Metadata standards and controlled vocabularies

Federico Bianchini
Centre for Bioinformatics, University of Oslo

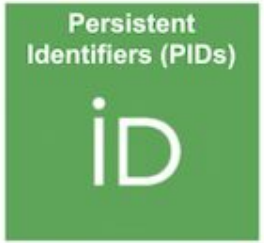
20 October 2023

Life Science RDM

What is a metadata standard?

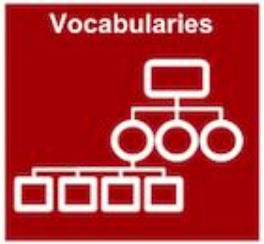
- How does a standard make data more FAIR?
- What are the ingredients required for defining a standard?
- Where to find metadata standards?
- Which tools can be used in connection with standards?

FAIR illustrations icons from ARDC 2018 licensed under Creative Commons Attribution 4.0 International License



- The metadata requires a persistent identifier
- Metadata standard compliant with repositories checklist
- Metadata fields used to identify/retrieve the data

- Persistency of metadata (also when the data is not available)



- Interoperability fully relies on metadata
- Vocabularies and ontologies ensure that standards are FAIR
- A well-defined standard can be linked with standards describing other type of data
- Focus on machine-actionability

- Richer metadata fields enhance reusability
- Metadata should describe data provenance
- Everything needs to follow community standards
  - Alignment with repository

ArrayExpress

MINSEQE
MIAME
...

PRIDE Archive
PRoteomics IDEntifications Database

HUPO-PSI
TraML
MIAPE
...

ENA
European Nucleotide Archive

SRA-XML

Sample Checklists

There is a minimum amount of information required during ENA sample registration and all samples must conform to a defined checklist of expected metadata values. The most suitable checklist for sample registration depends on the type of the sample.

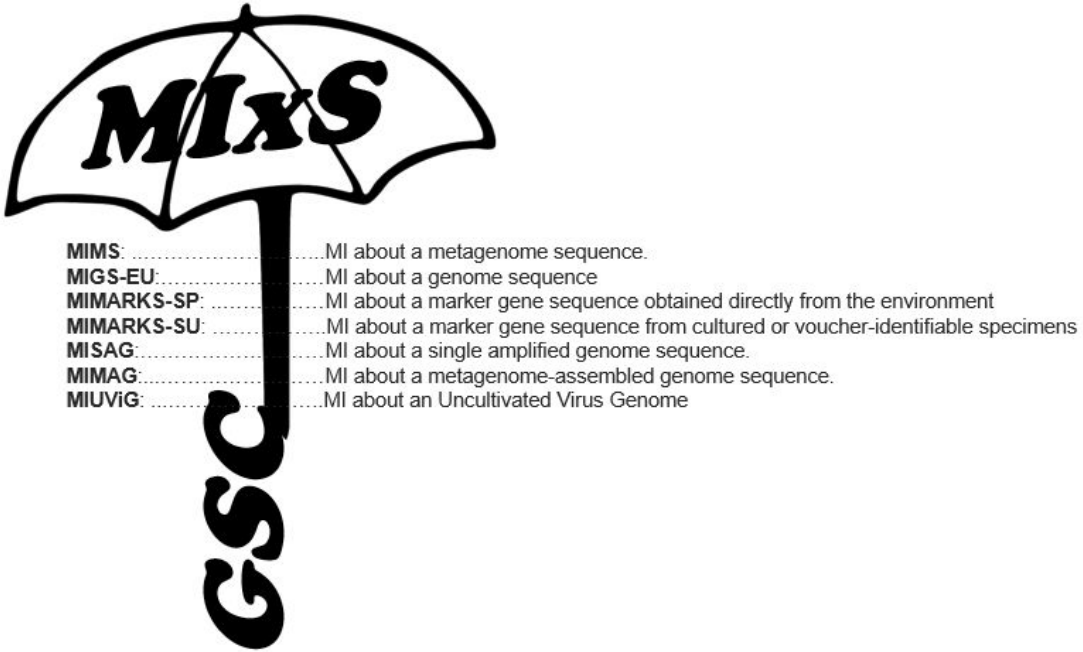
https://www.ebi.ac.uk/ena/browser/checklists

ENA
European Nucleotide Archive

genomic STANDARDS consortium

Sample Checklists

Accession	Name
ERC000012	GSC MIxS air
ERC000013	GSC MIxS host associated
ERC000014	GSC MIxS human associated



