Formation “Science Ouverte & PGD”
Module 3: Metadata

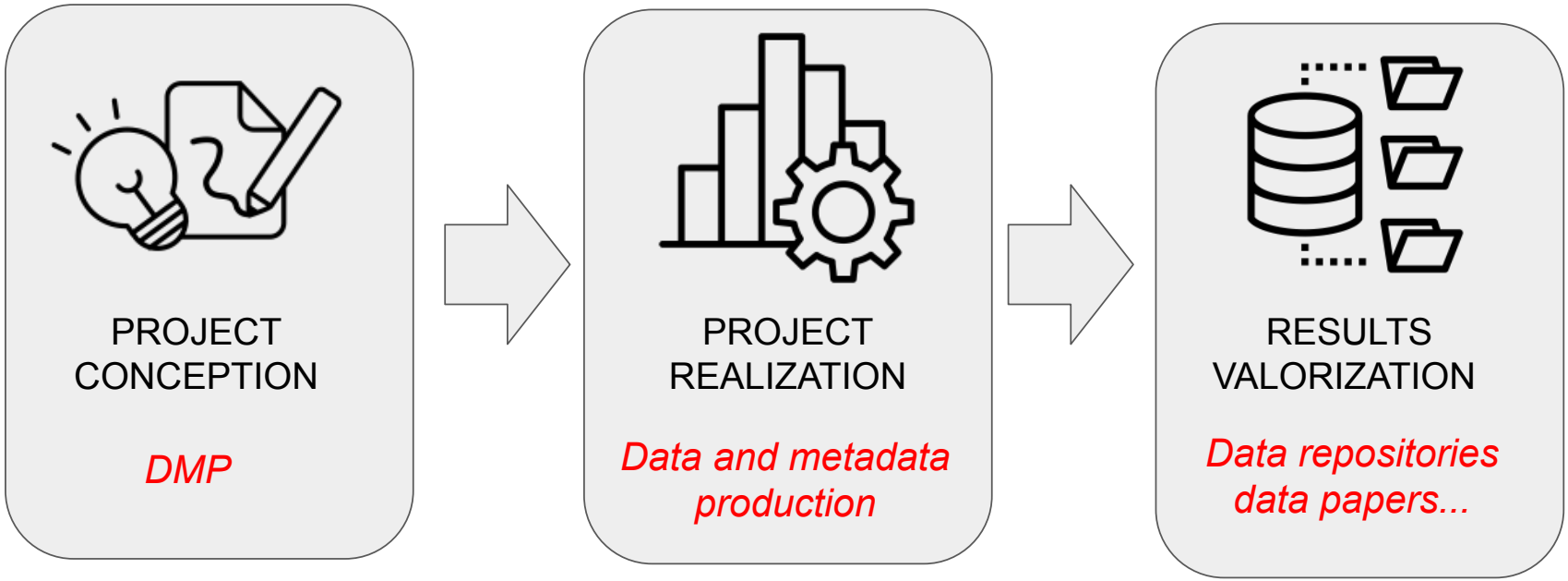
Standards de métadonnées en Sciences de la Vie

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

INSTITUT FRANÇAIS DE BIOINFORMATIQUE

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

https://scrumblr.ethibox.fr/metadata_standard

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

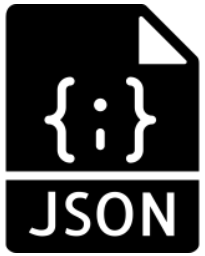
Three text formats frequently used for metadata

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

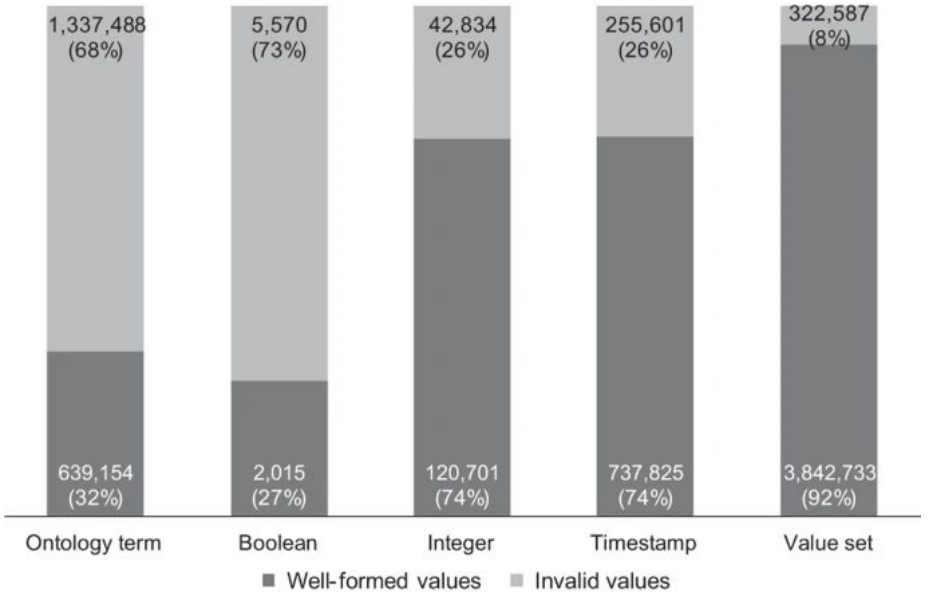
```
{
"SAMPLE_SET": {
     "SAMPLE":[
     {
          "alias": "A",
          "date":"20200802",
          "source":"blood"
     },
     {
          "alias": "B",
          "date":"20200802",
          "source":"feces"
     },
  {
          "alias": "C",
          "date":"20200802",
