

Formation “Science Ouverte & PGD”  
*Module 3: Metadata*



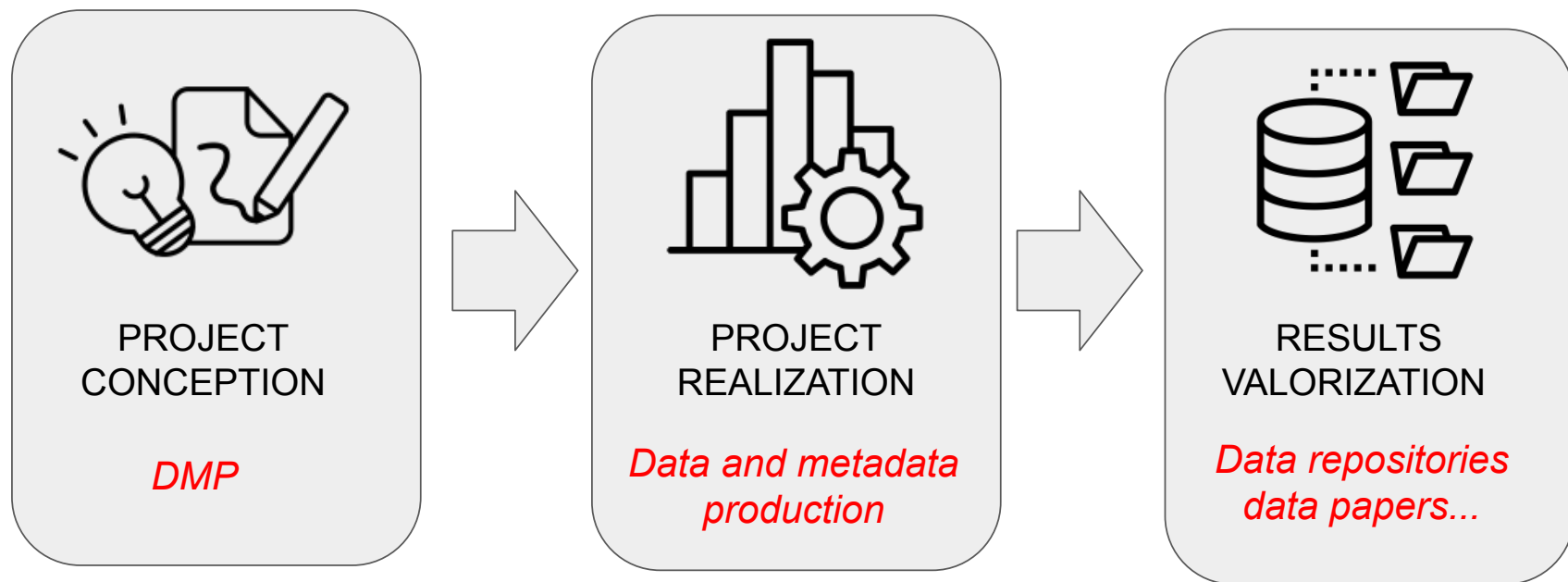
# Standards de métadonnées en Sciences de la Vie

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>

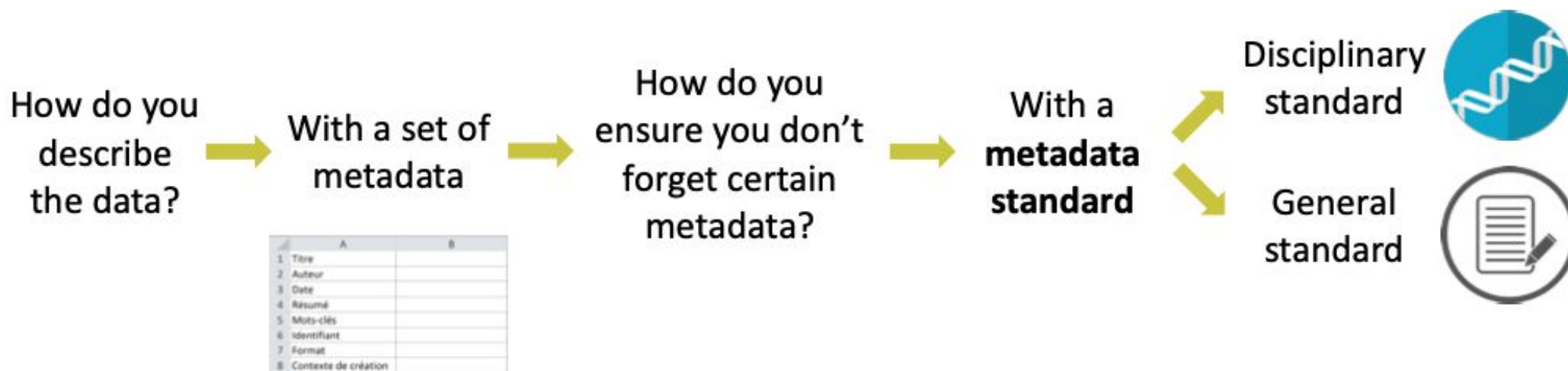


# Metadata during the project



**Metadata concern all steps of a scientific project !**

# How do I produce metadata?



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](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)

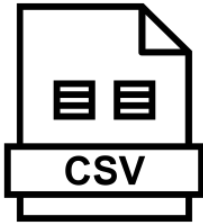
Metadata standards often depend on the repository you will use to publish data

> 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



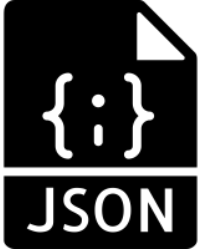
**Comma Separated Values**

```
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"
      }
    ]
  }
}
```