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

https://www.gensc.org/pages/standards-intro.html

ERC000014 **GSC MIxS human associated**

Field Name	Field Format (Field Restriction)
project name	free text
experimental factor	free text
ploidy	free text
number of replicons	restricted text regular expression

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

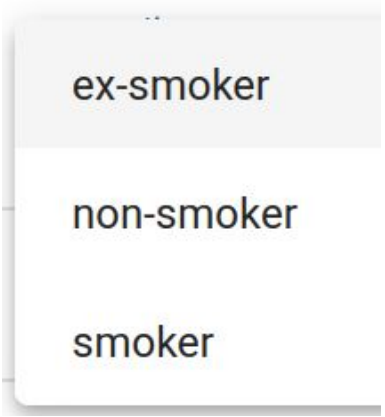
ERC000014	**GSC MIxS human associated**

number of replicons		restricted text	regular expression [+-]?[0-9]+
smoker		text choice	ex-smoker non-smoker smoker

https://www.ebi.ac.uk/ena/browser/view/ERC000014

Controlled Vocabularies

These lists of pre-selected answers for a metadata field are known as **controlled vocabularies**. They are useful to avoid situations such the one below, where things got out of control



How many ways can you say “female”?

18-day pregnant females
2 yr old female
400 yr. old female
adult female
asexual female
castrate female
cf.female
cystocarpic female
dikaryon
dioecious female
diploid female
female (lactating)
female (pregnant)
female (outbred)
female parent
female plant
female with eggs
female worker
female, 6-8 weeks old
female, virgin
female, worker
female(gynoecious)
individual female
lgb*cc females
mare
female (worker)
monosex female
ovigerous female
oviparous sexual females
worker bee
female enriched
pseudohermaprhoditic female
remale
worker caste (female)
sex: female
female, other
female child
femal
3 female
female (phenotype)
female mice
female, spayed
femlale
metafemale

How many ways can you say “female”?

18-day pregnant females
2 yr old female
400 yr. old female
adult female
asexual female
castrate female
cf.female
cystocarpic female
dikaryon
dioecious female
diploid female
f
famale
femail
female
female - worker
female (alate sexual)
female (calf)
hen
female (lactating)
female (pregnant)
female (outbred)
female parent
female plant
female with eggs
female worker
female, 6-8 weeks old
female, virgin
female, worker
female(gynoecious)
femele
female, pooled
femalen
females
females only
gynoecious
healthy female
individual female
lgb*cc females
mare
female (worker)
monosex female
ovigerous female
oviparous sexual females
worker bee
female enriched
pseudohermaprhoditic female
remale
semi-engorged female
sexual oviparous female
sterile female worker
strictly female
tetraploid female
thelytoky
female (gynoecious)
worker caste (female)
sex: female
female, other
female child
femal
3 female
female (phenotype)
female mice
female, spayed
femlale
metafemale
sterile female
normal female
sf
vitellogenic replete female
worker
hexaploid female
female (f-o)
probably female (based on morphology)

female (note: this sample was originally provided as a \"male\" sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a female individual)"

Courtesy of N. Silvester, European Nucleotide Archive, EMBL-EBI

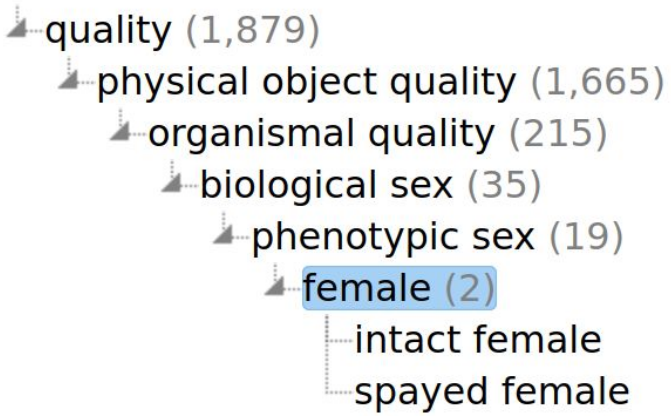
Ontologies

While **controlled vocabularies** offer more control over free text boxes, this solution is not ideal as different data sources could use different options with e.g. a different granularity.