          "source":"skin"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

Metadata exhibit questionable quality in biology

Submission in public resources is often a complex task

Submission procedures are heterogeneous

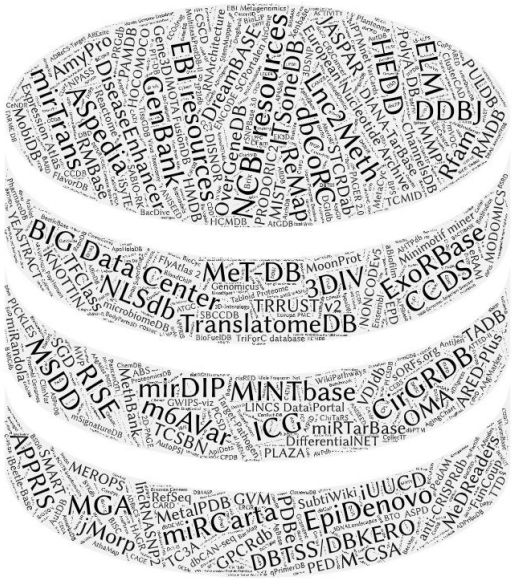
Metadata are often incomplete, inconsistent, redundant or not informative enough



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

Standard adoption and perennity

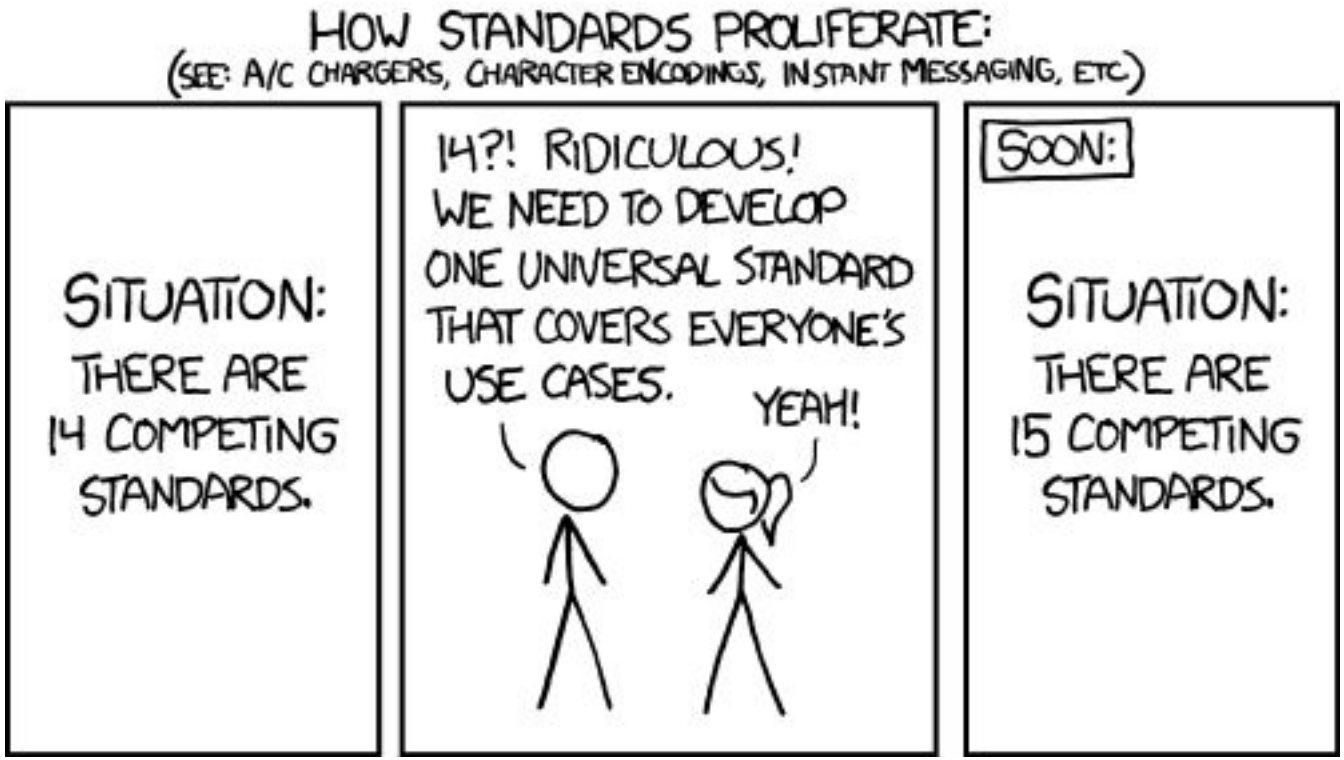
- There are thousand of databases, softwares and resources in biology with an **unequal level of standard adoption**
- Is is not always easy for life scientists and bioinformaticians to identify and use the most appropriate standards



1641 databases in NAR Database 2021

Rigden *et al*, 2021

Standard adoption and perennity



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

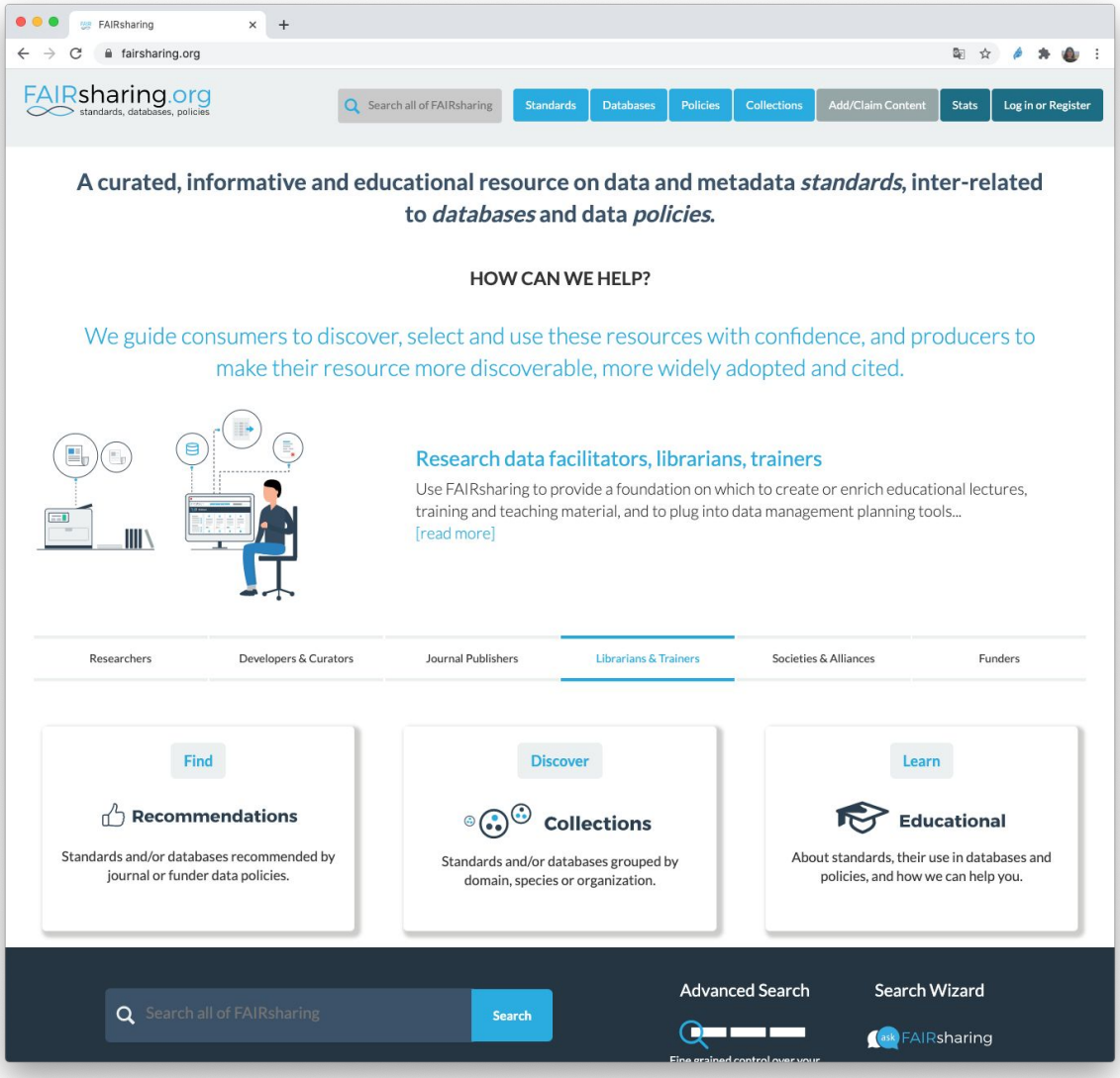
How do I find the standard I need?