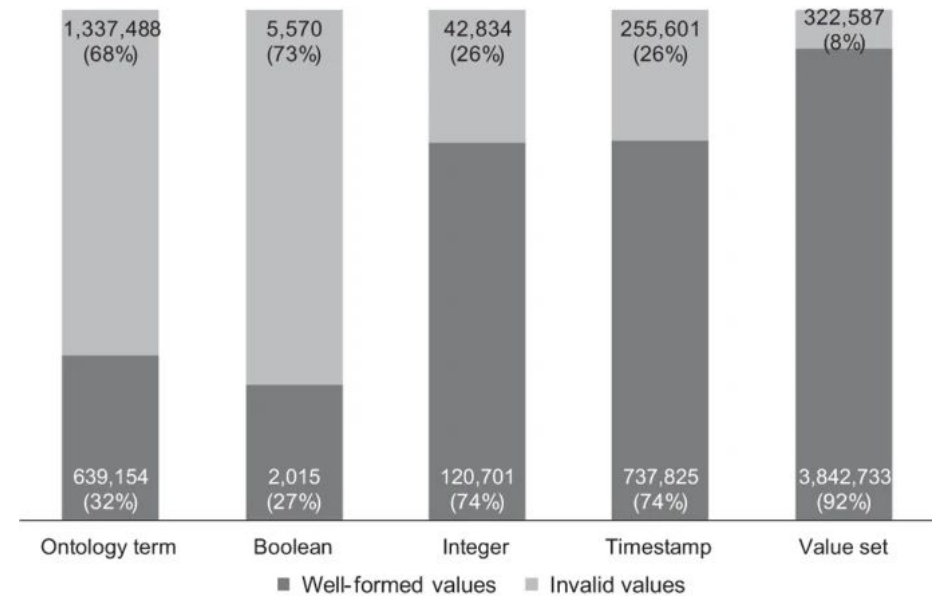


# Metadata exhibit questionable quality in biology

Submission in public resources is often a complex task

Submission procedures are heterogeneous

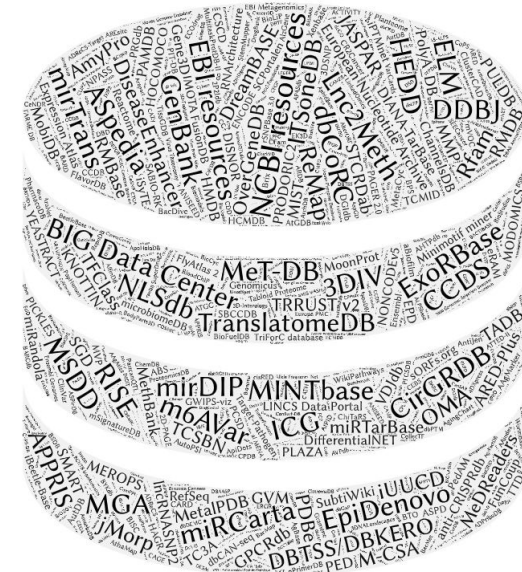
**Metadata are often incomplete, inconsistent, redundant or not informative enough**



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

# Standard adoption and perennity

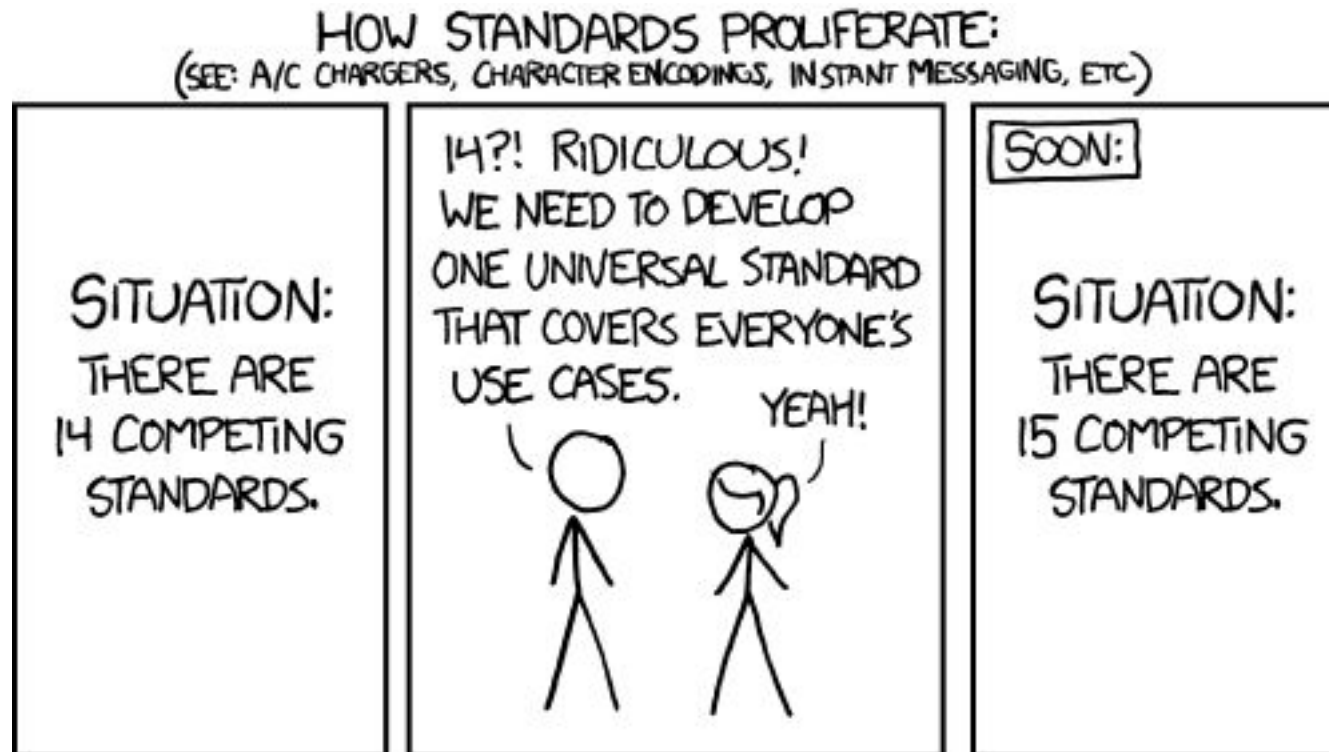
- There are thousand of databases, softwares and resources in biology with an **unequal level of standard adoption**
- Is is not always easy for life scientists and bioinformaticians to identify and use the most appropriate standards



