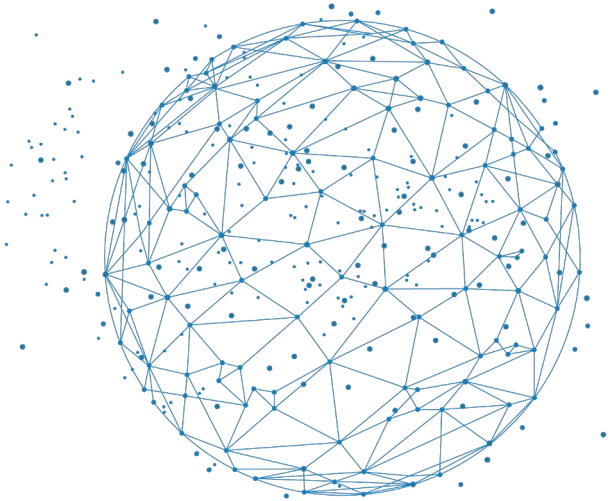
IBiSA Infrastructures en Biologie Santé et Agronomie



Standards in life sciences

Thomas Denecker
Institut Français de Bioinformatique

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE elixir FRANCE

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	Mot-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

Inspiré de 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

GENEONTOLOGY
Unifying Biology

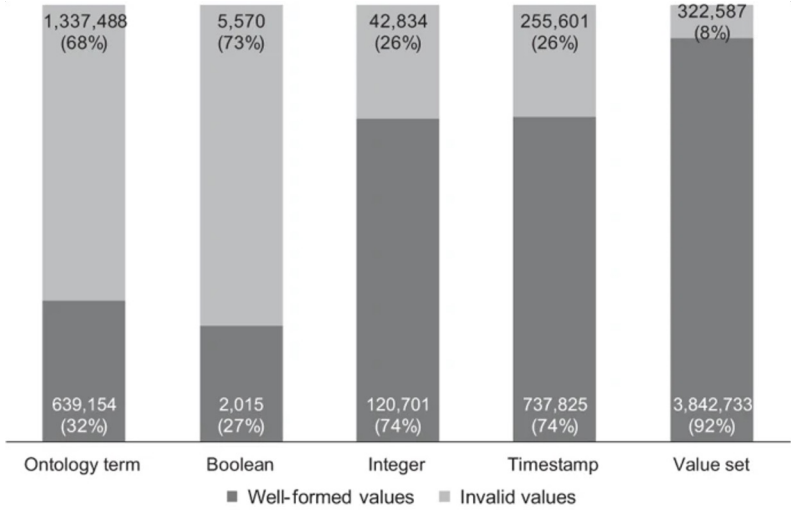
Do you know any standard in life sciences ?

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

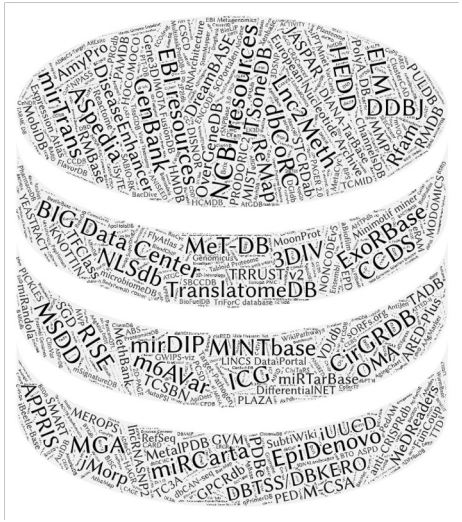


Quality of dictionary attributes in NCBI BioSample according to their type, in Gonçalves et al., 2019

Standard adoption and perenity

There are thousand of databases, softwares and resources in biology with **unequal level of standard adoption**

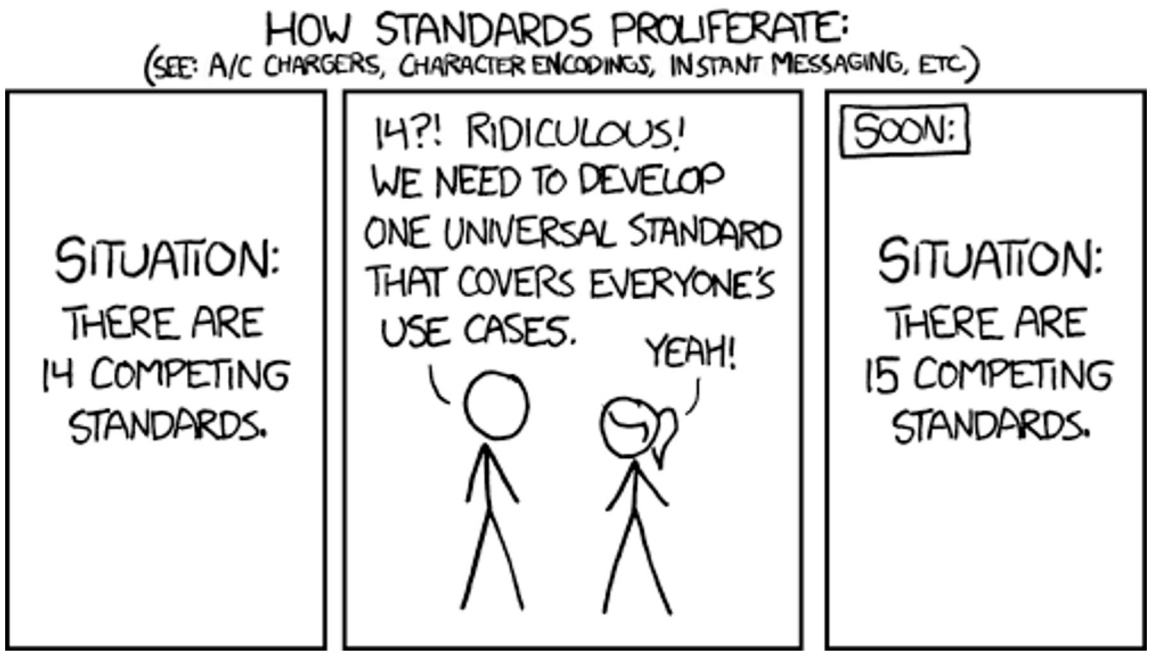
Is is not always easy for Life scientists and bioinformaticians to identify and use the most appropriate standards



