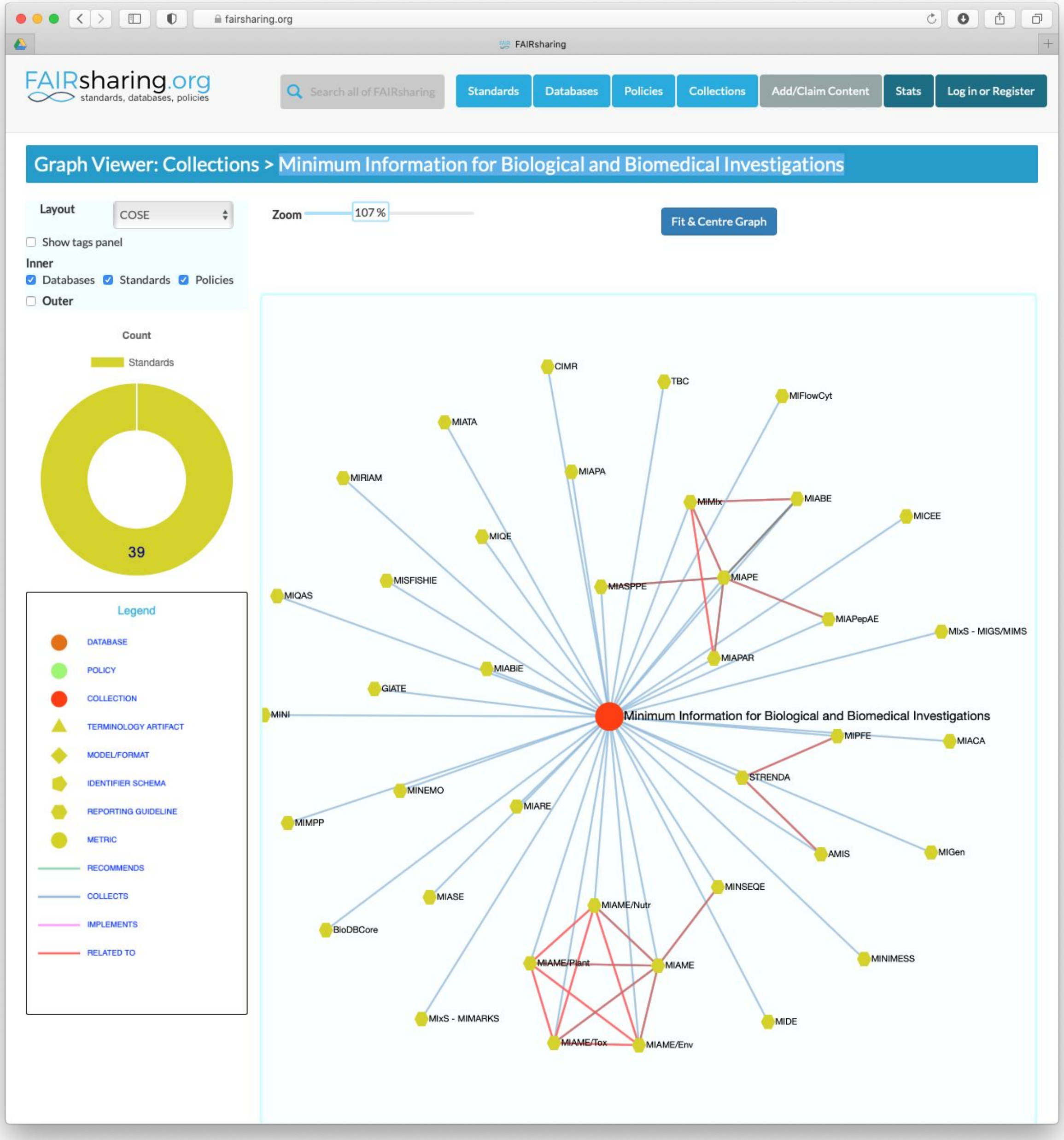
□ Outer

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

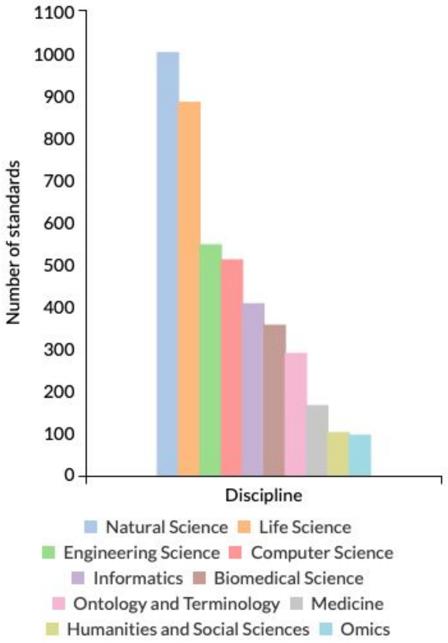
Standards: 39

Legend

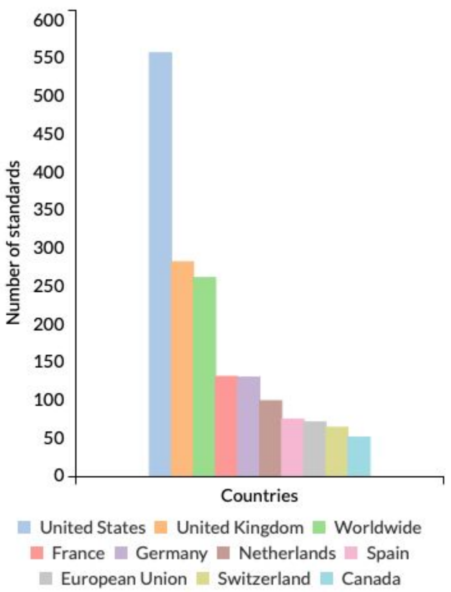
- ● DATABASE
- ● POLICY
- ● COLLECTION
- ▲ TERMINOLOGY ARTIFACT
- ◆ MODEL/FORMAT
- ⬢ IDENTIFIER SCHEMA
- ⬢ REPORTING GUIDELINE
- ● METRIC
- — RECOMMENDS
- — COLLECTS
- — IMPLEMENTS
- — RELATED TO