1641 databases in NAR Database 2021

[Rigden et al, 2021](#)

# Standard adoption and perennity



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

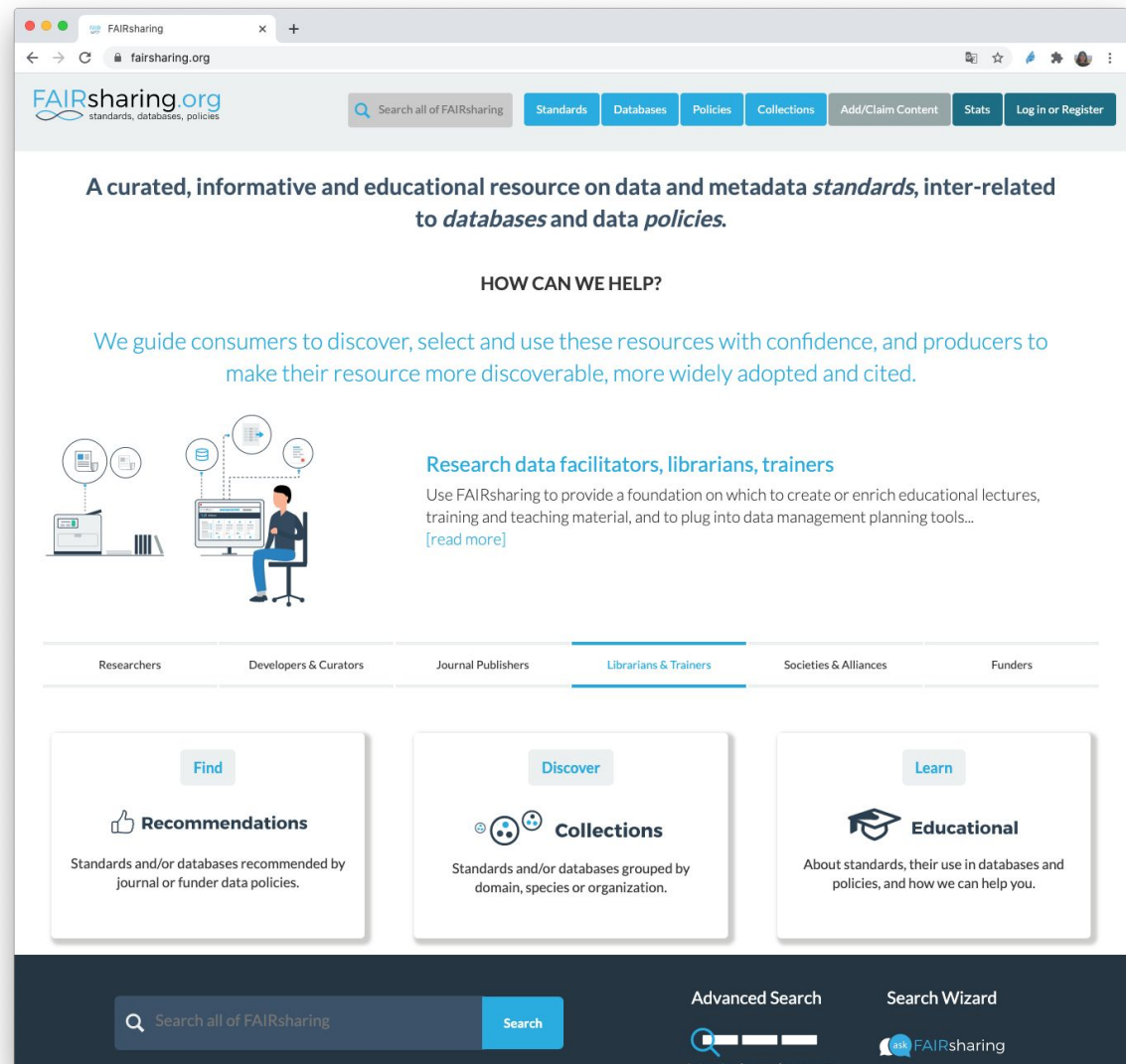
How do I find the standard I need?