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



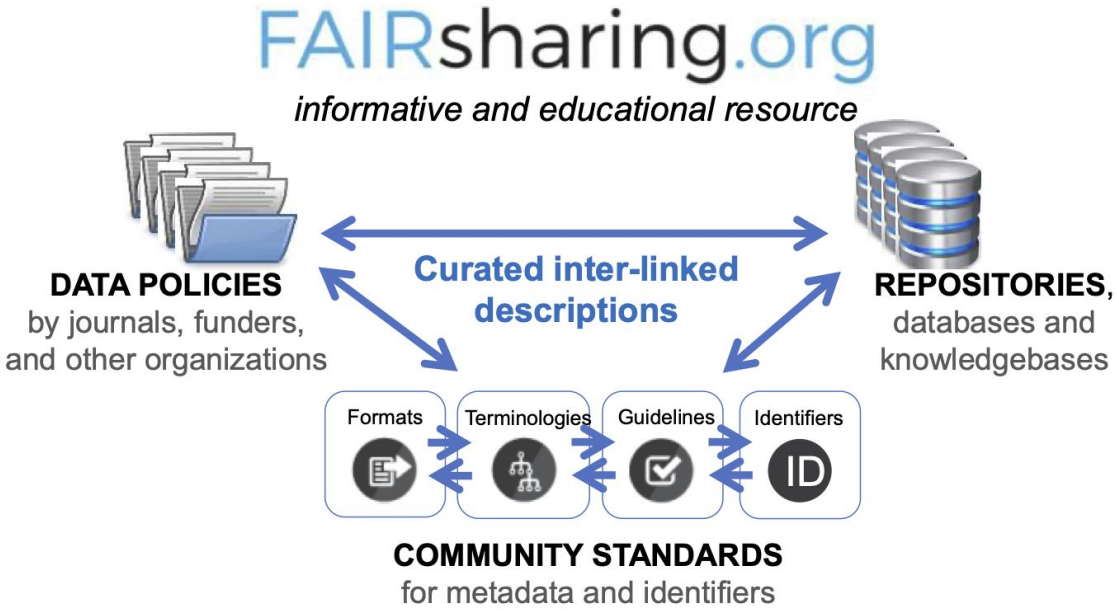
https://fairsharing.org

The FAIRsharing portal

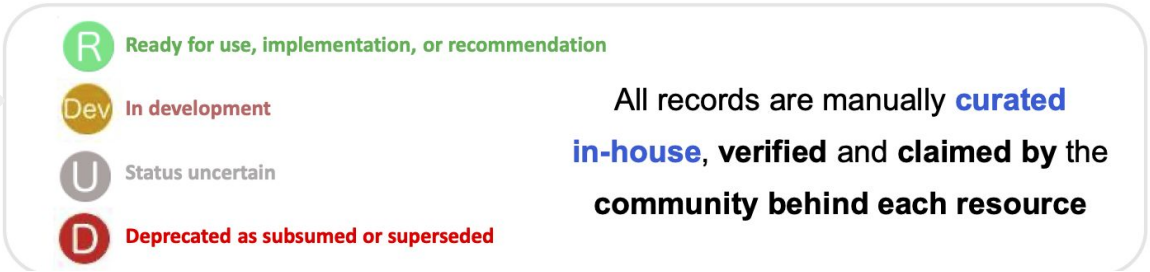
Citable *DOI* for all records

Accessible via *API* or *web interface*

Curation

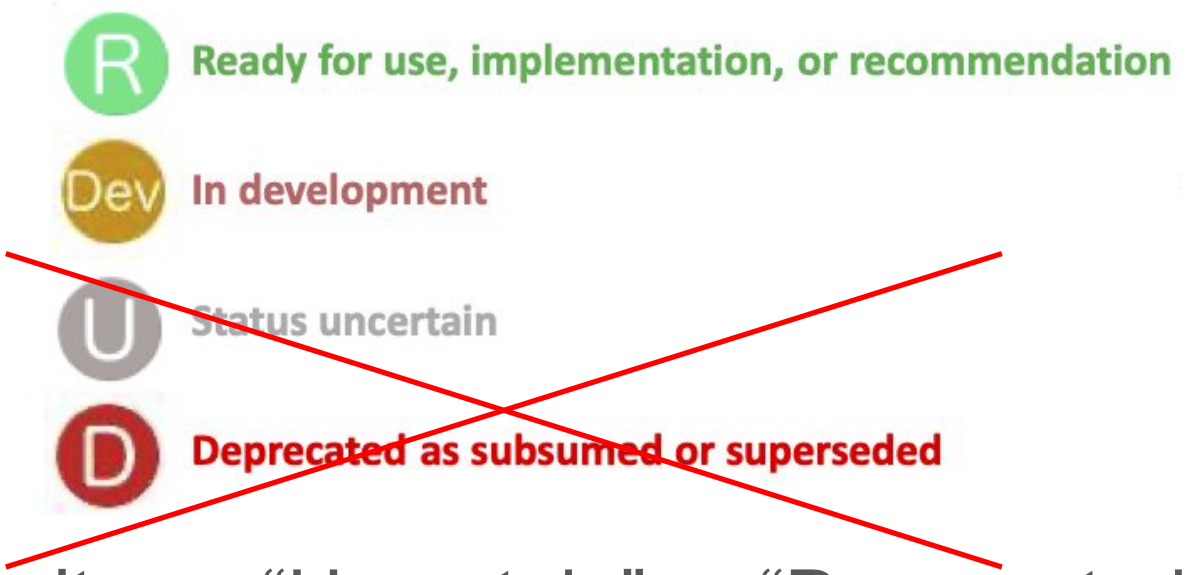


RECORD STATUS



https://fairsharing.org

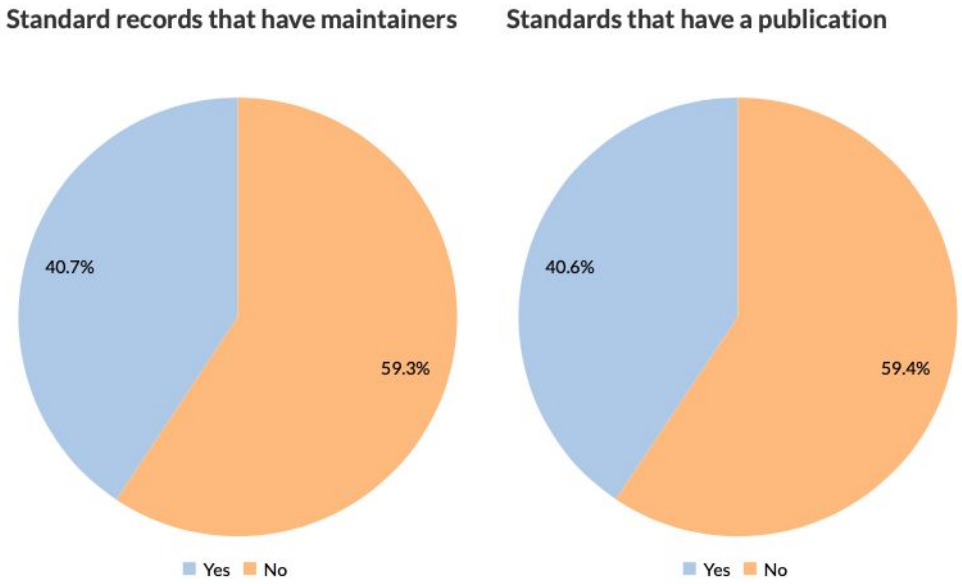
The FAIRsharing portal: record status



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