Summary statistics about standards

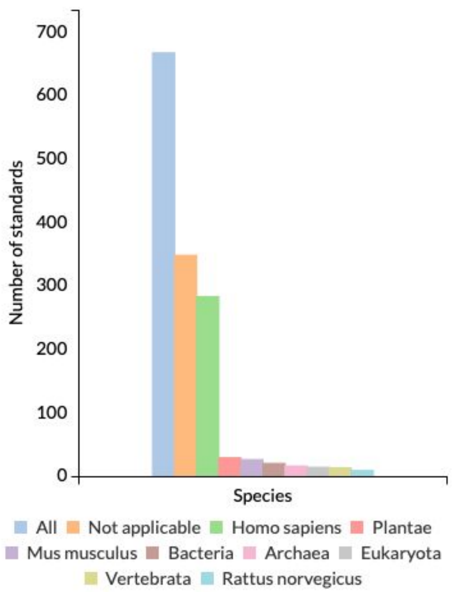
Top 10 disciplines covered by standards



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

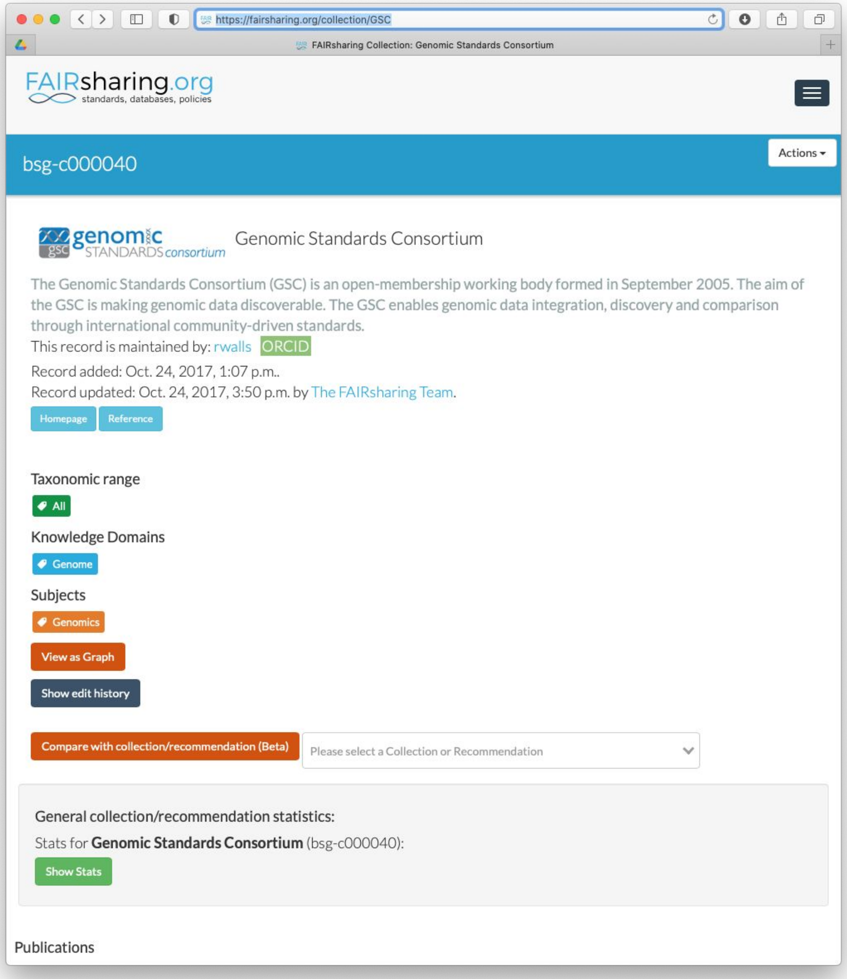
Find the *Genomic Standards Consortium (GSC) used by both ENA and SRA databases* in the **FAIRsharing collections**

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 ?

Source: https://gensc.org

The Genomic Standards 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

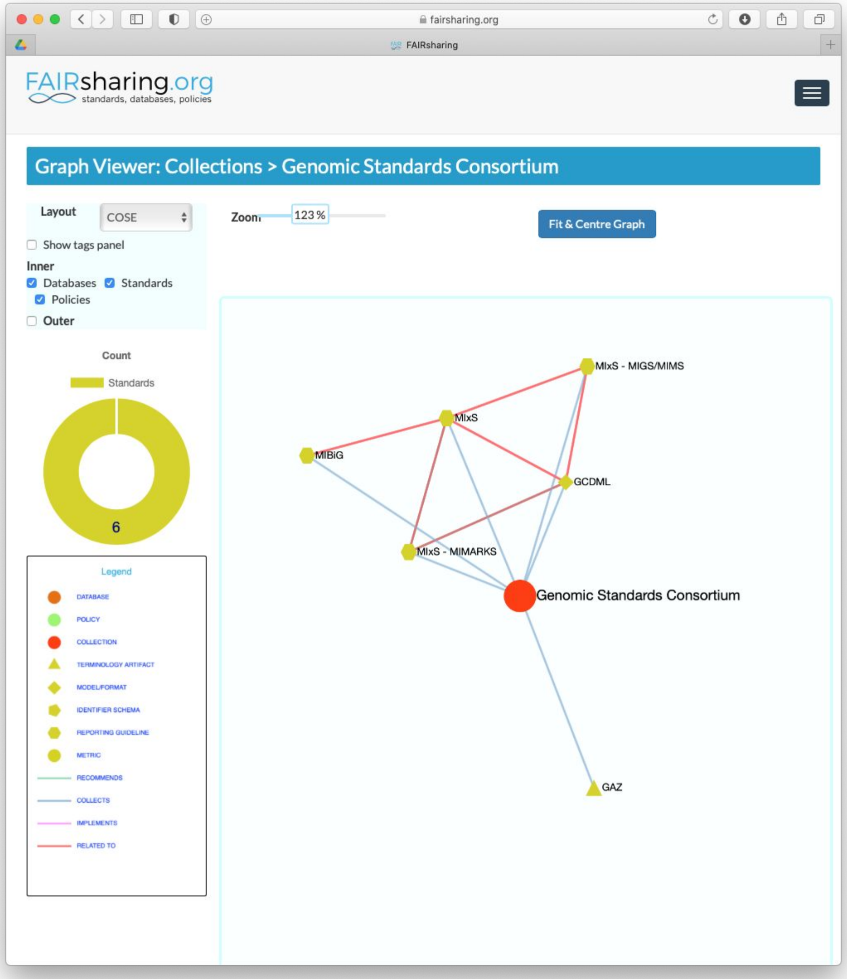
Compare with collection/recommendation (Beta) | Please select a Collection or Recommendation

General collection/recommendation statistics:

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

Show Stats

Publications



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

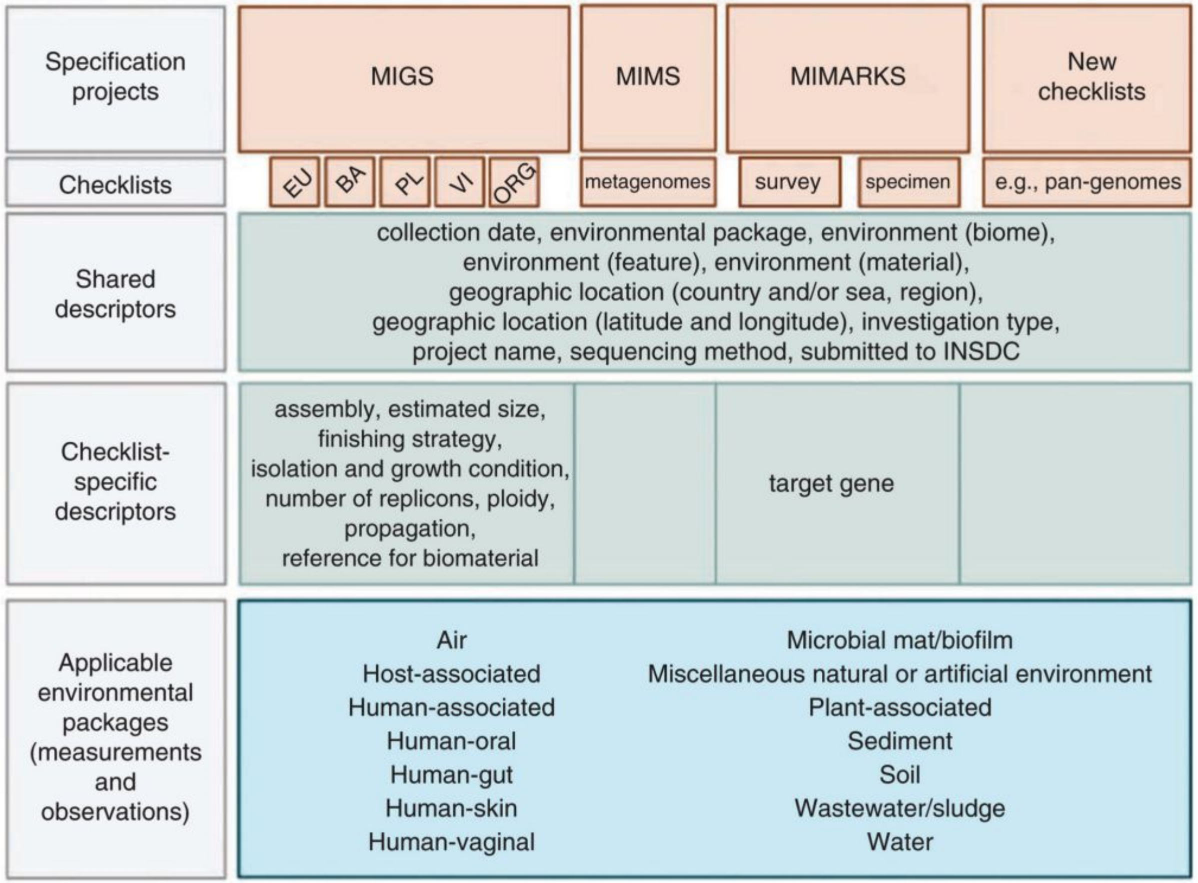
- ● DATABASE
- ● POLICY
- ● COLLECTION
- ▲ TERMINOLOGY ARTIFACT
- ◆ MODEL/FORMAT
- ⬟ IDENTIFIER SCHEMA
- ⬢ REPORTING GUIDELINE
- ● METRIC
- — RECOMMENDS
- — COLLECTS
- — IMPLEMENTS
- — RELATED TO

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

The 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



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

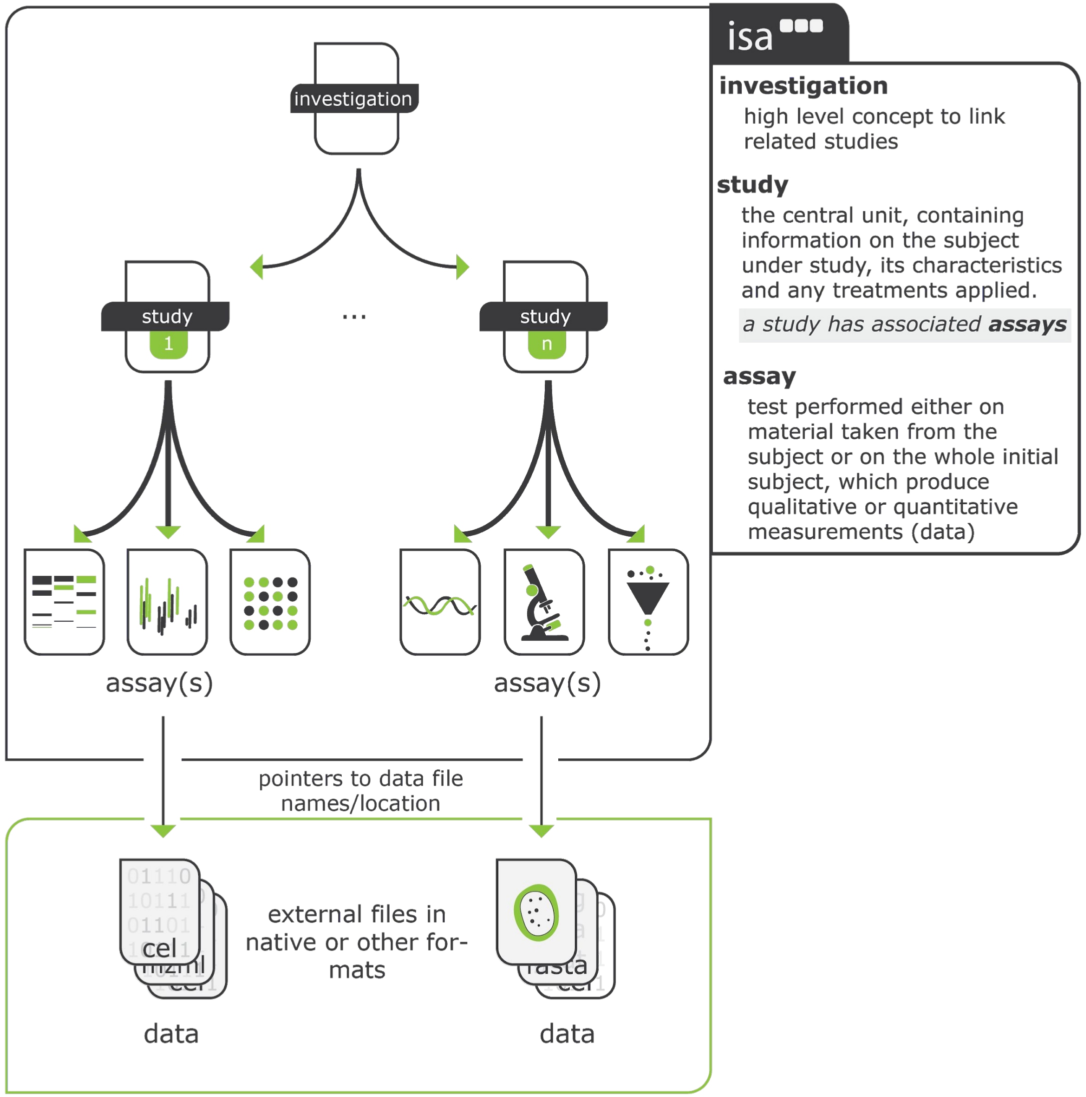
The ISA model

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
external files in native or other for-mats
cel
mzml
fasta
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