1641 databases in NAR Database 2021

Rigden *et al*, 2021

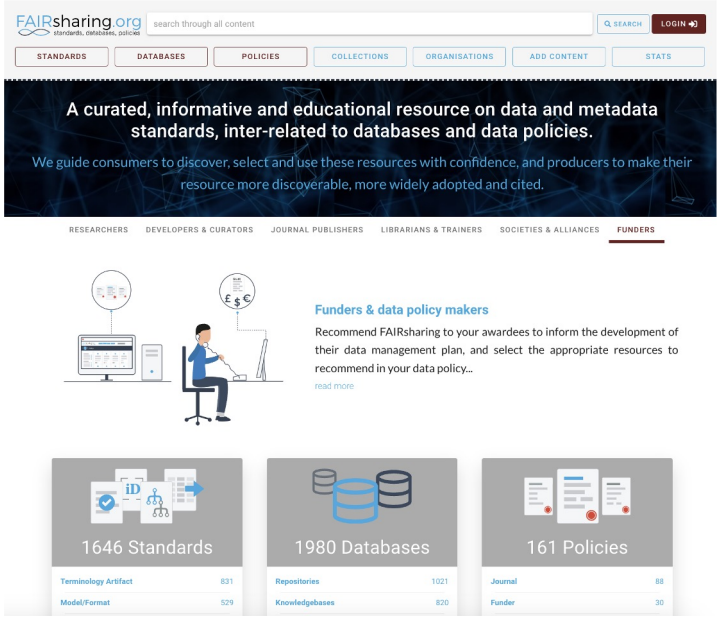
Standard adoption and perenity

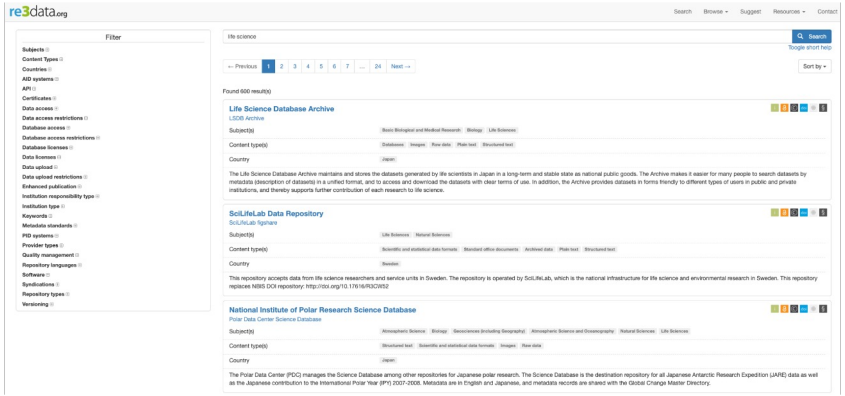


Source: https://xkcd.com/927/

How do I find the standard I need ?

FAIR sharing & re3data





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

re3data.org - Registry of Research Data Repositories. https://doi.org/10.17616/R3D last accessed: 2023-01-25

ENA

GENERAL INFORMATION

ENA European Nucleotide Archive

European Nucleotide Archive (ENA)

10.25504/FAIRsharing.dj8nt8

Type	Repository
Registry	Database
Description	The European Nucleotide Archive (ENA) is a globally comprehensive data resource for nucleotide sequence, spanning raw data, alignments and assemblies, functional and taxonomic annotation and rich contextual data relating to sequenced samples and experimental design. Serving both as the database of record for the output of the world's sequencing activity and as a platform for the management, sharing and publication of sequence data, the ENA provides a portfolio of services for submission, data management, search and retrieval across web and programmatic interfaces. The ENA is part of the International Nucleotide Sequence Database Collaboration (INSDC), which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at the NCBI. These three organizations exchange data on a daily basis.
Homepage	http://www.ebi.ac.uk/ena
Year of Creation	1980
Maintainers	cochrane
Countries developing this resource	Japan, United States, European Union
Subjects	Functional Genomics, Metagenomics, Genomics, Bioinformatics, Data Management, Biodiversity, Transcriptomics
Domains	DNA Sequence Data, Annotation, Sequence Annotation, Genomic Assembly, Histone, Deoxyribonucleic Acid, Ribonucleic Acid, Nucleotide, Sequencing, Amino Acid Sequence, Data Storage
Taxonomic Range	All
User Defined Tags	Data Coordination

VIEW RELATION GRAPH

How to cite this record

FAIRsharing.org: ENA; European Nucleotide Archive, **DOI:** 10.25504/FAIRsharing.dj8nt8, **Last Edited: Wednesday, May 4th 2022, 14:41, Last Editor:** delphinedauga, **Last Accessed:** Wednesday, January 25th 2023, 16:04

Publication for citation

European Nucleotide Archive in 2016.; Toribio AL, Alako B., et al., (2016) 10.1093/nar/gkw1106

The following community curators have contributed to this record: kccpas@gbif.org

Record ID 1890 | **Record created at** Tuesday, November 4th 2014, 16:23 | **Record updated at** Wednesday, July 20th 2022, 11:30

COLLECTIONS & RECOMMENDATIONS

Search through in collections

IN COLLECTIONS (7) | IN POLICIES (46)

- Data Citation Implementation — Data Citation Implementation **collects** European Nucleotide Archive
- EOSC-Life — EOSC-Life **collects** European Nucleotide Archive

RELATED CONTENT

Search through related standards

RELATED STANDARDS (13) | RELATED DATABASES (32)

- Short Read Archive eXtensible Markup Language — European Nucleotide Archive **implements** Short Read Archive eXtensible Markup Language
- NCBI Taxonomy — European Nucleotide Archive **implements** NCBI Taxonomy

https://fairsharing.org/FAIRsharing.e1byny

Repository details

European Nucleotide Archive

General | Institutions | Terms | Standards

Name of repository	**European Nucleotide Archive**
Additional name(s)	ENA
Repository URL	https://www.ebi.ac.uk/ena/browser/home
Subject(s)	Biology, Medicine, General Genetics, Bioinformatics and Theoretical Biology, Microbiology, Virology and Immunology, Life Sciences, Basic Biological and Medical Research
Description	The European Nucleotide Archive (ENA) captures and presents information relating to experimental workflows that are based around nucleotide sequencing. A typical workflow includes the isolation and preparation of material for sequencing, a run of a sequencing machine in which sequencing data are produced and a subsequent bioinformatic analysis pipeline. ENA records this information in a data model that covers input information (sample, experimental setup, machine configuration), output machine data (sequence traces, reads and quality scores) and interpreted information (assembly, mapping, functional annotation). Data arrive at ENA from a variety of sources. These include submissions of raw data, assembled sequences and annotation from small-scale sequencing efforts, data provision from the major European sequencing centres and routine and comprehensive exchange with our partners in the International Nucleotide Sequence Database Collaboration (INSDC). Provision of nucleotide sequence data to ENA or its INSDC partners has become a central and mandatory step in the dissemination of research findings to the scientific community. ENA works with publishers of scientific literature and funding bodies to ensure compliance with these principles and to provide optimal submission systems and data access tools that work seamlessly with the published literature.
Contact	https://www.ebi.ac.uk/ena/browser/support
Content type(s)	Scientific and statistical data formats, Networkbased data, Structured graphics, Plain text, Software applications, Raw data, Structured text
Keyword(s)	CRAM, DNA, RNA, epigenomics, genome, human genetics, metagenomics, protein coding, samples, taxonomy data, transcripts
Persistent identifier(s) of the repository	RRID:nif-0000-32981 RRID:SCR_006515 OMICS_01029 MIR:00000372 FAIRsharing_doi:10.25504/FAIRsharing.dj8nt8
Repository type(s)	disciplinary
Mission statement for designated community	https://www.ebi.ac.uk/ena/browser/about
Research data repository language(s)	English
Data and/or service provider	data provider service provider