The way to address this is to replace words in this list with **persistent identifiers** provided by an external sources. This would enable cross-linking between data resources i.e. **interoperability**

Ontologies

Granularity is very important to find the right terminology. Using a well-defined hierarchy usually helps finding the right information.



Phenotype And Trait Ontology

An ontology of phenotypic qualities (properties, attributes or characteristics)

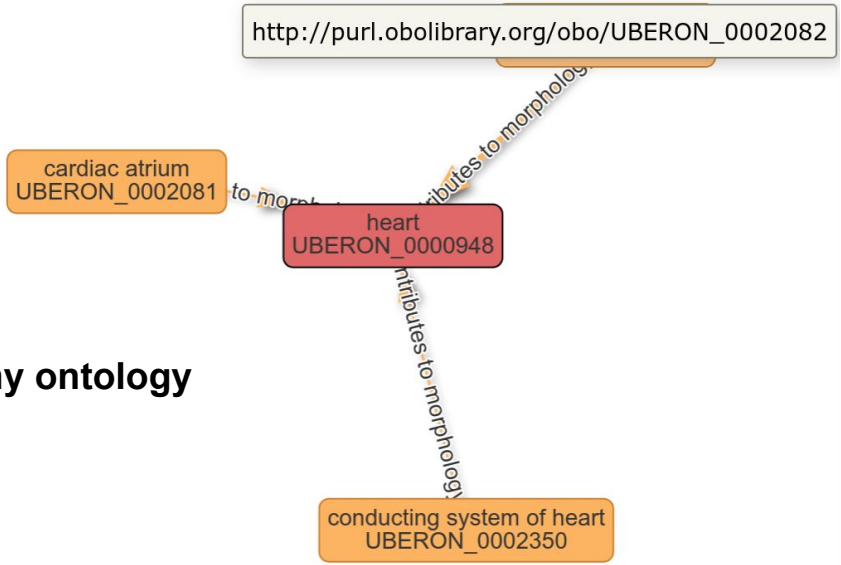
PID

http://purl.obolibrary.org/obo/PATO_0000383

https://www.ebi.ac.uk/ols4/ontologies/pato/classes/http%253A%252F%252Fpurl.obolibrary.org%252Fobo%252FPATO_0000383

Ontologies

On top of the **hierarchy**, formal relations between terms are also relevant to interlink data
These can be visualised as **graphs**



Uber-anatomy ontology

surrounds
in right side of
in left side of
contributes to morphology of
has part
continuous with
channels_from
results in formation of
results in growth of
has soma location
develops from

https://www.ebi.ac.uk/ols4/ontologies/pato/classes/http%253A%252F%252Fpurl.obolibrary.org%252Fobo%252FUBERON_0000948

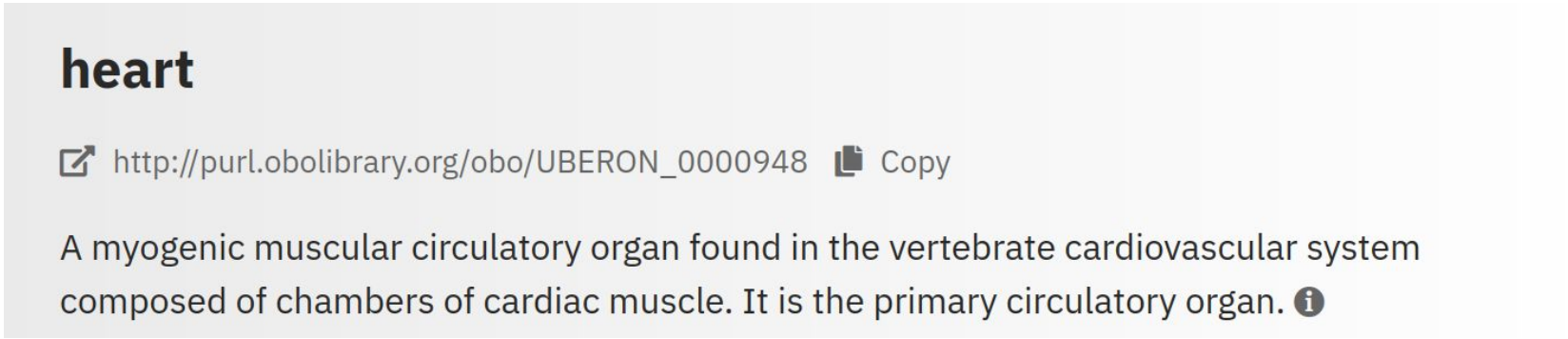
Ontologies

Cross-linking of terms across ontologies allows for an improved standardisation of terms and definitions.

https://www.ebi.ac.uk/ols4/ontologies/uberon/classes/http%253A%252F%252Fpurl.obolibrary.org%252Fobo%252FUBERON_0000948

Ontologies

Definition and synonyms to facilitate search functionalities.