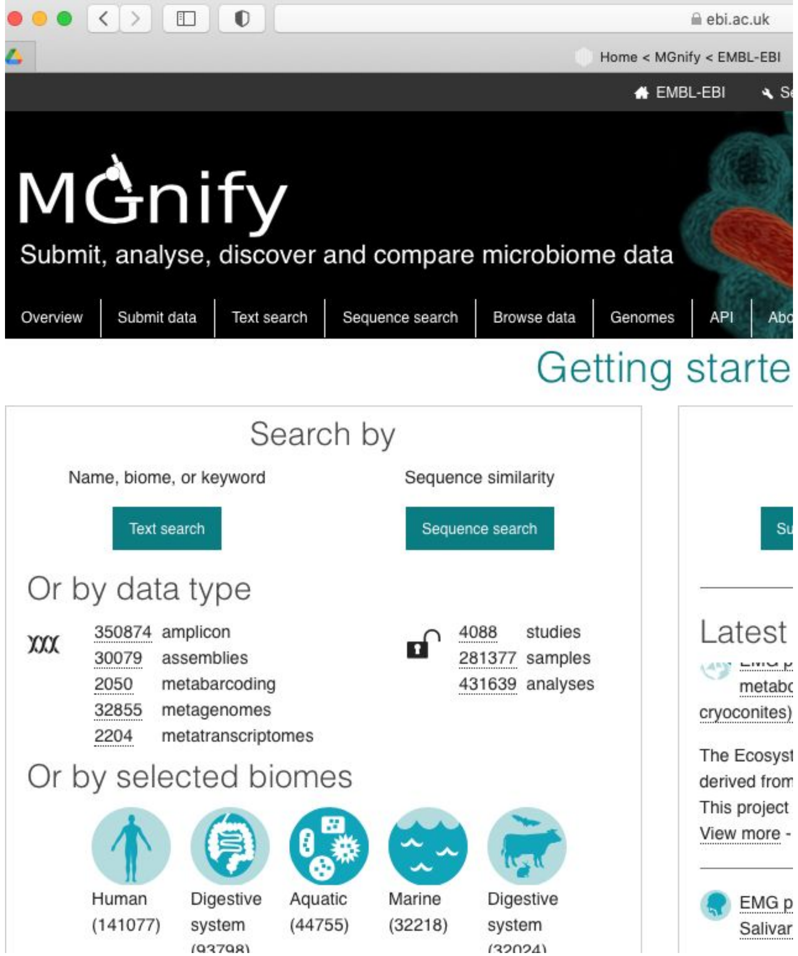
European Nucleotide Archive (ENA) submission

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]



MGnify

Submit, analyse, discover and compare microbiome data

Getting starte

Search by

Name, biome, or keyword

Sequence similarity

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)

Latest

metabo

cryoconites)