Back to search | Submit a change request | Get a badge

DataCite

Cite this re3data.org record:

re3data.org: European Nucleotide Archive; editing status 2021-12-13; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R3HW3J last accessed: 2023-01-25

https://www.re3data.org/repository/r3d100010527

Pride

GENERAL INFORMATION

PRIDE — PRoteomics IDEntifications database (PRIDE)

10.25504/FAIRsharing.e1byny

Type	Repository
Registry	Database
Description	The PRIDE PRoteomics IDEntifications (PRIDE) Archive database is a centralized, standards compliant, public data repository for mass spectrometry proteomics data, including protein and peptide identifications and the corresponding expression values, post-translational modifications and supporting mass spectra evidence (both as raw data and peak list files). PRIDE is a core member in the ProteomeXchange (PX) consortium, which provides a standardised way for submitting mass spectrometry based proteomics data to public-domain repositories.
Homepage	http://www.ebi.ac.uk/pride/
Year of Creation	2004
Maintainers	prideebi
Countries developing this resource	Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Lithuania, Luxembourg, Malta, Montenegro, Netherlands, Norway, Portugal, Slovakia, Spain, Sweden, Switzerland, United Kingdom
Subjects	Proteomics
Domains	Mass Spectrum, Peptide Identification, Protein Identification, Peptide, Post-translational Protein Modification, Mass Spectrometry Assay, Curated Information, Protein
Taxonomic Range	All
User Defined Tags	None

VIEW RELATION GRAPH

How to cite this record

FAIRsharing.org: PRIDE, PRoteomics IDEntifications database, **DOI:** 10.25504/FAIRsharing.e1byny, **Last Edited:** Monday, October 3rd 2022, 9:07, **Last Editor:** allysonlister, **Last Accessed:** Wednesday, January 25th 2023, 15:58

Publication for citation

The PRIDE database and related tools and resources in 2019: improving support for quantification data.: Perez-Riverol Y,Csordas A,Bai J,Bernal-Llinares M,Hewapathirana S,Kundu DJ,Inuganti A,Griss J,Mayer G,Eisenacher M,Perez E,Uszkoreit J,Pfeuffer J,Sachsenberg T,Yilmaz S,Tiwary S,Cox J,Audain E,Walzer M,Jarnuczak AF,Ternent T,Brazma A,Vizcaino JA (2018) 10.1093/nar/gky1106

Record ID 1863 | **Record created at** Tuesday, November 4th 2014, 16:23 | **Record updated at** Monday, October 3rd 2022, 9:07

COLLECTIONS & RECOMMENDATIONS

Search through in collections

IN COLLECTIONS (4) | IN POLICIES (36)

- Data Citation Implementation — Data Citation Implementation collects PRoteomics IDEntifications database
- EOSC-Life — EOSC-Life collects PRoteomics IDEntifications database

RELATED CONTENT

Search through related standards

RELATED STANDARDS (15) | RELATED DATABASES (16)

- mz Quantitative Markup Language — PRoteomics IDEntifications database implements mz Quantitative Markup Language
- Transition Markup Language — PRoteomics IDEntifications database implements Transition Markup Language

https://fairsharing.org/FAIRsharing.e1byny

Repository details

PRoteomics IDEntifications Database

General | Institutions | Terms | Standards

Name of repository	**PRoteomics IDEntifications Database**
Additional name(s)	PRIDE
Repository URL	https://wwwdev.ebi.ac.uk/pride/
Subject(s)	Biology, Biochemistry, Bioinformatics and Theoretical Biology, Analytical Chemistry, Method Development (Chemistry), Life Sciences, Basic Biological and Medical Research, Analytical Chemistry, Method Development (Chemistry), Chemistry, Natural Sciences
Description	The PRIDE PRoteomics IDEntifications database is a centralized, standards compliant, public data repository for proteomics data, including protein and peptide identifications, post-translational modifications and supporting spectral evidence. PRIDE encourages and welcomes direct user submissions of mass spectrometry data to be published in peer-reviewed publications.
Content type(s)	Structured text, Standard office documents, Images, Structured graphics, Plain text, Scientific and statistical data formats, Software applications, Archived data
Keyword(s)	amino acids, mass spectra, mass spectrometry, peptide identifications, protein identifications, proteomics, sequencing
Persistent identifier(s) of the repository	RRID:nif-0000-03336 RRID:SCR_003411 OMICS_02456 MIR:00100094 FAIRsharing_doi:10.25504/FAIRsharing.e1byny
Repository size	13.226 datasets
Repository type(s)	disciplinary
Mission statement for designated community	https://wwwdev.ebi.ac.uk/pride/markdownpage/citationpage
Research data repository language(s)	English
Data and/or service provider	data provider

Back to search | Submit a change request | Get a badge

DataCite

Cite this re3data.org record:

re3data.org: PRoteomics IDEntifications Database; editing status 2021-11-09; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R3JG6V last accessed: 2023-01-25

https://www.re3data.org/repository/r3d100010137

BioImage Archive

GENERAL INFORMATION

BioImage Archive (BioImage Archive)

10.25504/FAIRsharing.x38D2k

Type	Repository
Registry	Database
Description	The BioImage Archive stores and distributes biological images that are useful to life-science researchers. Its development will provide data archiving services to the broader bioimaging database community. This includes added-value bioimaging data resources such as EMPIAR, Cell-IDR and Tissue-IDR.
Homepage	https://www.ebi.ac.uk/bioimage-archive/
Year of Creation	2019
Maintainers	ugis, matthewh
Countries developing this resource	Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Lithuania, Luxembourg, Malta, Montenegro, Netherlands, Norway, Portugal, Slovakia, Spain, Sweden, Switzerland, United Kingdom
Subjects	Molecular Biology, Clinical Studies, Life Science, Cell Biology, Biomedical Science, Preclinical Studies
Domains	Bioimaging, Cell, Microscopy, Light Microscopy, Electron Microscopy, Imaging, Super-resolution Microscopy, Image, Histology, Whole Mount Tissue, Tissue, High-content Screen
Taxonomic Range	All
User Defined Tags	Pre-Clinical Imaging