https://fairsharing.org

Standard maintenance is a key point



59.3 % of standards have no maintainer

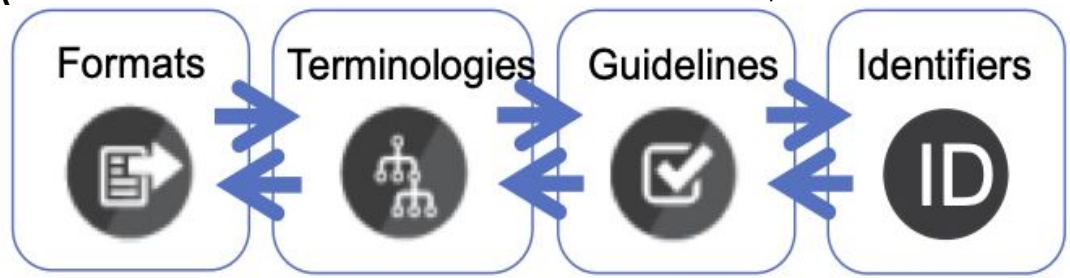
59.4% of standard has no publication

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

Types of data standards

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

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

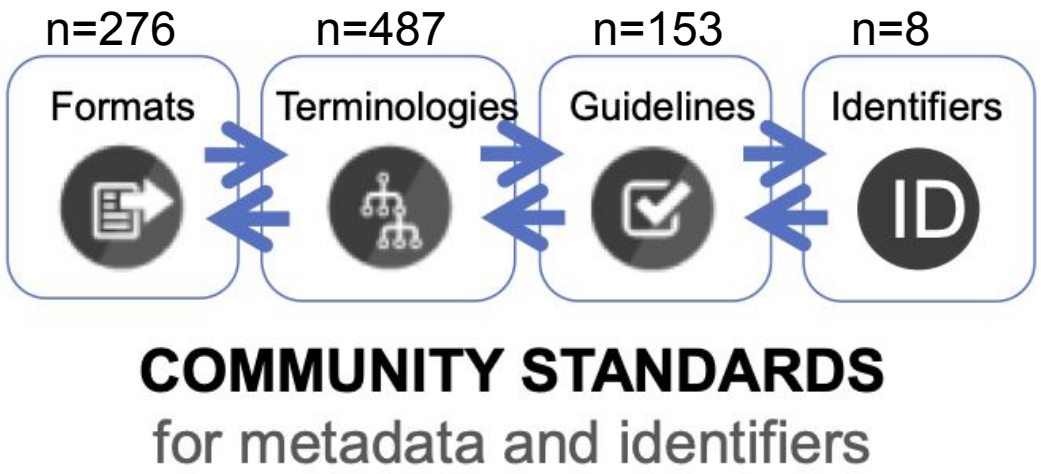


Controlled vocabularies, taxonomies, ontologies…
e.g. Gene Ontology