The Ecosyst

derived from

This project

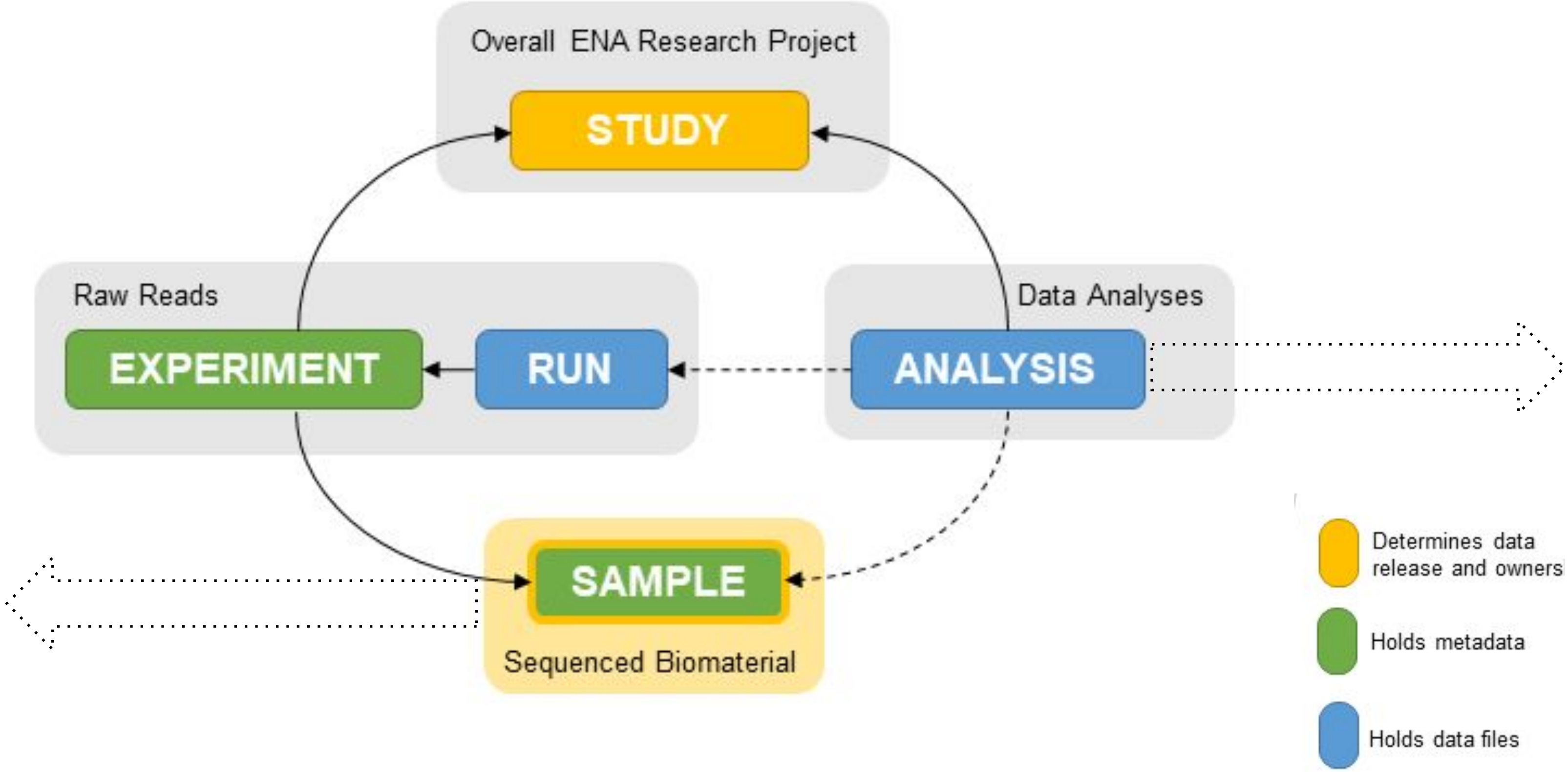
View more -

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

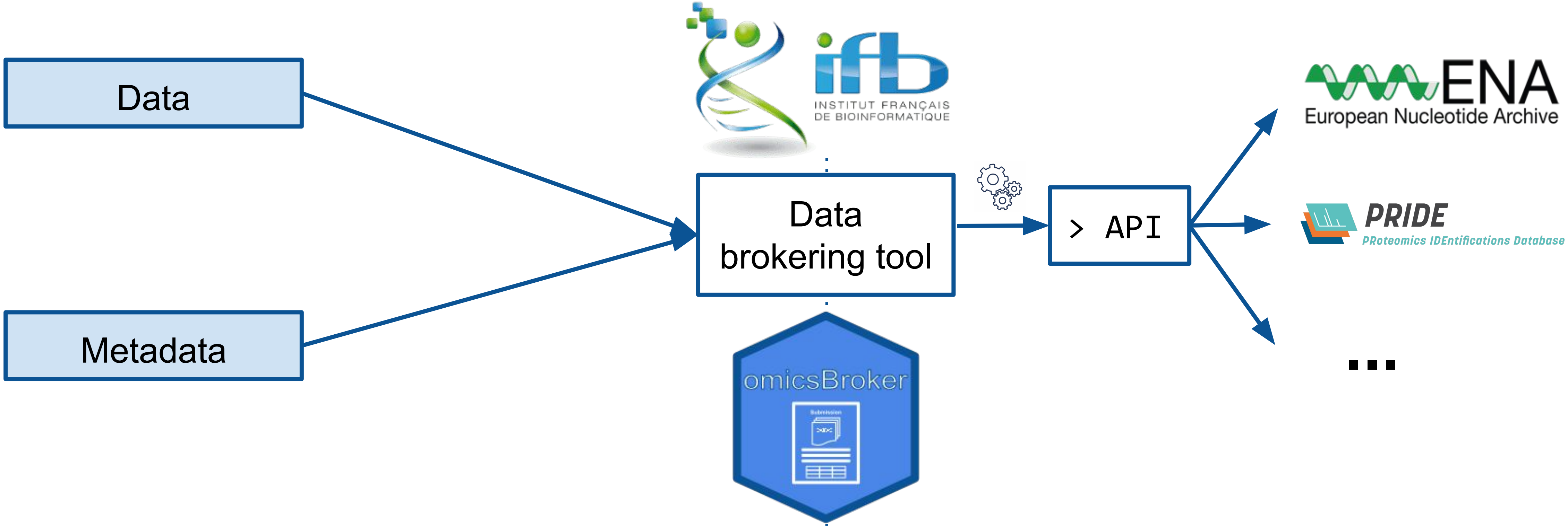
Main idea: offer a national service of **data brokering at IFB** to simplify and rationalize data exchange between international resources and the french Elixir node IFB.

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



Data
Metadata
INSTITUT FRANÇAIS DE BIOINFORMATIQUE
Data brokering tool
omicsBroker
> 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 (http://134.158.247.47:443 or http://134.158.247.47:443/app/omicsBroker) 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)
- SARS-Cov2 genome (IFB Training : Sars-CoV-2)
- plant transcriptome (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

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

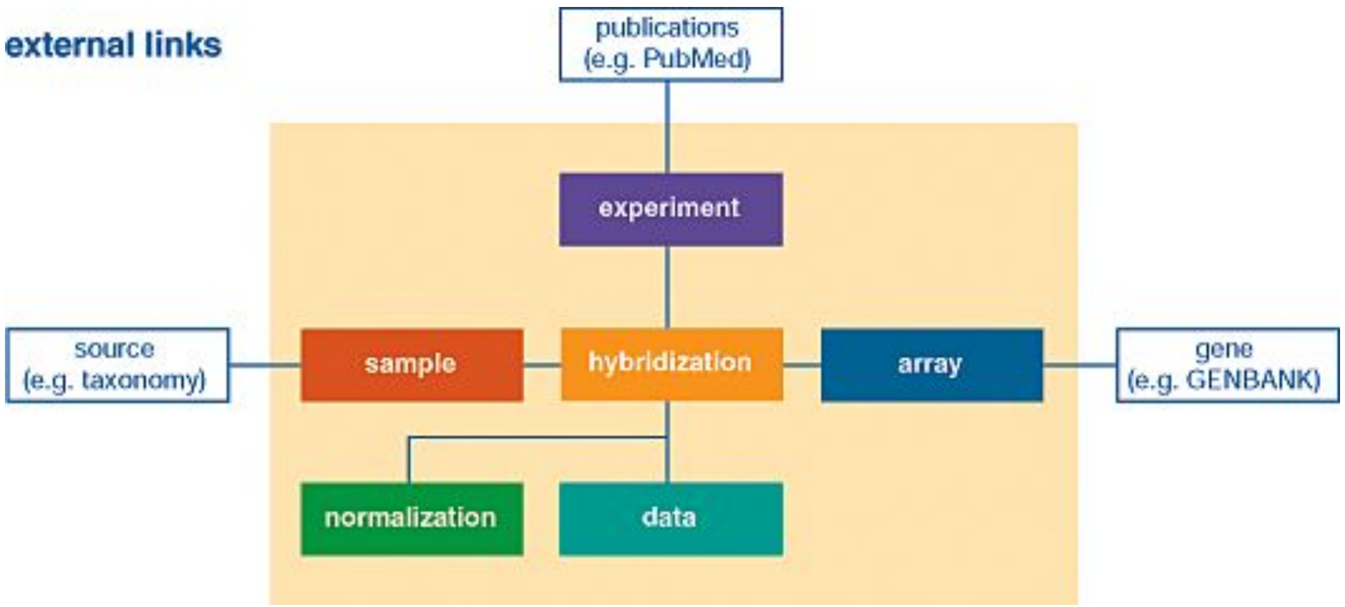
Identifier schemata

https://fairsharing.org

The Minimum information standard initiative

- A set of **guidelines** for reporting data derived by relevant methods in biosciences.