VIEW RELATION GRAPH

How to cite this record

FAIRsharing.org: BioImage Archive; BioImage Archive, **DOI:** 10.25504/FAIRsharing.x38D2k, **Last Edited: Thursday, November 24th 2022, 14:48, Last Editor:** ramon.granell, **Last Accessed:** Tuesday, January 31st 2023, 13:49

Publication for citation

A call for public archives for biological image data.: Ellenberg J,Swedlow JR,Barlow M,Cook CE,Sarkans U,Patwardhan A,Brazma A,Birney E (2018) 10.1038/s41592-018-0195-8

Record ID 2830 | **Record created at** Thursday, October 17th 2019, 18:23 | **Record updated at** Wednesday, December 7th 2022, 16:57

COLLECTIONS & RECOMMENDATIONS

Search through in collections

IN COLLECTIONS (5) | IN POLICIES (1)

EOSC-Life

EOSC-Life collects BioImage Archive

RELATED CONTENT

Search through related standards

RELATED STANDARDS (1) | RELATED DATABASES (3)

Recommended Metadata for Biological Images

BioImage Archive implements Recommended Metadata for Biological Images

https://fairsharing.org/FAIRsharing.x38D2k

Repository details

BioImage Archive

General | Institutions | Terms | Standards

Name of repository	**BioImage Archive**
Additional name(s)	BIA
Repository URL	https://www.ebi.ac.uk/bioimage-archive
Subject(s)	Basic Biological and Medical Research, Life Sciences, Bioinformatics and Theoretical Biology, Biology
Description	The BioImage Archive stores and distributes life sciences imaging datasets. It supports deposition of biological imaging data associated with publications for the whole research community, as well as reference imaging datasets. All data deposited to the BioImage Archive is made openly accessible to the scientific community.
Contact	bioimage-archive@ebi.ac.uk https://www.ebi.ac.uk/bioimage-archive/contact-us/
Content type(s)	Images, Raw data, Scientific and statistical data formats, Archived data
Keyword(s)	bioimaging, high-content screening, microscopy
Persistent identifier(s) of the repository	FAIRsharing_doi:10.25504/FAIRsharing.x38D2k
Repository size	694 datasets
Repository type(s)	disciplinary
Mission statement for designated community	https://www.ebi.ac.uk/bioimage-archive/scope/
Research data repository language(s)	English
Data and/or service provider	data provider service provider

Back to search | Submit a change request | Get a badge

DataCite

Cite this re3data.org record:

re3data.org: BioImage Archive; editing status 2022-08-13; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R31NJN99 last accessed: 2023-01-31

https://www.re3data.org/repository/r3d100013949

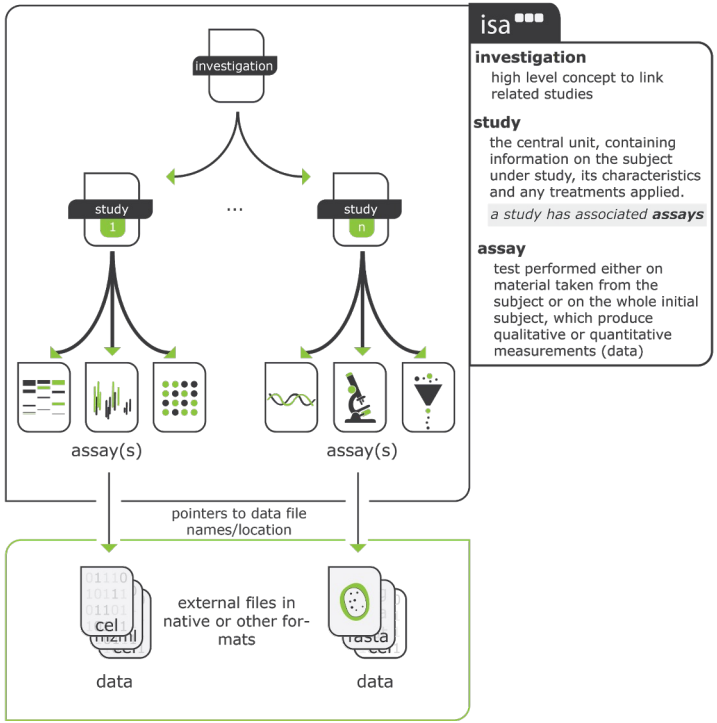
Focus

The ISA model

A standard for Life Science Data

A **model** to capture **experimental metadata** through **3 core entities**:

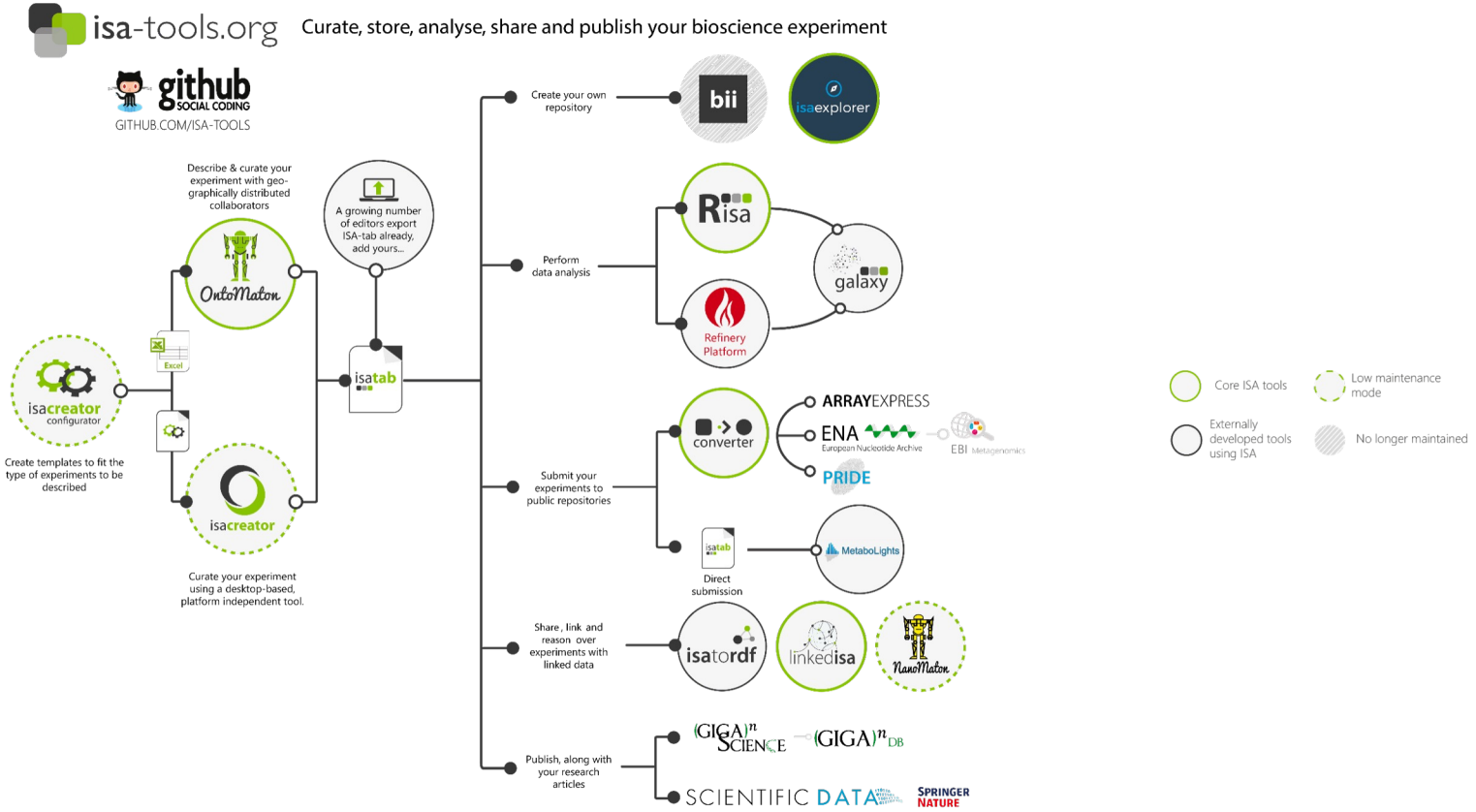
- **Investigation**: the project context
- **Study**: an experimentation in one location
- **Assay:** a specific measurement that targets a trait with a method and a scale