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



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

genomic STANDARDS consortium

isatools

GENEONTOLOGY Unifying Biology

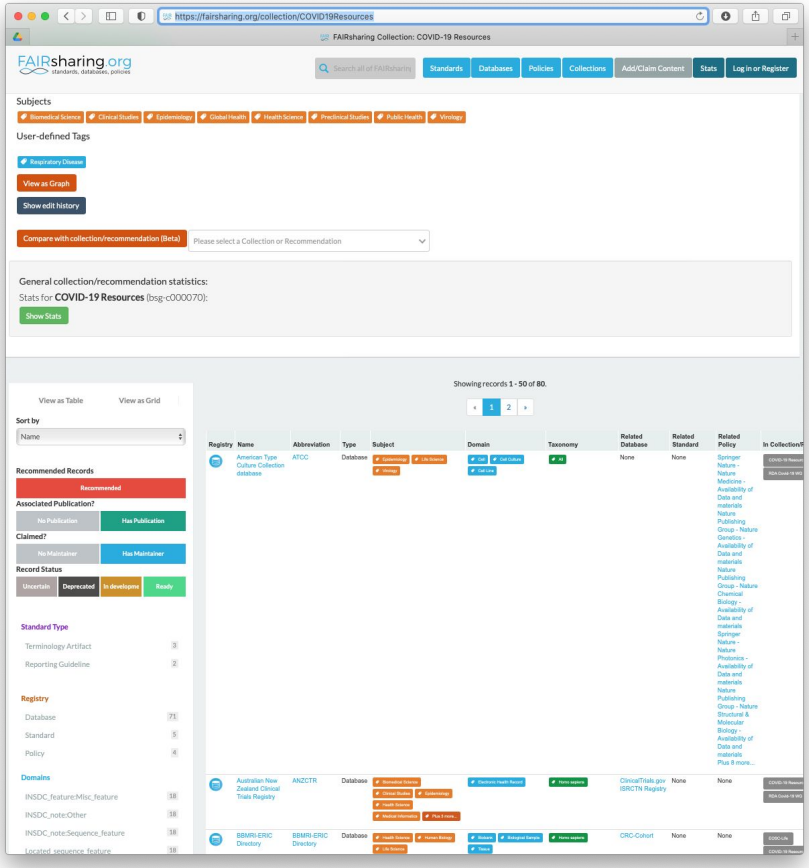
Crop Ontology for agricultural data

DISEASE ONTOLOGY

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



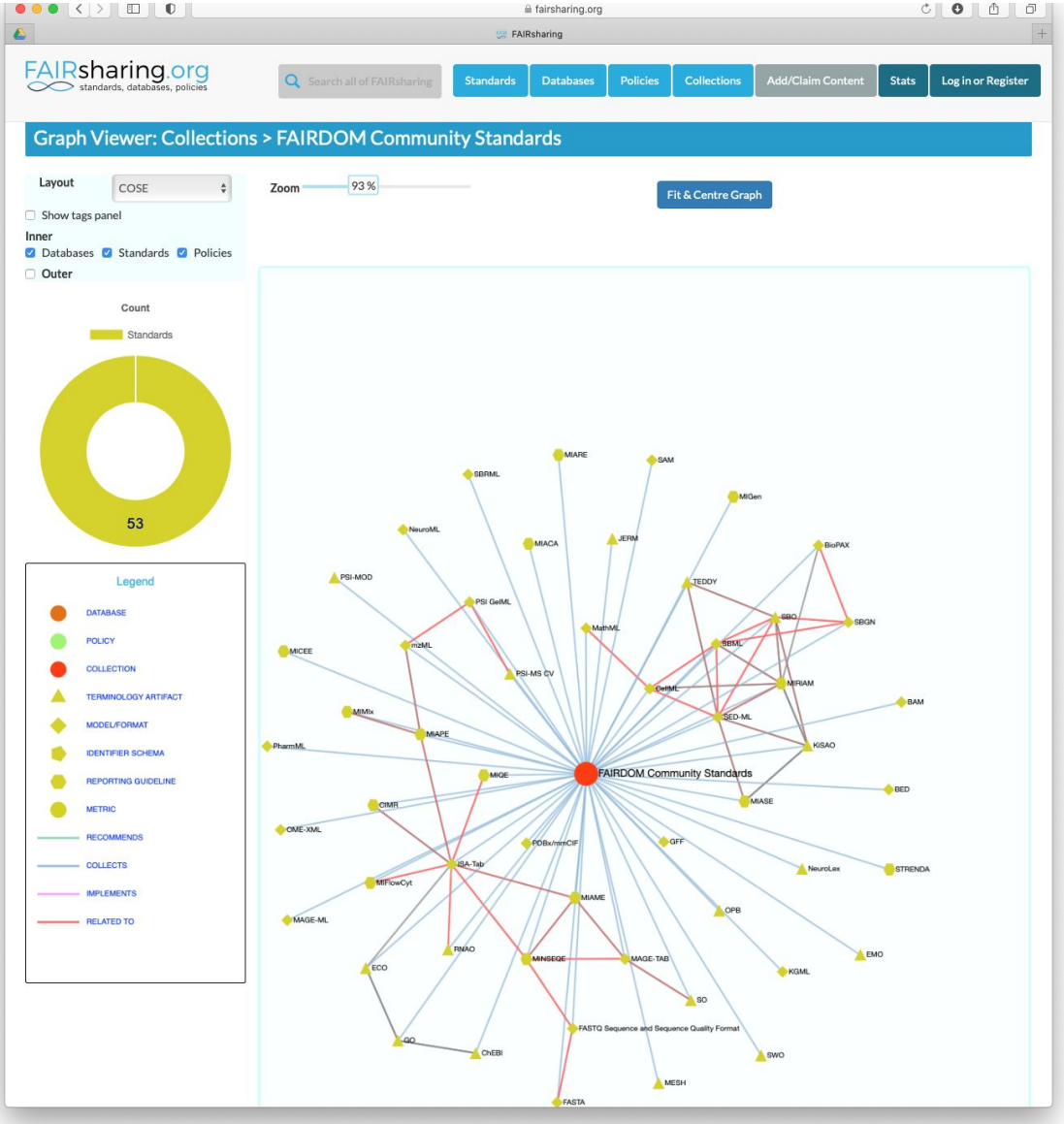
https://fairsharing.org/collections/

Collections in Life Sciences

53 collections related to Life Science standards in FAIRsharing