# 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>



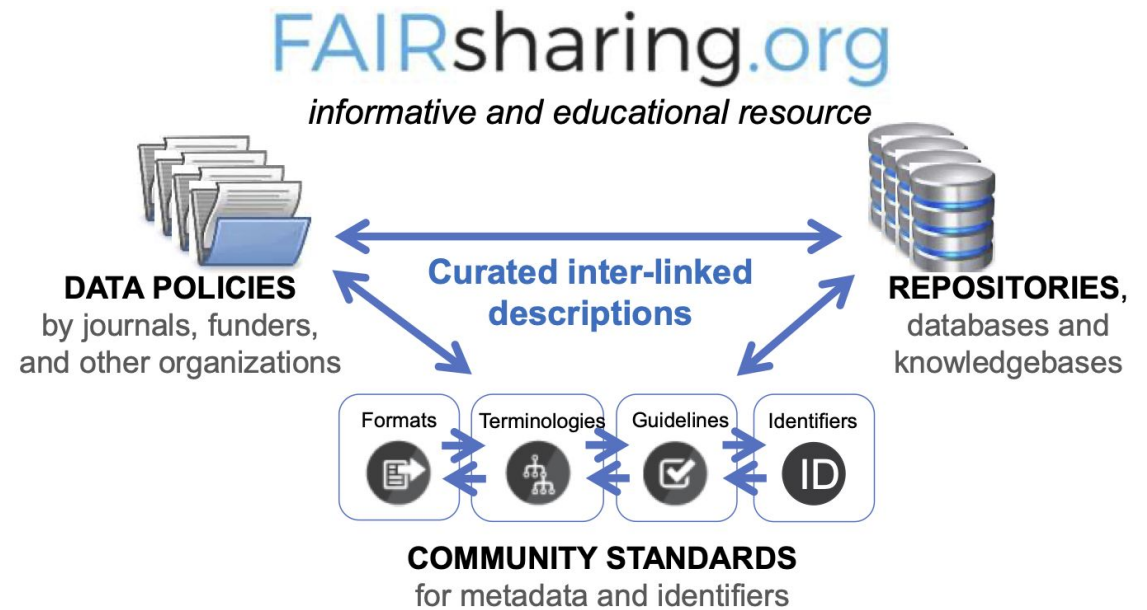
<https://fairsharing.org>

# The FAIRsharing portal

Citable DOI for all records

Accessible via API or web interface

Curation



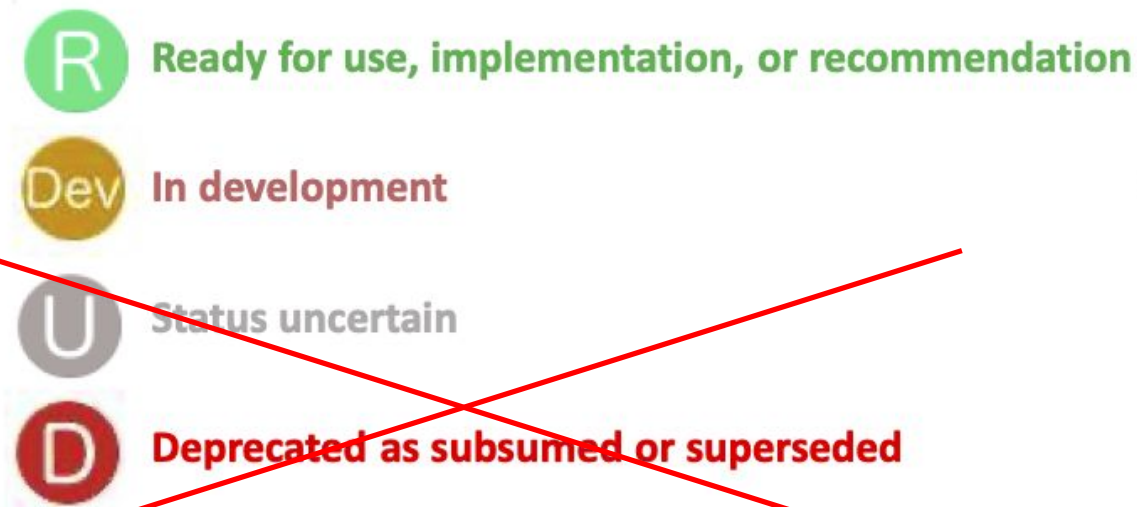
**RECORD  
STATUS**

- 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

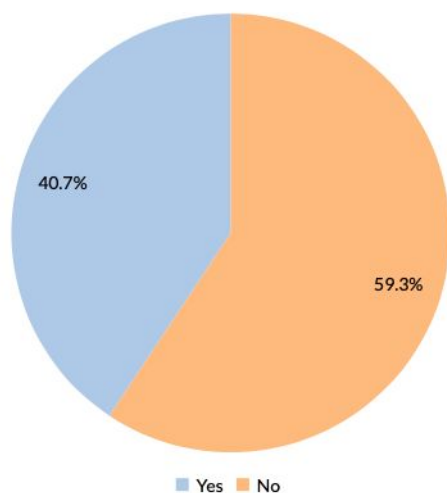


Please don't use "Uncertain" or "Deprecated" standards

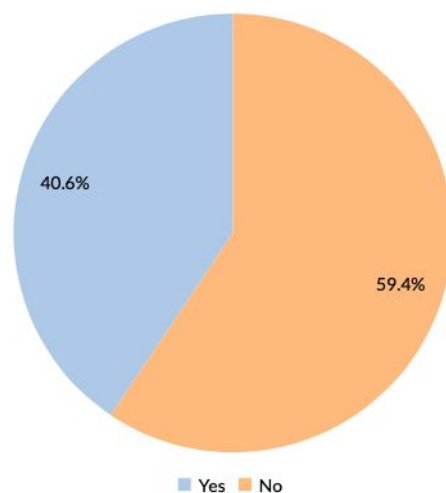
<https://fairsharing.org>

# Standard maintenance is a key point

Standard records that have maintainers



Standards that have a publication



59.3 % of standards have no maintainer

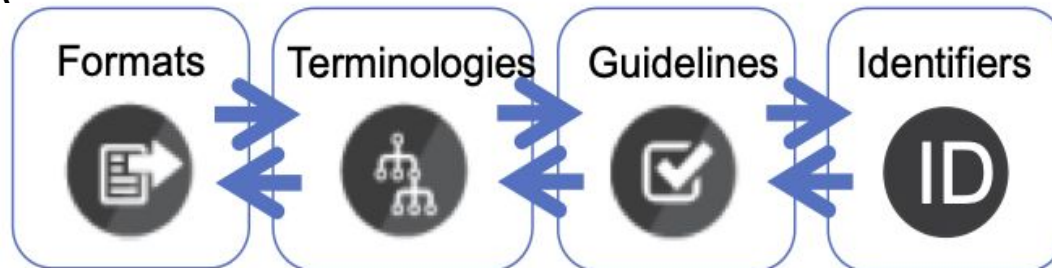
59.4% of standard has no publication

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



# Types of data standards

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

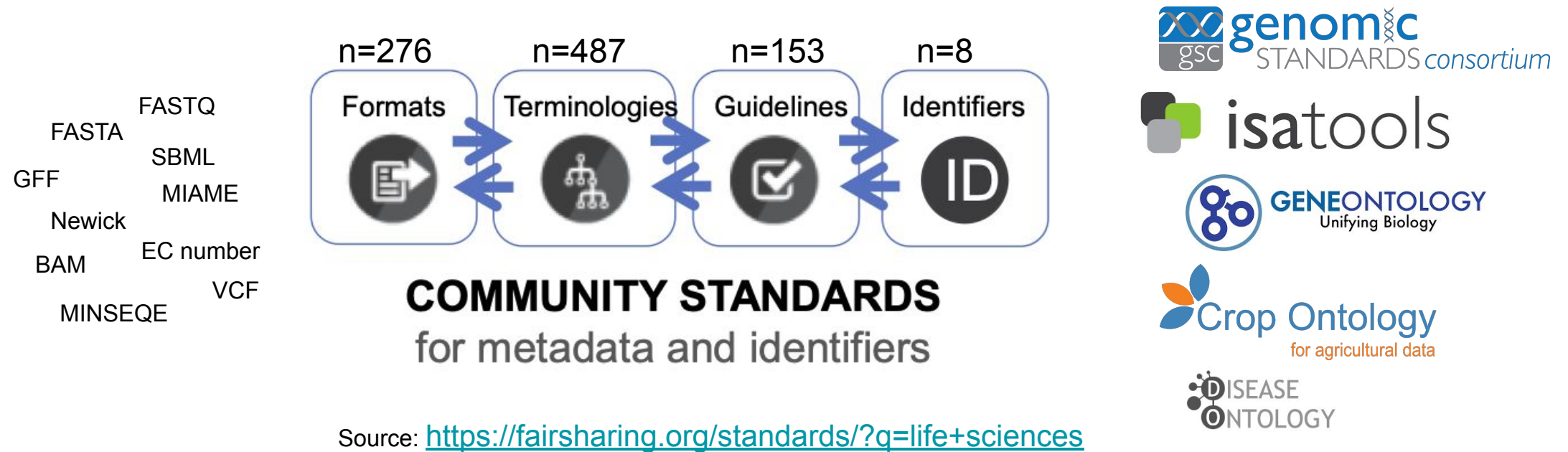


**Minimum information reporting requirements, checklists...**  
e.g. MIAME guidelines

**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



# 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 displays the FAIRsharing portal interface for a collection titled "COVID-19 Resources". The page features a navigation bar with tabs for Standards, Databases, Policies, Collections, Add/Claim Content, Stats, and Log In or Register. Below the navigation, there