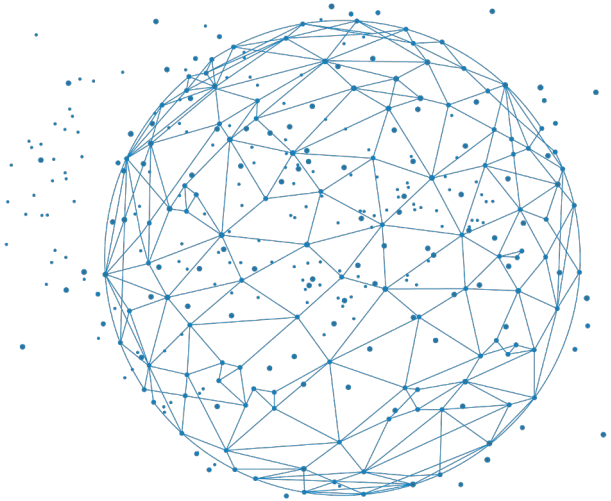


ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Rocca-Serra P et al. **Bioinformatics 2010**. https://doi.org/10.1093/bioinformatics/btq415

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

ISA tools


isa-tools.org
Curate, store, analyse, share and publish your bioscience experiment
github
SOCIAL CODING
GITHUB.COM/ISA-TOOLS
Describe & curate your experiment with geo-graphically distributed collaborators
OntoMaton
A growing number of editors export ISA-tab already, add yours...
isatab
Excel
isacreator
configurator
Create templates to fit the type of experiments to be described
isacreator
Curate your experiment using a desktop-based, platform independent tool.
Create your own repository
bii
isaexplorer
Perform data analysis
Risa
galaxy
Refinery Platform
Submit your experiments to public repositories
converter
ARRAYEXPRESS
ENA
European Nucleotide Archive
EBI Metagenomics
PRIDE
isatab
Direct submission
MetaboLights
Share, link and reason over experiments with linked data
isatordf
linkedisa
NanoMaton
Publish, along with your research articles
(GIGA)n SCIENCE
(GIGA)n DB
SCIENTIFIC DATA
SPRINGER NATURE
Core ISA tools
Low maintenance mode
Externally developed tools using ISA
No longer maintained