Ontologies

To summarise, an ontology is a set of concept and categories in a subject area that:

- Shows and define properties
- Shows the relation between these properties
- Provides a persistent identifier to these properties
- Follows a specific tree-like hierarchy
- Is cross-linked with other ontologies

Ontologies - where to find them

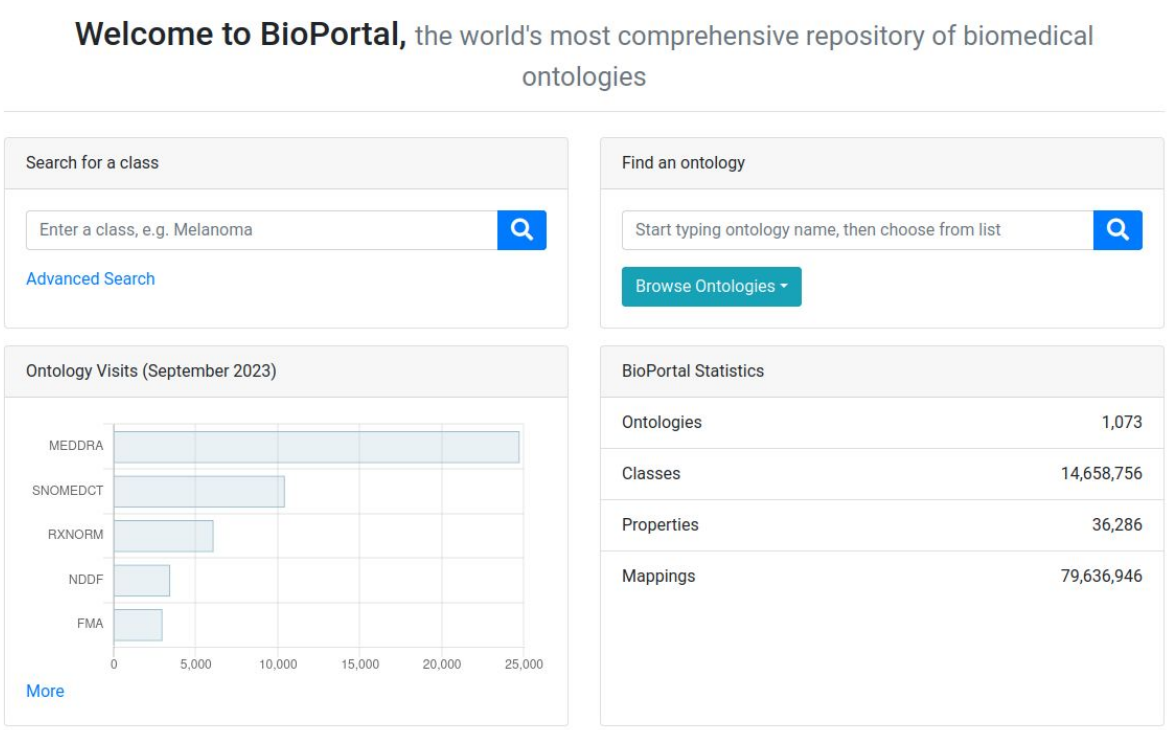
Data Content

Updated 19 Oct 2023 Thu 13:39(+02:00)

- 247 ontologies
- 7,943,142 classes
- 43,673 properties
- 684,962 individuals

https://www.ebi.ac.uk/ols4/

Ontologies - where to find them 2



https://bioportal.bioontology.org/

ERC000014	GSC MIxS human associated

experimental factor ⓘ free text

Experimental factors are essentially the variable aspects of an experiment design which can be used to describe an experiment, or set of experiments, in an increasingly detailed manner. This field accepts ontology terms from Experimenta Factor Ontology (EFO) and/or Ontology for Biomedical Investigations (OBI). For a browser of EFO (v 2.95) terms, please see http://purl.bioontology.org/ontology/EFO; for a browser of OBI (v 2018-02-12) terms please see http://purl.bioontology.org /ontology/OBI. E.g. time series design [EFO:EFO_0001779]

On ENA checklists, many “free text” boxes would actually require ontology terms.