are sections for Subjects, User-defined Tags, and a "Compare with collection/recommendation (Beta)" dropdown. A "General collection/recommendation statistics" section shows stats for "COVID-19 Resources ( Insg-c000070)". The main content area displays a table of records, with the first record being "American Type Culture Collection database". The table columns include Registry Name, Abbreviation, Type, Subject, Domain, Taxonomy, Related Database, Related Standard, and Related Policy. The table shows 80 records in total, with the first page displaying 1-50 of 80.

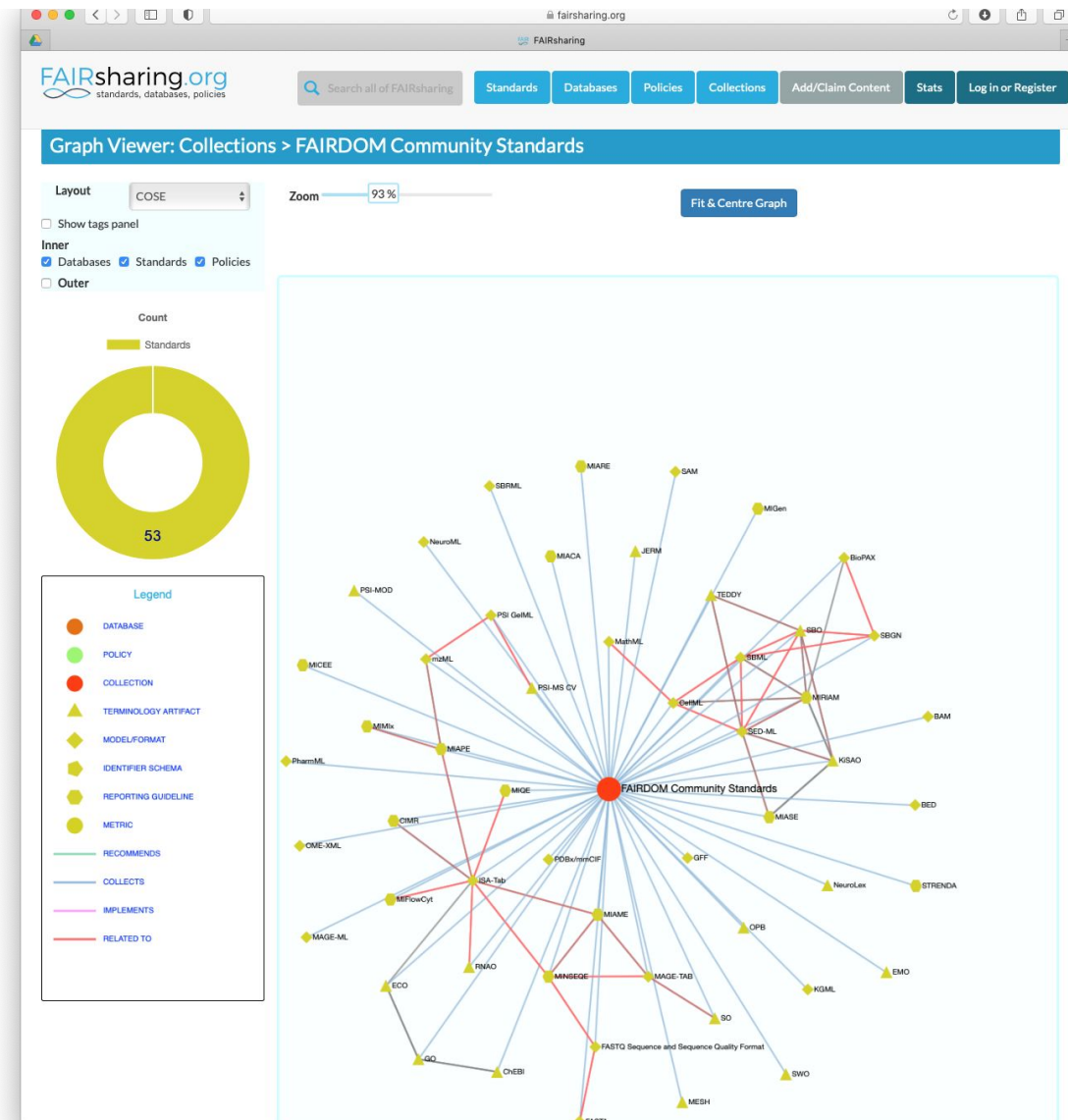
Registry Name	Abbreviation	Type	Subject	Domain	Taxonomy	Related Database	Related Standard	Related Policy	In Collection?
American Type Culture Collection database	ATCC	Database				None	None	Springer Nature Nature Medicine - Availability of Data and materials Nature Publishing Group - Nature Genetics - Availability of Data and materials Nature Publishing Group - Nature Chemistry - Availability of Data and materials Springer Nature - Nature Preprints - Availability of Data and materials Nature Publishing Group - Nature Structure & Molecular Biology - Availability of Data and materials Plus 8 more...	
Australian New Zealand Clinical Trials Registry	ANZCTR	Database						ClinicalTrials.gov ISRCTN Registry	
EBM/ERIC Directory	EBM/ERIC Directory	Database						ERIC - Culture	