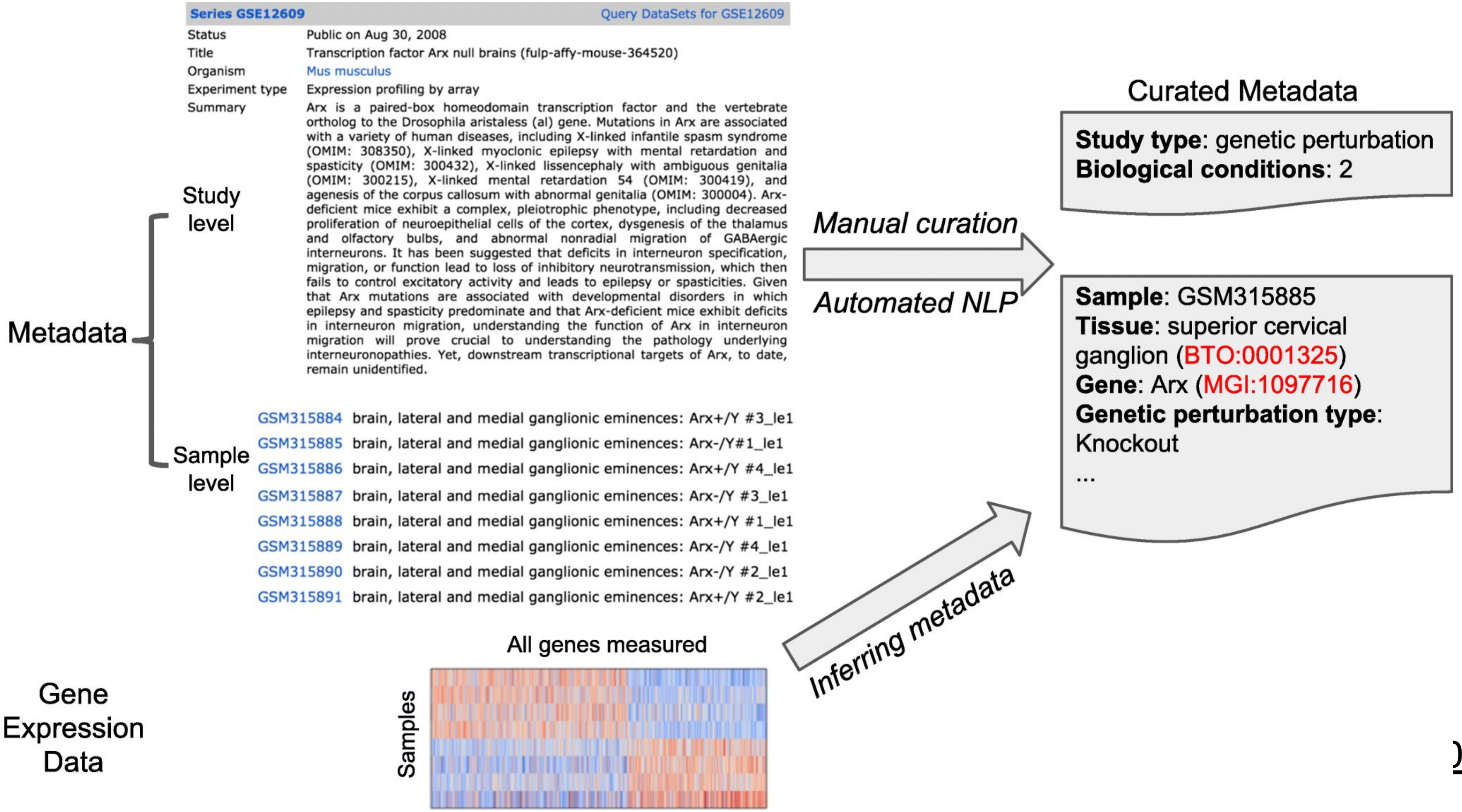
- Example : the **Minimum Information About a Microarray Experiment (MIAME)**



A schematic representation of six components of a microarray experiment.

https://en.wikipedia.org/wiki/Minimum_information_standard
10.1038/ng1201-365

Example 2: GEO (Gene Expression Omnibus) data & metadata



Series GSE12609 Query DataSets for GSE12609

Status	Public on Aug 30, 2008
Title	Transcription factor Arx null brains (fulp-affy-mouse-364520)
Organism	Mus musculus
Experiment type	Expression profiling by array
Summary	Arx is a paired-box homeodomain transcription factor and the vertebrate ortholog to the Drosophila aristaless (al) gene. Mutations in Arx are associated with a variety of human diseases, including X-linked infantile spasm syndrome (OMIM: 308350), X-linked myoclonic epilepsy with mental retardation and spasticity (OMIM: 300432), X-linked lissencephaly with ambiguous genitalia (OMIM: 300215), X-linked mental retardation 54 (OMIM: 300419), and agenesis of the corpus callosum with abnormal genitalia (OMIM: 300004). Arx-deficient mice exhibit a complex, pleiotrophic phenotype, including decreased proliferation of neuroepithelial cells of the cortex, dysgenesis of the thalamus and olfactory bulbs, and abnormal nonradial migration of GABAergic interneurons. It has been suggested that deficits in interneuron specification, migration, or function lead to loss of inhibitory neurotransmission, which then fails to control excitatory activity and leads to epilepsy or spasticities. Given that Arx mutations are associated with developmental disorders in which epilepsy and spasticity predominate and that Arx-deficient mice exhibit deficits in interneuron migration, understanding the function of Arx in interneuron migration will prove crucial to understanding the pathology underlying interneuronopathies. Yet, downstream transcriptional targets of Arx, to date, remain unidentified.

GSM315884	brain, lateral and medial ganglionic eminences: Arx+/Y #3_le1
GSM315885	brain, lateral and medial ganglionic eminences: Arx-/Y#1_le1
GSM315886	brain, lateral and medial ganglionic eminences: Arx+/Y #4_le1
GSM315887	brain, lateral and medial ganglionic eminences: Arx-/Y #3_le1
GSM315888	brain, lateral and medial ganglionic eminences: Arx+/Y #1_le1
GSM315889	brain, lateral and medial ganglionic eminences: Arx-/Y #4_le1
GSM315890	brain, lateral and medial ganglionic eminences: Arx-/Y #2_le1
GSM315891	brain, lateral and medial ganglionic eminences: Arx+/Y #2_le1

Metadata

Study level

Sample level

Gene Expression Data

All genes measured

Samples

Manual curation

Automated NLP

Inferring metadata

Curated Metadata

Study type: genetic perturbation
Biological conditions: 2

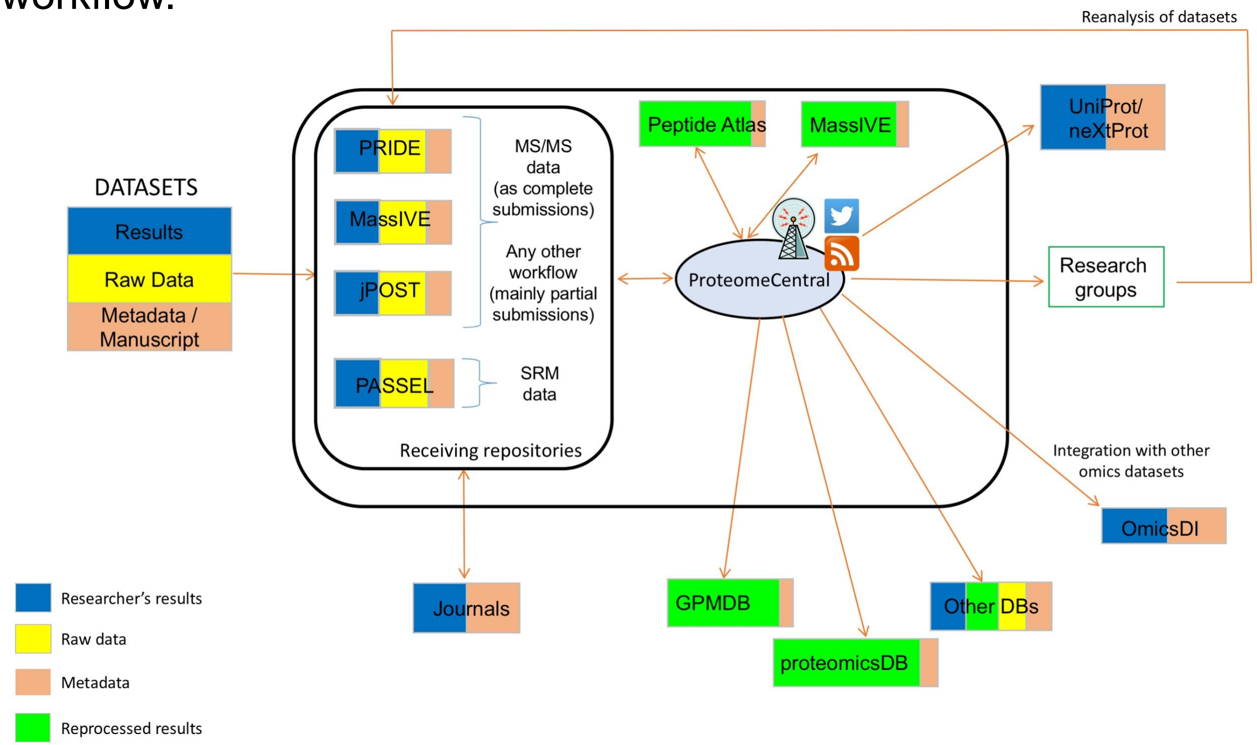
Sample: GSM315885
Tissue: superior cervical ganglion (BTO:0001325)
Gene: Arx (MGI:1097716)
Genetic perturbation type: Knockout
...