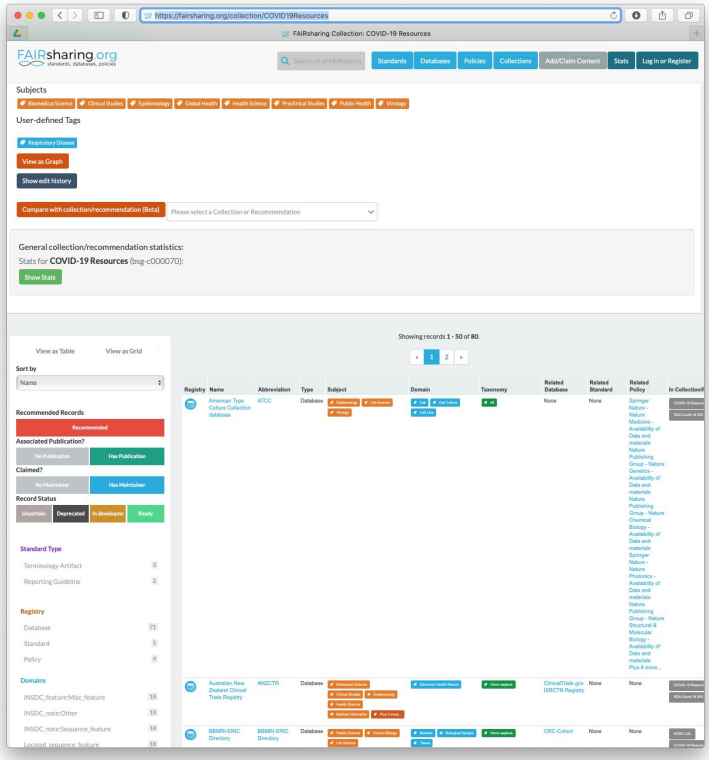
Example 1: the *FAIRdom community Standards collection* (System biology)

https://fairsharing.org/collection/FAIRDOM

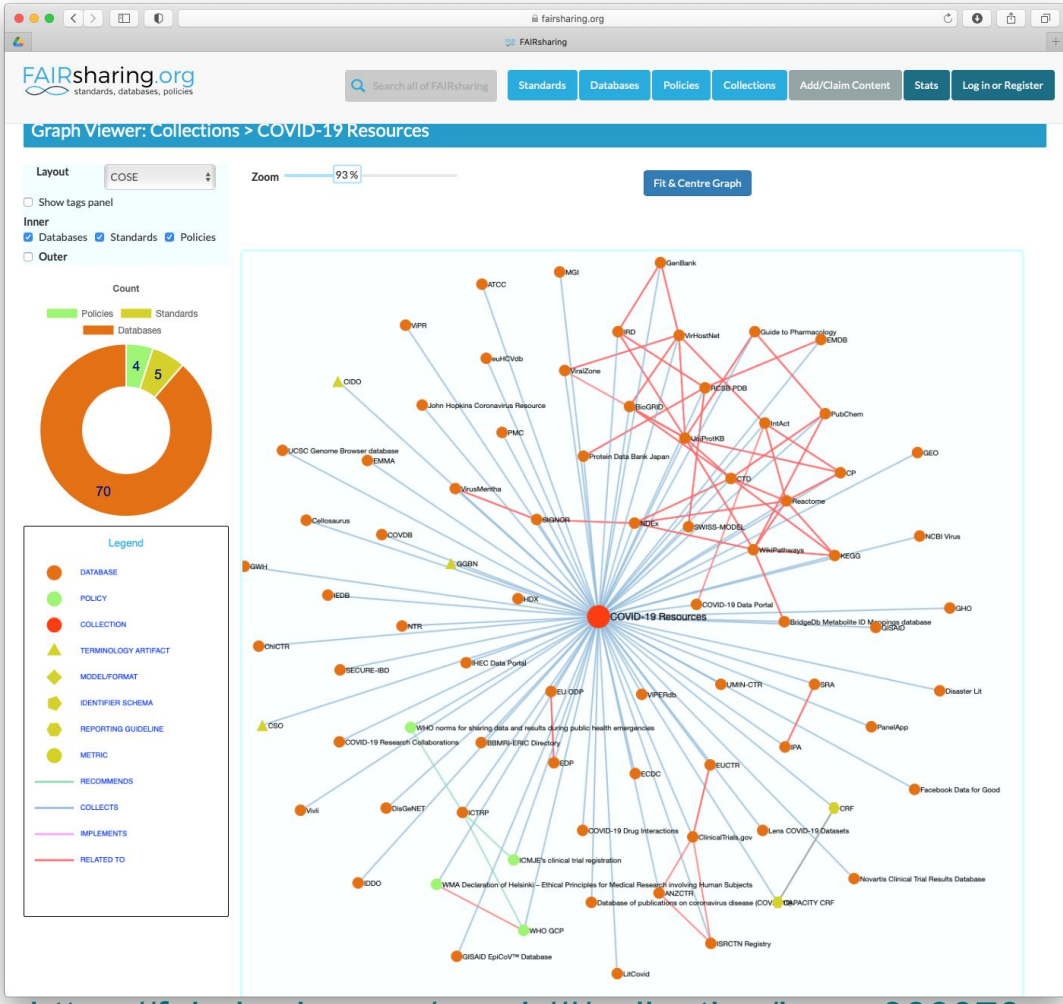


Some collections are recent

Example 2: The *Covid-19* collection



https://fairsharing.org/collection/COVID19Resources

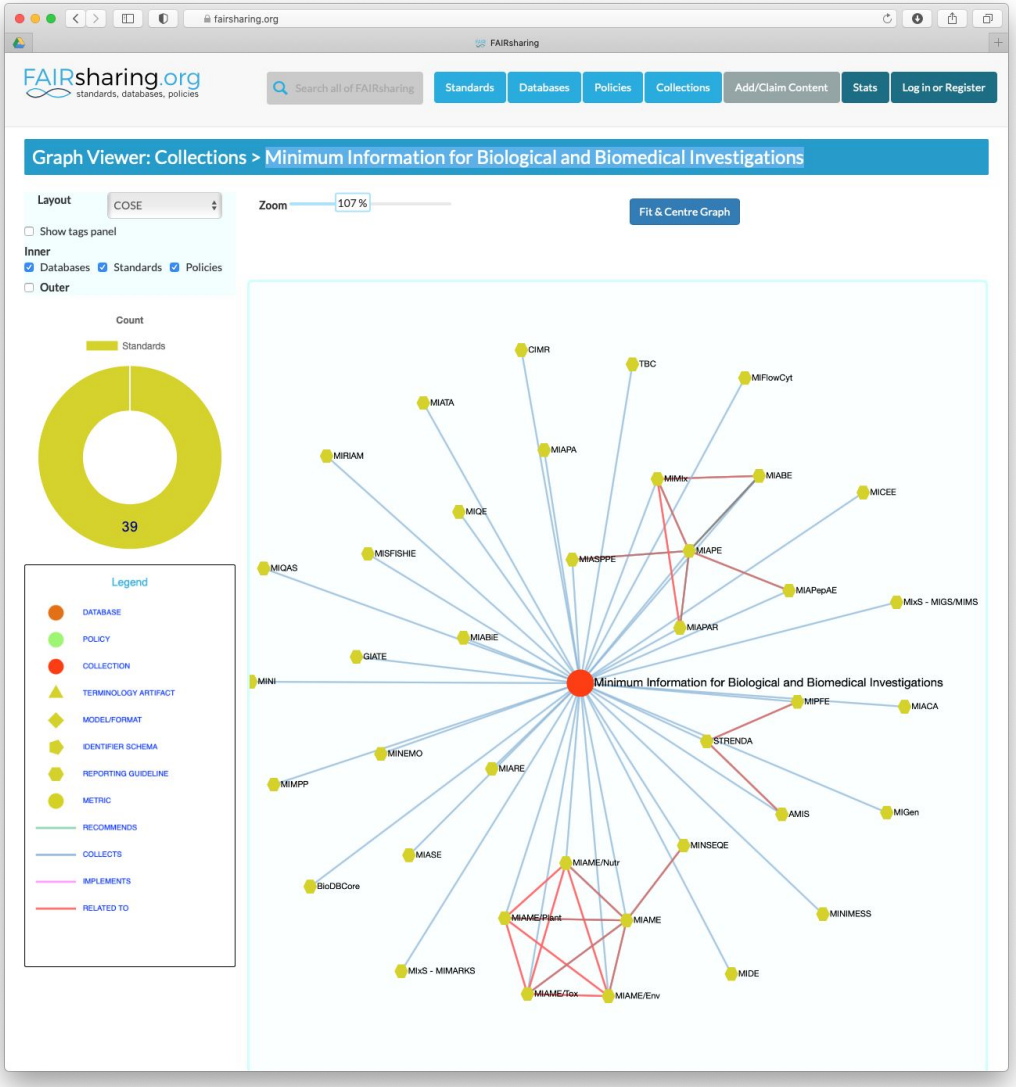


https://fairsharing.org/graph/#/collection/bsg-c000070

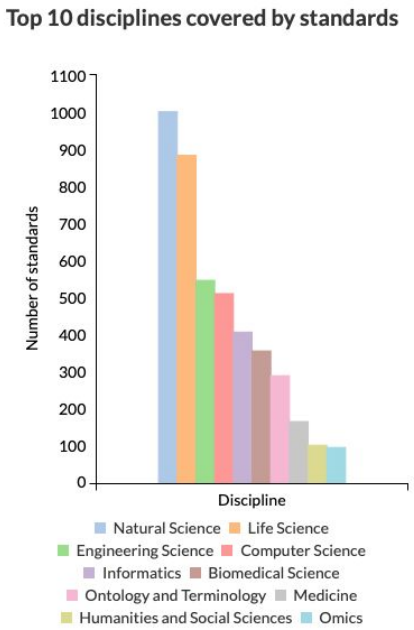
What about the minimum required metadata in biology?