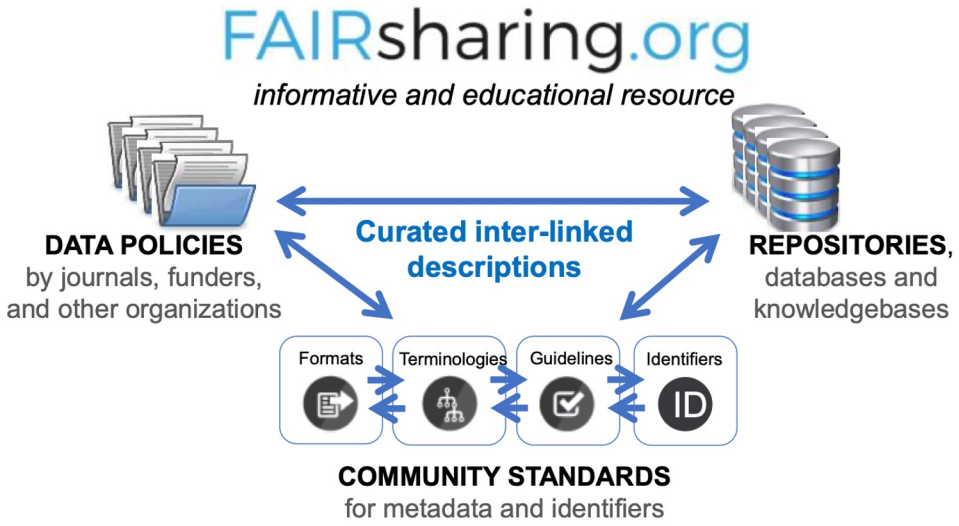
Supplementary slides

The FAIRsharing portal

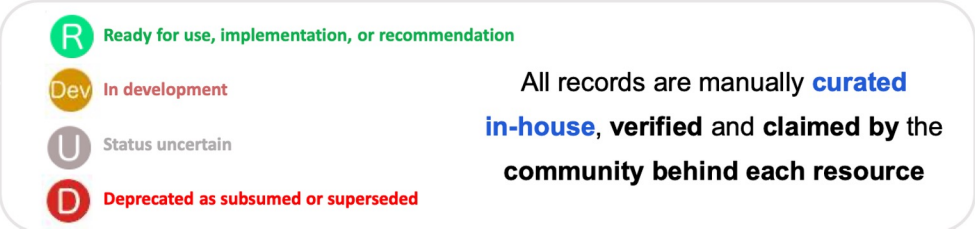
Citable *DOI* for all records

Accessible via *API* or *web interface*

Curation

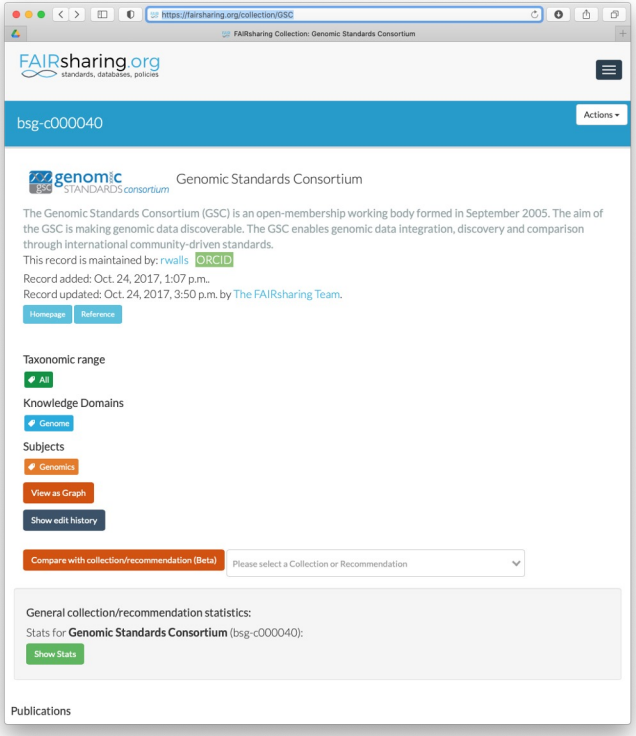


RECORD STATUS

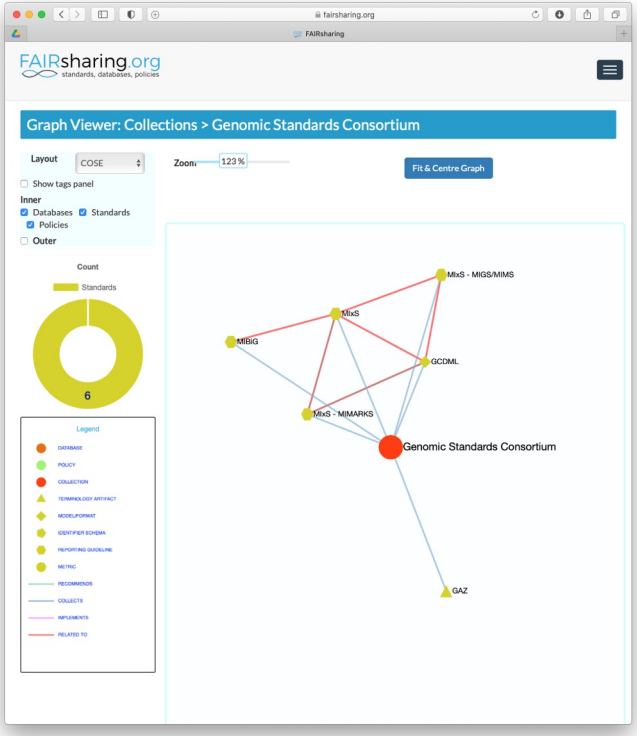


https://fairsharing.org

The Genomic Standards Consortium (GSC)