)7/s12551-018-0490-8/figures/2

Example 3: The ProteomeXChange initiative

- **An international consortium** devoted to mass spectrometry (MS)-based proteomics data
- Overall objective: provide a common framework and infrastructure for the **cooperation of proteomics resources** by **defining and implementing consistent, harmonized, user-friendly data deposition** and **exchange procedures** among the members

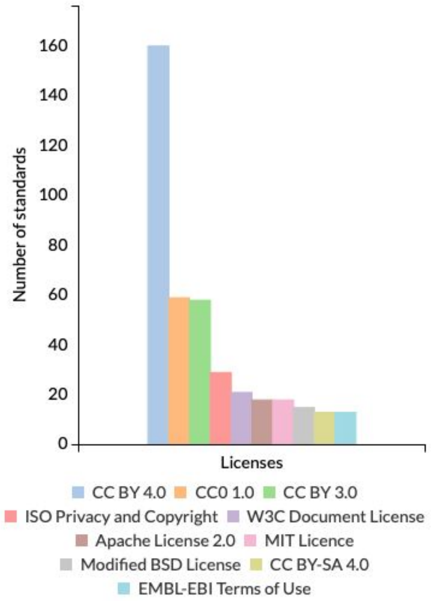
Figure 1. Schematic representation of the ProteomeXchange data workflow.



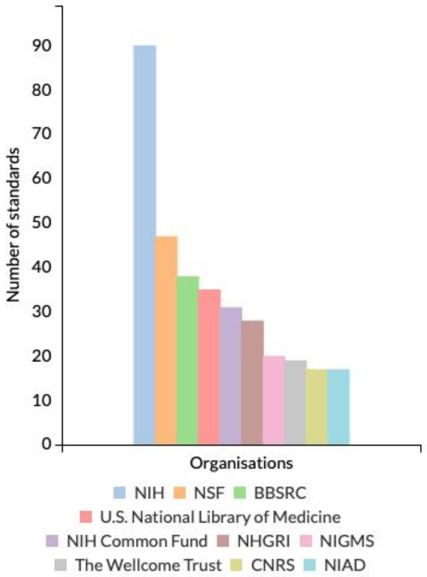
Source: **The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition**
Nucleic Acids Res, Volume 45, Issue D1, January 2017, Pages D1100–D1106, https://doi.org/10.1093/nar/gkw936

Summary statistics about standards

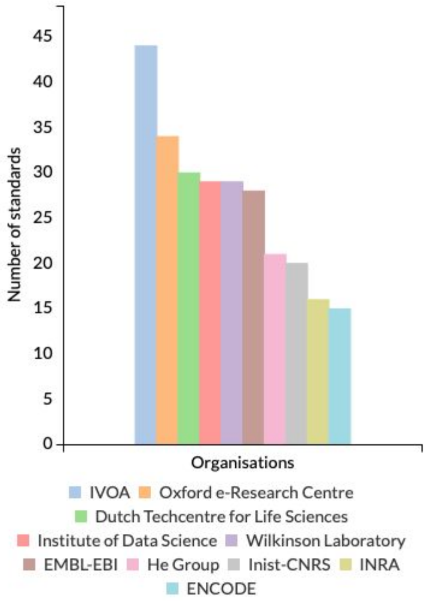
Top 10 licenses for standards



Top 10 funders of standards



Top 10 organisations (excluding funder) of standards



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

ENA proposes 3 types of submission

- Be careful: it is not possible to submit all objects using the 3 submission types

- IFB is currently being developing **brokering services to simplify submission to ENA**

	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	Y	Y	Y
Other Analyses	N	N	Y

An ENA submission step by step

1. Register a **submission account**

https://www.ebi.ac.uk/ena/submit/sra/#home

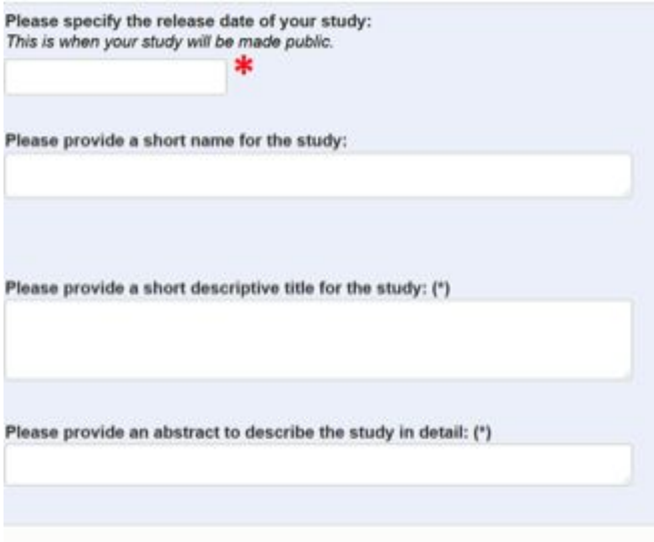
2. Register a **Study** (~a Project)

either *Interactively* or *Programmatically*

Using either test or production service:
https://wwwdev.ebi.ac.uk/ena/submit/sra *or*
https://www.ebi.ac.uk/ena/submit/sra

3. Choose a **Checklist**

https://www.ebi.ac.uk/ena/browser/checklists



Please specify the release date of your study:
This is when your study will be made public.
*

Please provide a short name for the study:

Please provide a short descriptive title for the study: (*)

Please provide an abstract to describe the study in detail: (*)

An ENA submission step by step

4. Register **samples** using the chosen Checklist and taxonomy

either *Interactively* or *Programmatically*

Using either test or production service

5. Register **experiments** and submit **raw data files** using the **run** object

either *Interactively* or *Programmatically* or *with the Webin-CLI application*

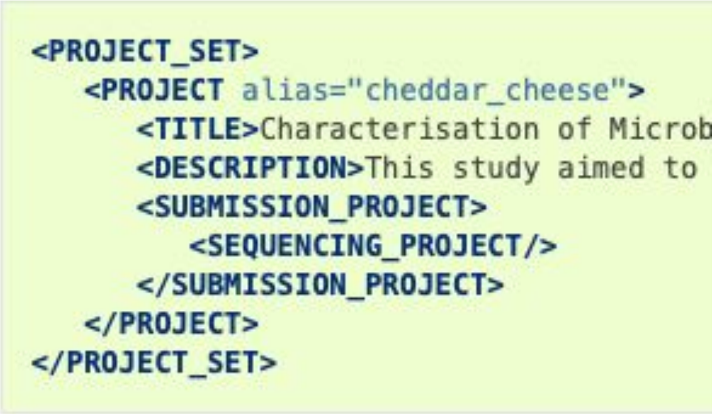
the run object includes the raw data filename and checksum code