# 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

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

Count  
Policies 4  
Standards 5  
Databases 70

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

Registry Name	Abbreviation	Type	Subject	Domain	Taxonomy	Related Database	Related Standard	Related Policy	In Collection?
Atkinson Type Culture Collection database	ATCC	Database							
Australian New Zealand Clinical Trial Registry	ANZCTR	Database							
EBM-COVID Directory	EBM-COVID	Database							

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

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

Count  
Policies 4  
Standards 5  
Databases 70

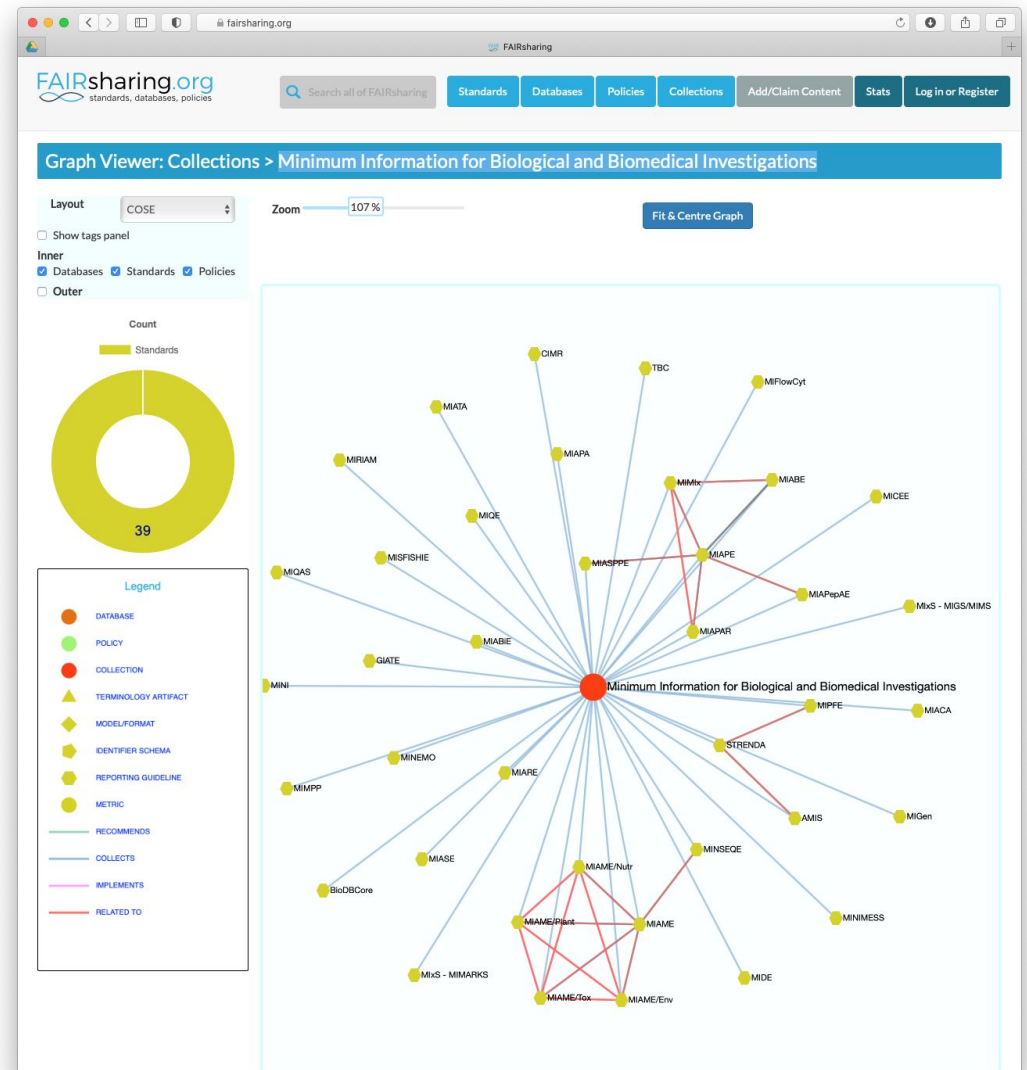
Legend  
● DATABASE  
● POLICY  
● COLLECTION  
▲ TERMINOLOGY ARTIFACT  
◆ MODEL/FORMAT  
◆ IDENTIFIER SCHEMA  
◆ REPORTING GUIDELINE  
● METRIC  
— RECOMMENDS  
— COLLECTS  
— IMPLEMENTS  
— RELATED TO

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

# What about the minimum required metadata in biology?

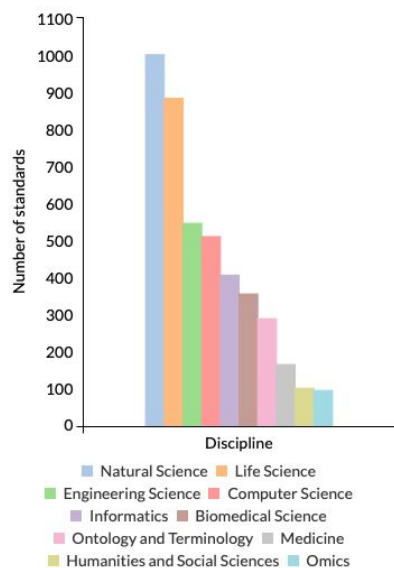
Example 3: the *Minimum Information for Biological and Biomedical Investigations* collection