https://fairsharing.org/collection/GSC



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

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

R Ready for use, implementation, or recommendation

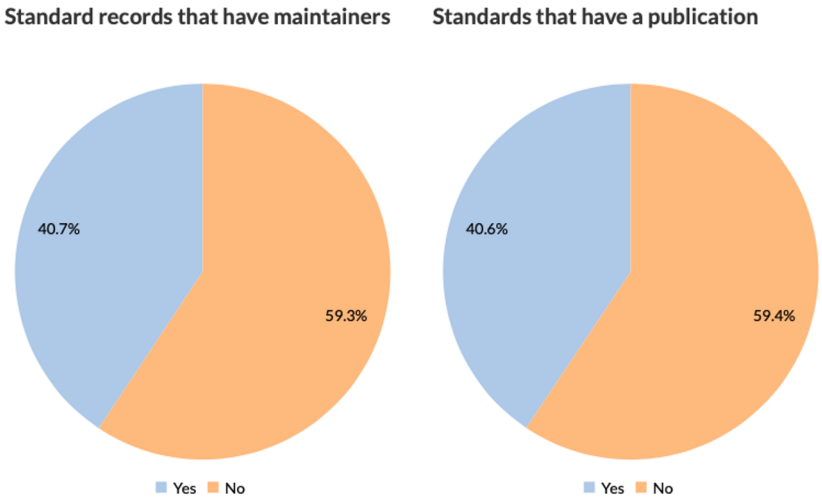
Dev In development

U Status uncertain

D Deprecated as subsumed or superseded

https://fairsharing.org

Standard maintenance is a key point



59.3 % of standards have no maintainer

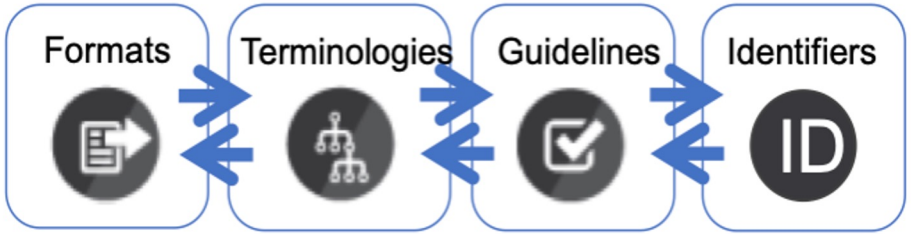
59.4% of standard has no publication

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

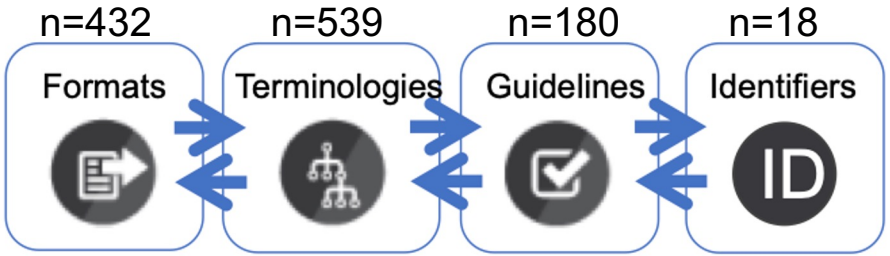


Controlled vocabularies, taxonomies, ontologies…
e.g. Gene Ontology

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



COMMUNITY STANDARDS
for metadata and identifiers

genomic STANDARDS consortium

isatools

GENEONTOLOGY
Unifying Biology

Crop Ontology
for agricultural data

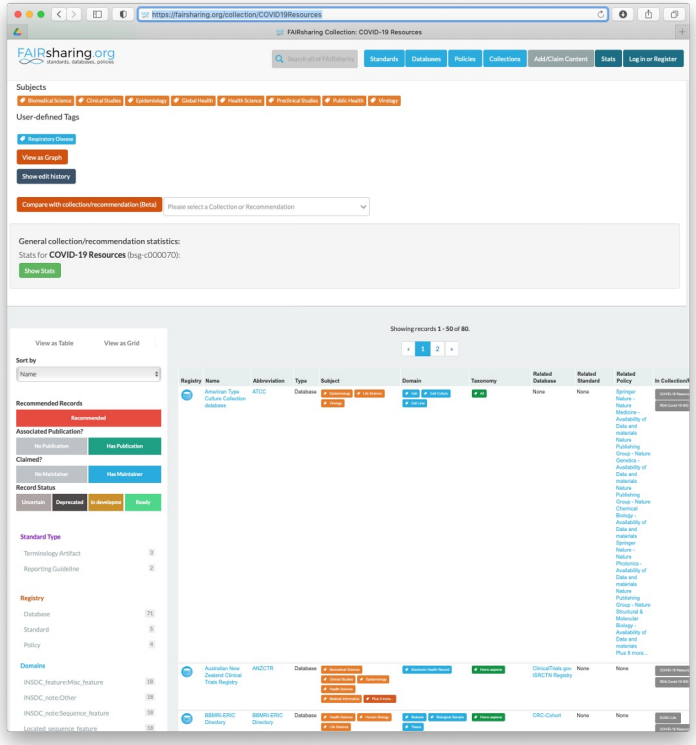
DISEASE ONTOLOGY

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

Collections in the FAIRsharing portal

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

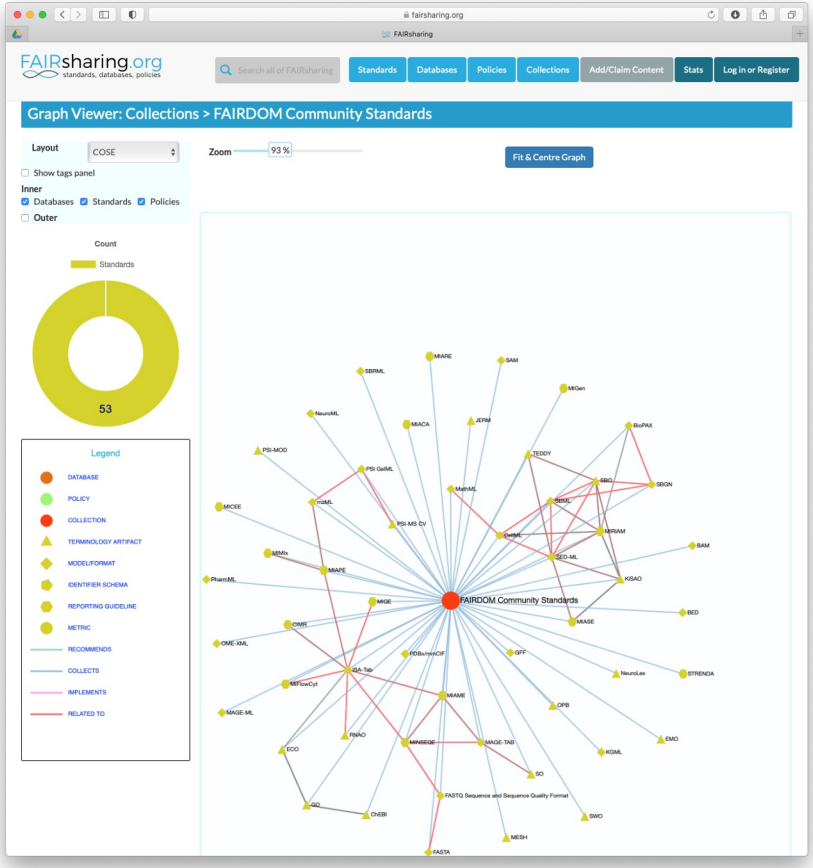
Graph view to visualize relationship links between resources



https://fairsharing.org/collections/

Collections in Life Sciences

63 collections related to Life Science standards in FAIRsharing

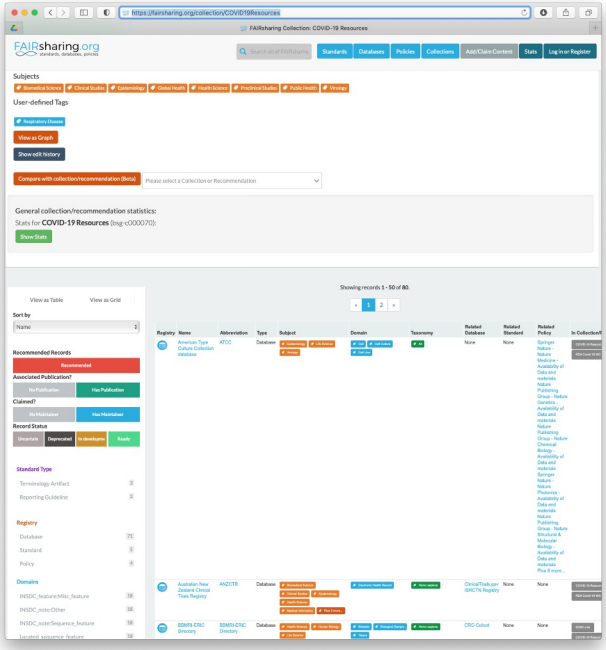


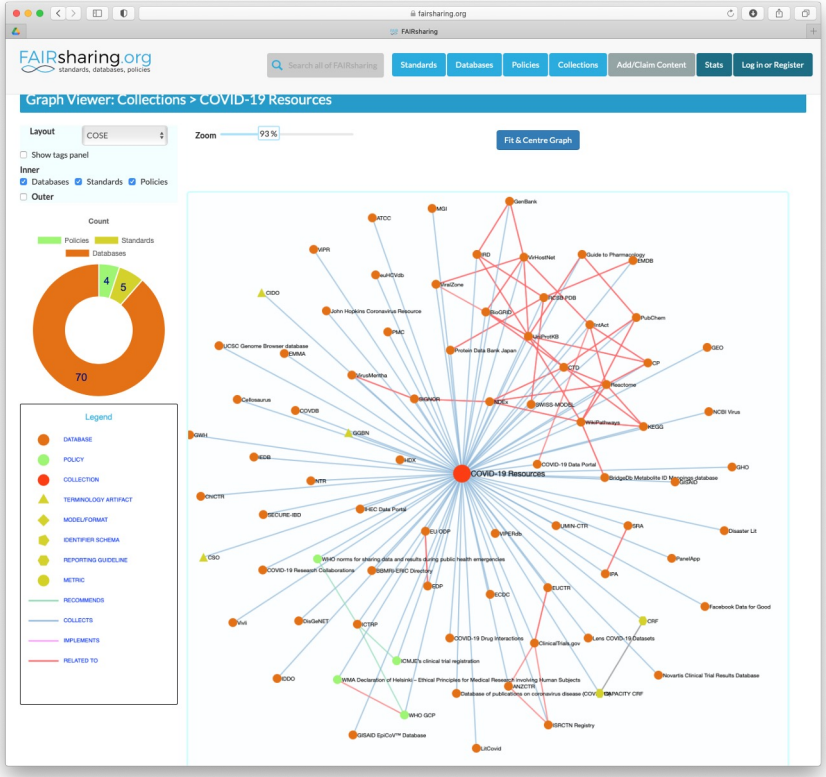
Example 1: the *FAIRdom community Standards collection* (System biology)