6. Optionally describe **analyses** (assemblies, annotations,... not discussed in this training session)

Formats for an ENA submission

Metadata

- **Tabular** (Spreadsheats) files for the interactive mode
- **XML** files for the programmatic mode



```
<PROJECT_SET>
   <PROJECT alias="cheddar_cheese">
      <TITLE>Characterisation of Microb
      <DESCRIPTION>This study aimed to
      <SUBMISSION_PROJECT>
         <SEQUENCING_PROJECT/>
      </SUBMISSION_PROJECT>
   </PROJECT>
</PROJECT_SET>
```

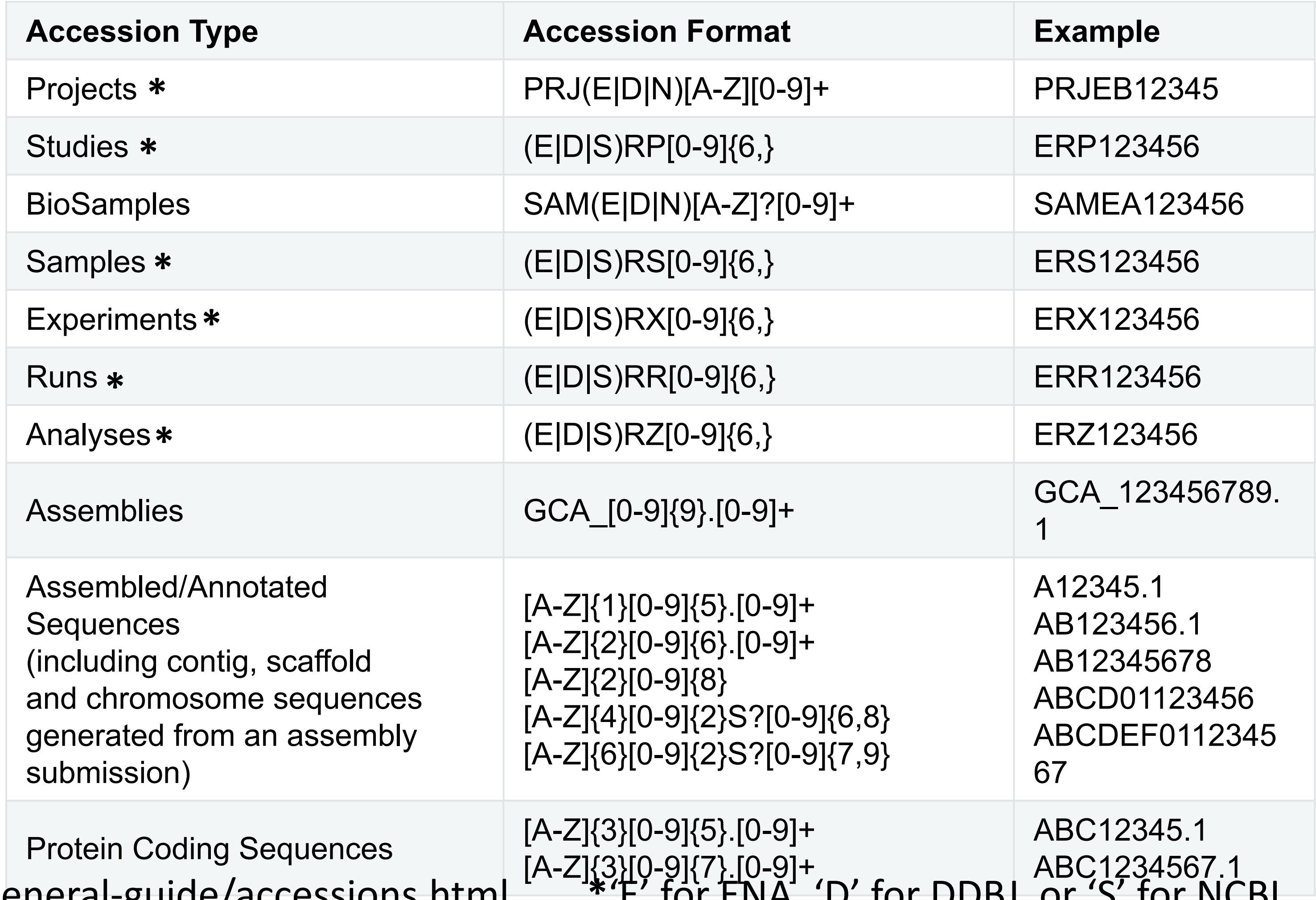
Data

- **Raw files:** standards formats like bam, cram, fastq, see https://ena-docs.readthedocs.io/en/latest/submit/fileprep/reads.html
- **Analysis files**
  - Assemblies: fasta file + manifest file + AGP file
  - Annotations: standards formats like bed or gff, see https://ena-docs.readthedocs.io/en/latest/submit/analyses.html

An ENA submission produce accession numbers

ENA project citation:

“the data for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEBxxxx (https://www.ebi.ac.uk/ena/browser/view/PRJEBxxxx).”



Accession Type	Accession Format	Example
Projects *	PRJ(E|D|N)[A-Z][0-9]+	PRJEB12345
Studies *	(E|D|S)RP[0-9]{6,}	ERP123456
BioSamples	SAM(E|D|N)[A-Z]?[0-9]+	SAMEA123456
Samples *	(E|D|S)RS[0-9]{6,}	ERS123456
Experiments *	(E|D|S)RX[0-9]{6,}	ERX123456
Runs *	(E|D|S)RR[0-9]{6,}	ERR123456
Analyses *	(E|D|S)RZ[0-9]{6,}	ERZ123456
Assemblies	GCA_[0-9]{9}.[0-9]+	GCA_123456789.1

Assembled/Annotated Sequences (including contig, scaffold and chromosome sequences generated from an assembly submission)	[A-Z]{1}[0-9]{5}.[0-9]+
	[A-Z]{2}[0-9]{6}.[0-9]+
	[A-Z]{2}[0-9]{8}
	[A-Z]{4}[0-9]{2}S?[0-9]{6,8}
	[A-Z]{6}[0-9]{2}S?[0-9]{7,9}
		A12345.1
		AB123456.1
		AB12345678
		ABCD01123456
		ABCDEF011234567

Protein Coding Sequences	[A-Z]{3}[0-9]{5}.[0-9]+ [A-Z]{3}[0-9]{7}.[0-9]+	ABC12345.1 ABC1234567.1

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

*‘E’ for ENA, ‘D’ for DDBJ, or ‘S’ for NCBI