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

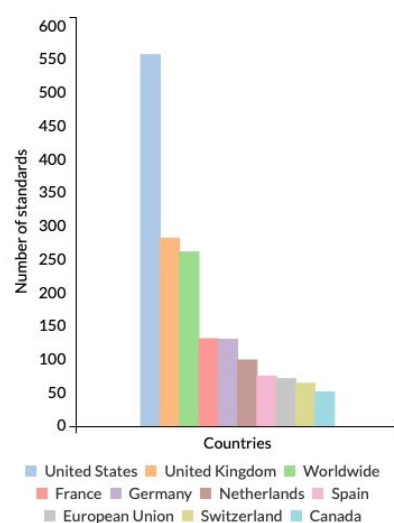


# Summary statistics about standards

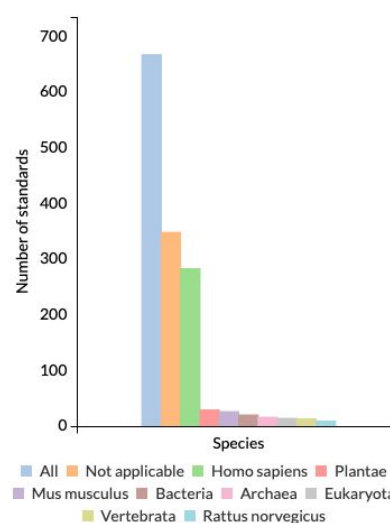
Top 10 disciplines covered by standards



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>



## Practice => Answers

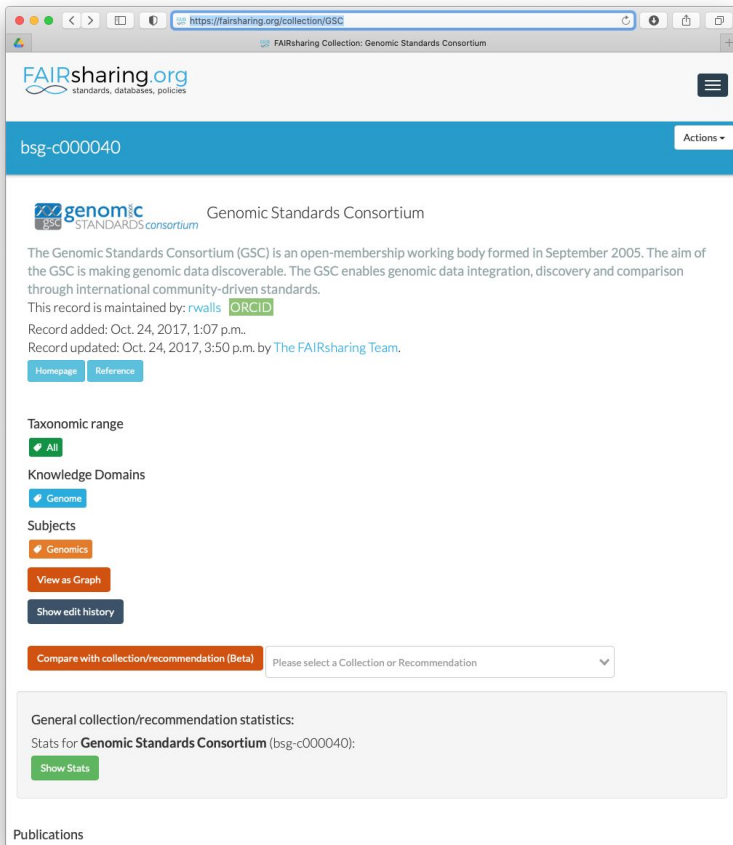
Find the *Genomic Standards Consortium (GSC)* used by both *ENA* and *SRA* databases in the **FAIRsharing collections**

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**

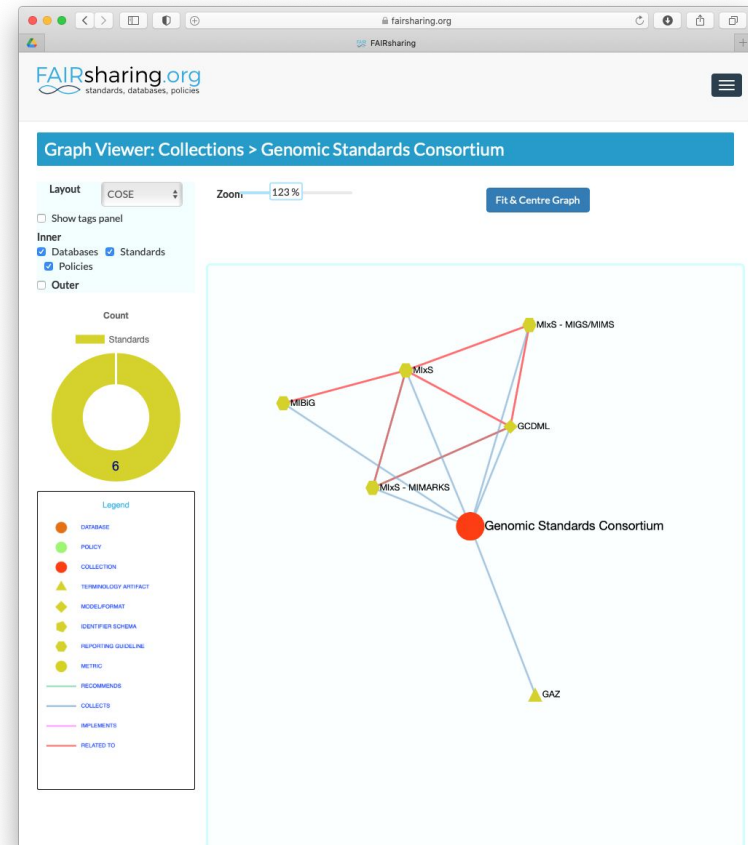
Source: <https://gensc.org>

# The Genomic Standards Consortium (GSC)



The screenshot shows the FAIRsharing.org interface for the collection 'bsg-c000040'. The page includes the FAIRsharing.org logo, the collection ID, and the Genomic Standards Consortium logo. A descriptive paragraph states: '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'. It also provides record dates: 'Record added: Oct. 24, 2017, 1:07 p.m.' and 'Record updated: Oct. 24, 2017, 3:50 p.m. by The FAIRsharing Team.' Navigation buttons for 'Homepage' and 'Reference' are visible. A 'Taxonomic range' section shows 'All' selected. 'Knowledge Domains' includes 'Genomic'. 'Subjects' includes 'Genomics'. A 'View as Graph' button is present. A 'Compare with collection/recommendation (Beta)' section has a dropdown menu. A 'General collection/recommendation statistics' section shows 'Stats for Genomic Standards Consortium (bsg-c000040):' with a 'Show Stats' button. A 'Publications' section is partially visible at the bottom.