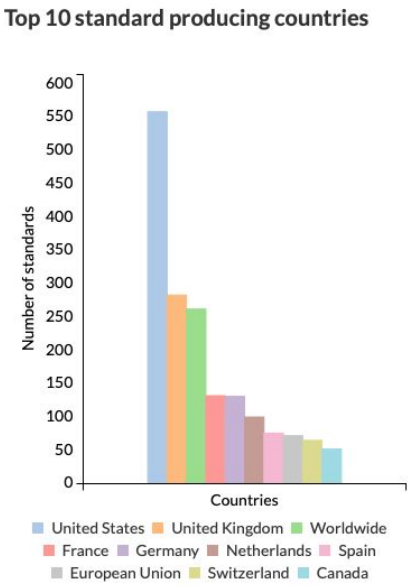
Example 3: the *Minimum Information for Biological and Biomedical Investigations* collection

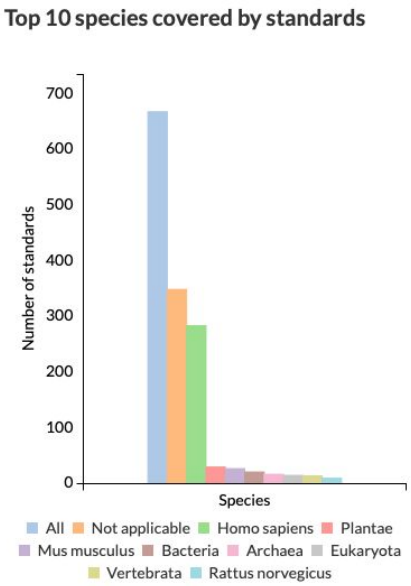
https://fairsharing.org/collection/MIBBI



Summary statistics about 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

Practice => Answers

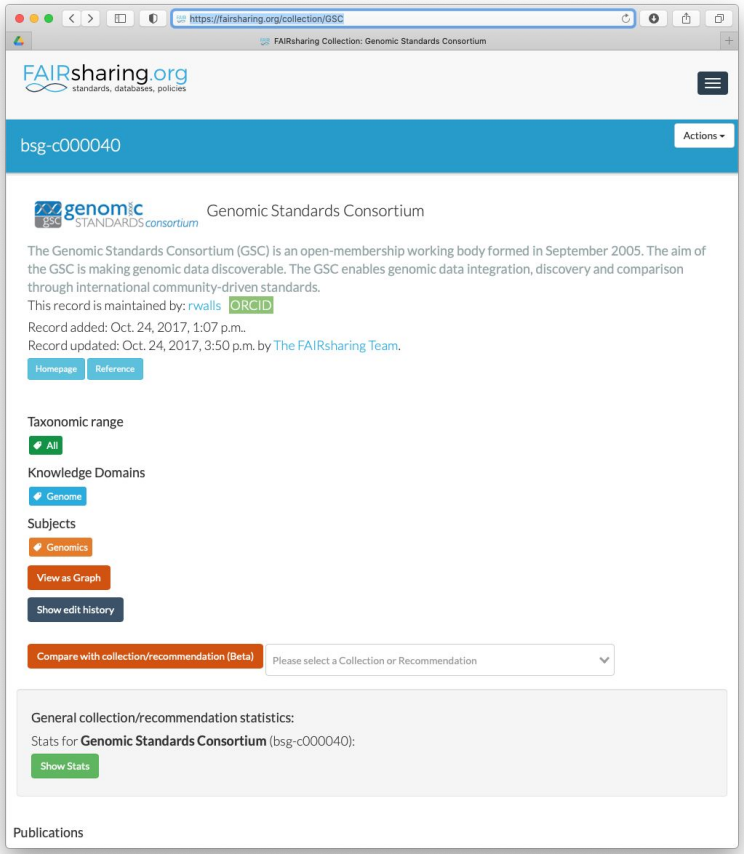
Find the *Genomic Standards Consortium (GSC) used by both ENA and SRA databases* in the **FAIRsharing collections**

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

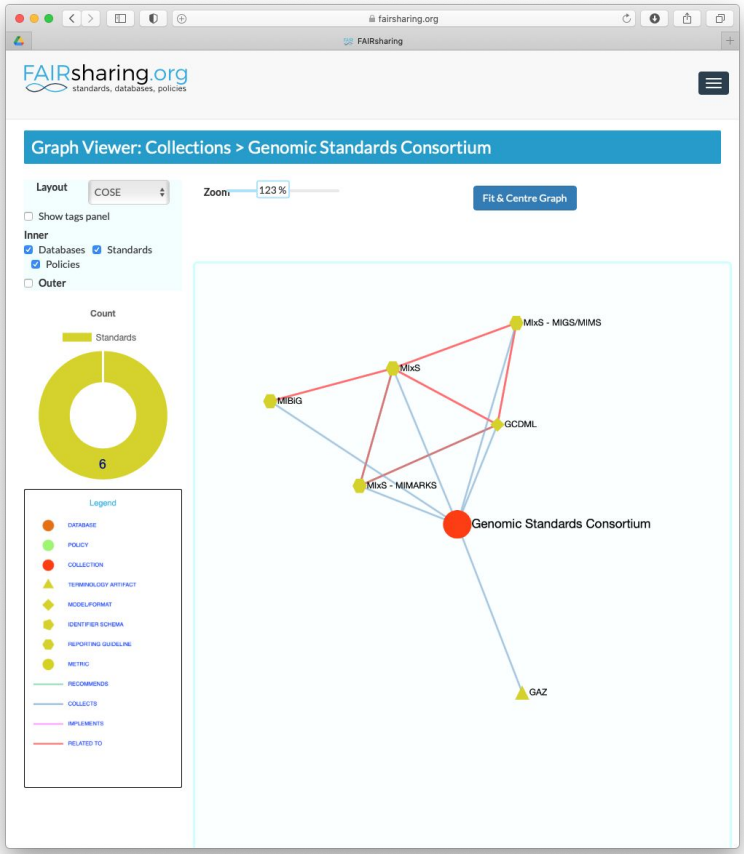
1. How many records (*i.e.* standards) are associated to the *GSC* ? => 6