https://fairsharing.org/collection/FAIRDOM

Some collections are recent

Example 2: The *Covid-19* collection



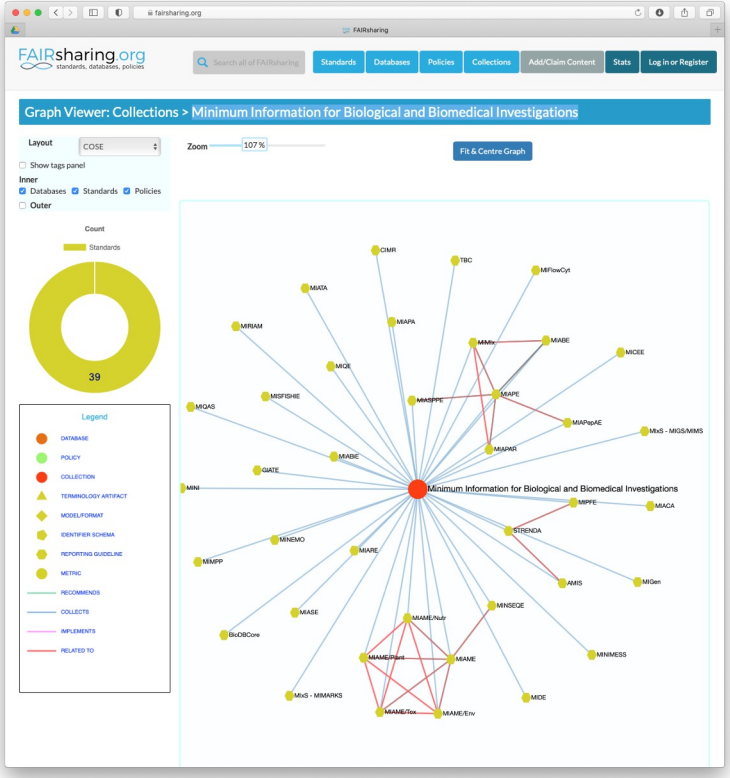


https://fairsharing.org/collection/COVID19Resources

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

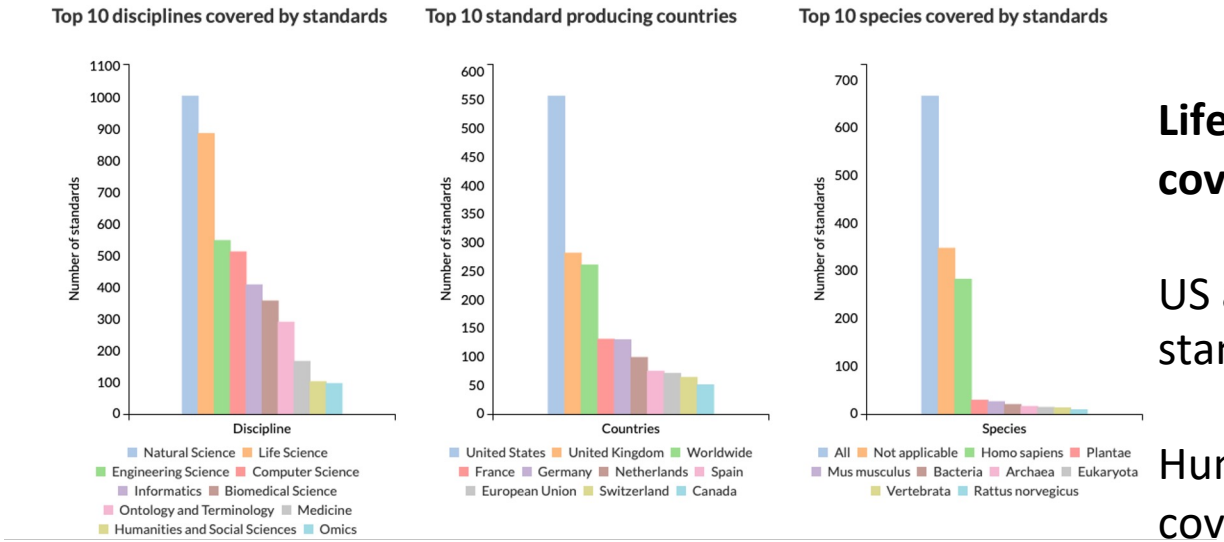
What about the minimum required metadata in biology?

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

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:

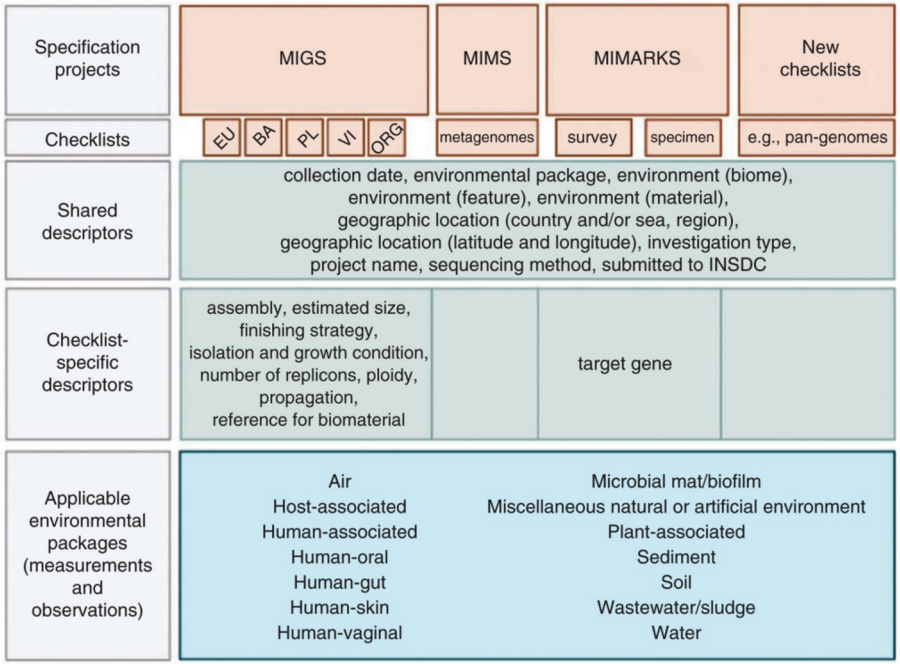
1. How many records (*i.e.* standards) are associated to the *GSC* ?
2. What type of standard is *Minimum Information about any (x) Sequence (MiXS)* ?
3. What is the record status of the *GAZ* record ?

Source: https://gensc.org

The Genomic Standards Consortium (GSC)

- 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