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



The screenshot shows the 'Graph Viewer: Collections > Genomic Standards Consortium' interface. It features a 'Layout' dropdown set to 'COSE', a 'Zoom' slider at 123%, and a 'Fit & Centre Graph' button. A 'Show tags panel' checkbox is present. The 'Inner' section has checkboxes for 'Databases', 'Standards', and 'Policies', all of which are checked. The 'Outer' section has an unchecked checkbox. A donut chart shows a count of 6 for 'Standards'. A network graph displays nodes and relationships. The nodes include 'MIBIG', 'MixS', 'MixS - MIMARKS', 'MixS - MIGS/MIMS', 'GCDML', 'Genomic Standards Consortium', and 'GAZ'. The 'Genomic Standards Consortium' node is a red circle, while others are yellow circles or triangles. A legend on the left defines the symbols: orange circle for DATABASE, green circle for POLICY, red circle for COLLECTION, yellow triangle for TERMINOLOGY ABSTRACT, blue triangle for MODELFORMAT, yellow diamond for IDENTIFIER SCHEMA, yellow circle for REPORTING GUIDELINE, yellow circle for METRIC, green line for RECOMMENDS, blue line for COLLECTS, pink line for IMPLEMENTS, and red line for RELATED TO.

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

# The Genomic Standards Consortium (GSC)

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

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

[Yilmaz et al, 2011](#)

Source: <https://gensc.org>

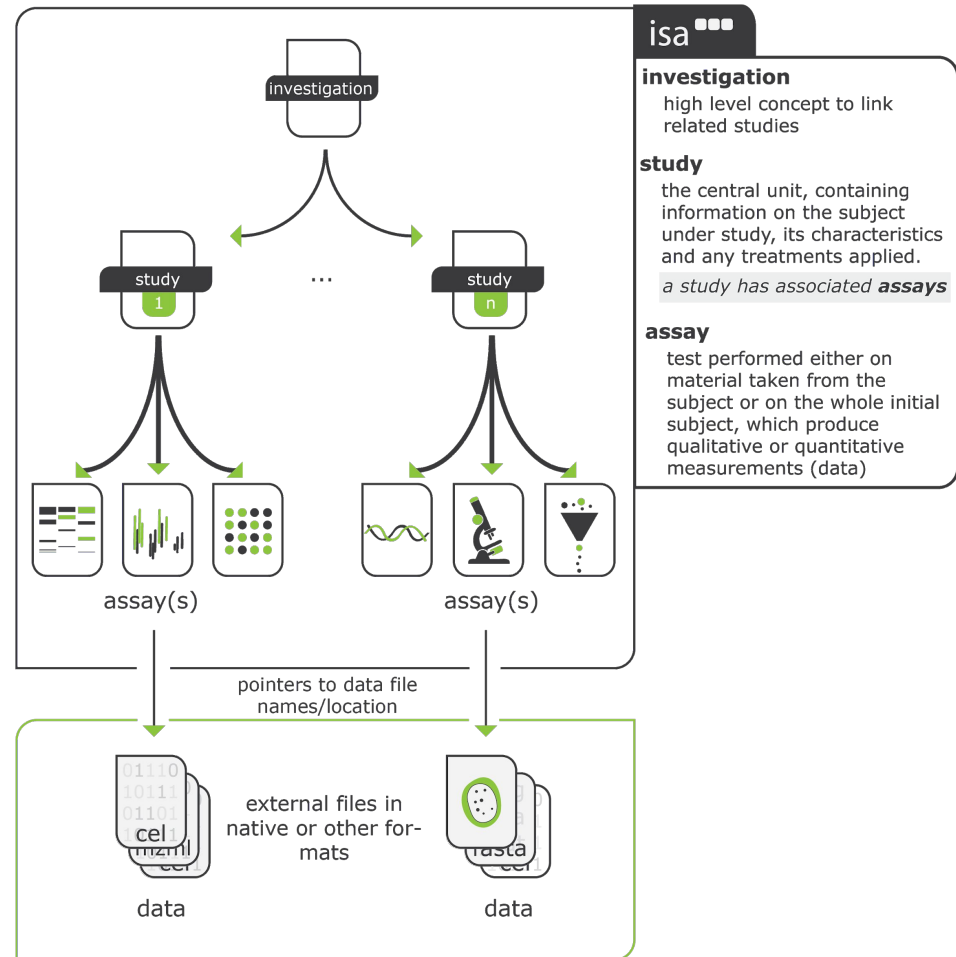
# The ISA model

## 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	<a href="https://fairsharing.org">https://fairsharing.org</a>
Investigation, Study, Assay (ISA) resource: A standard model and a set of tools to capture experimental data in life sciences	ISAtools	<a href="https://isa-tools.org">https://isa-tools.org</a>
Genomics Standard Consortium (GSC): An international consortium developing standards and checklists in genomics	GSC	<a href="https://gensc.org">https://gensc.org</a>
RDMkit: Documentation and metadata	RDMkit documentation and metadata	<a href="https://rdmkit.elixir-europe.org/metadata_management.html">https://rdmkit.elixir-europe.org/metadata_management.html</a>

# Thanks



Paulette Lieby



Jean-François Dufayard



Frédéric de Lamotte