2. What type of standard is *Minimum Information about any (x) Sequence (MiXS)* ? => Reporting guideline
3. What is the record status of the *GAZ* record ?=>Uncertain

Source: https://gensc.org

The Genomic Standards Consortium (GSC)



https://fairsharing.org/collection/GSC

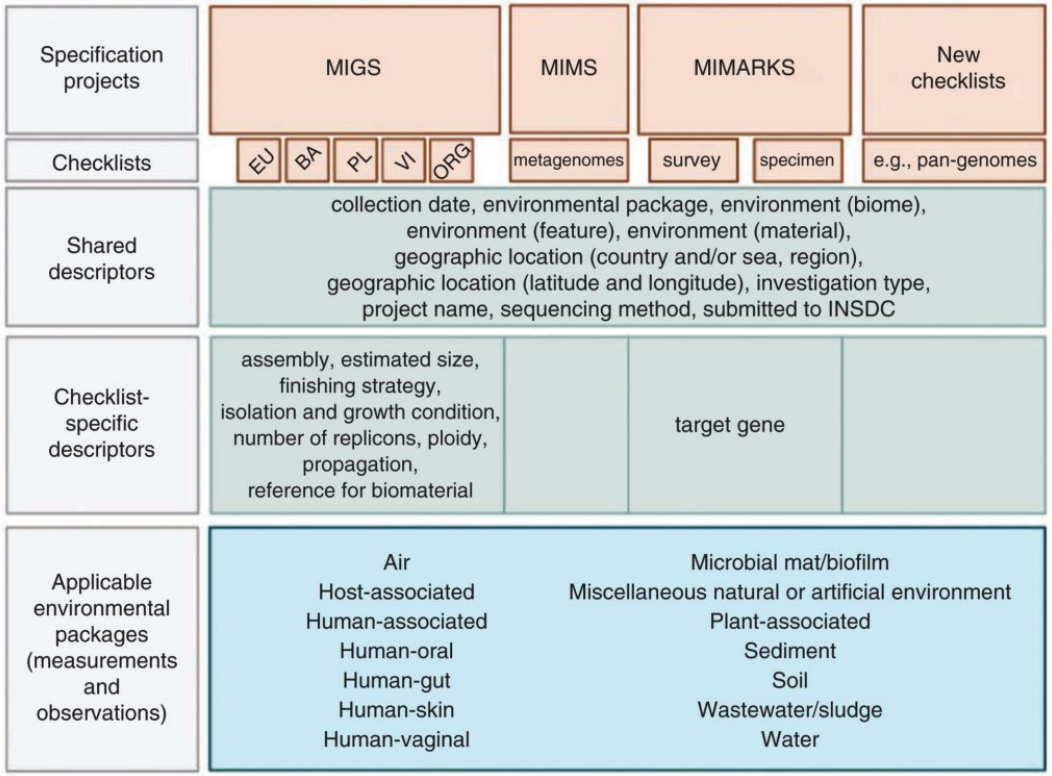


https://fairsharing.org/graph/#/collection/bsg-c000040

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

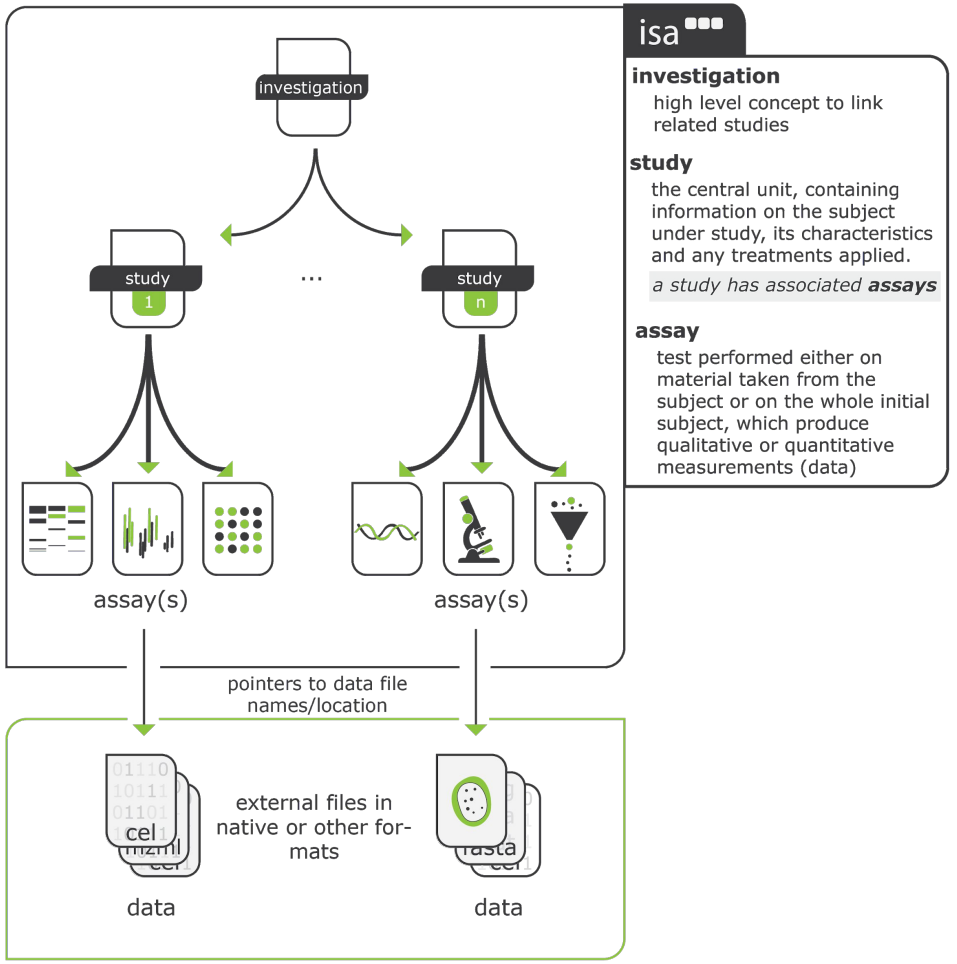
Source: https://gensc.org

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

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