How do I make sure at an early stage that I am collecting the right metadata?

Metadata tracking platforms

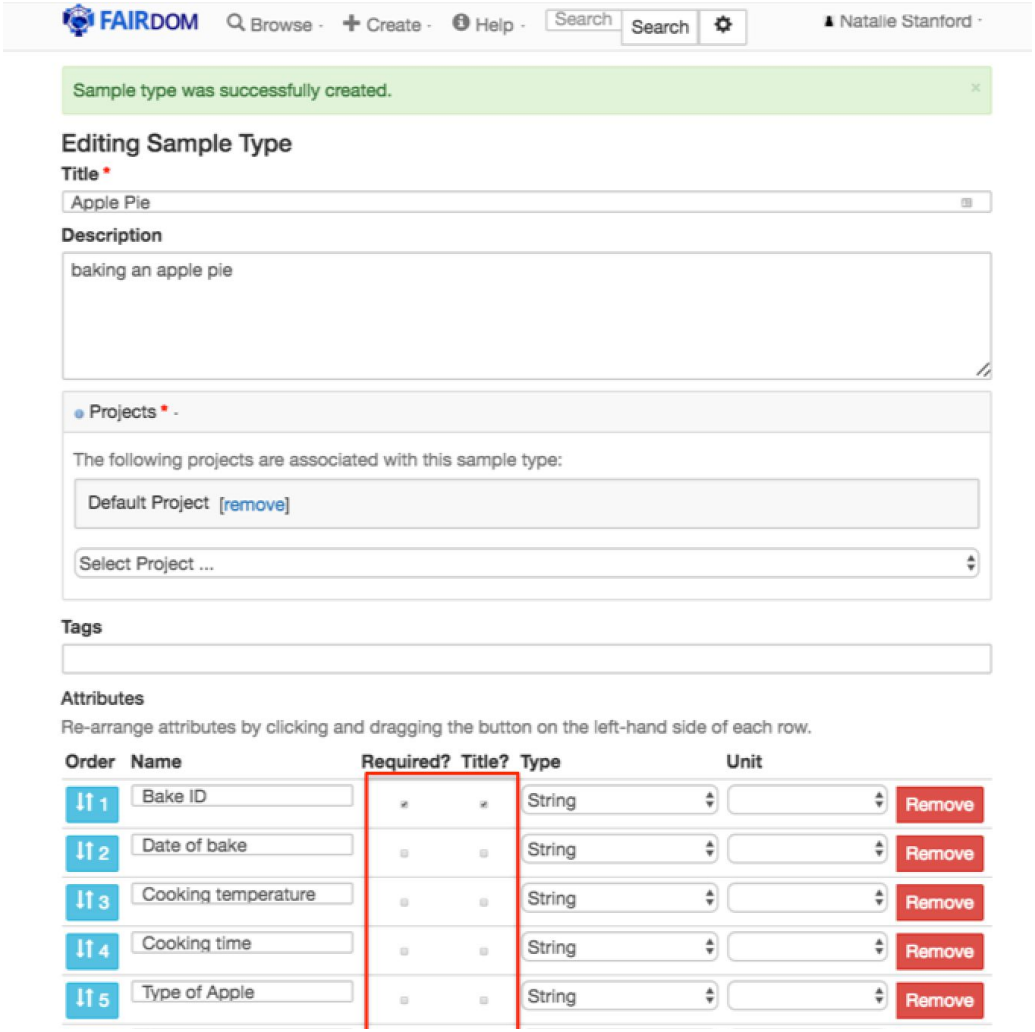
Domain specific:

- COPO for plant sciences
- MOLGENIS for biobanking
- Omero for imaging data

Customisable (domain expertise required)

- Proprietary ELNs/LIMS -
  - often poor support for ontologies
- openBIS - open source ELN/LIMS
- FAIRDOM SEEK

https://copo-project.org/
https://www.molgenis.org/
https://openbis.ch/
https://seek4science.org/



FAIRDOM SEEK

The SEEK platform is a web-based resource for sharing heterogeneous scientific research datasets, models or simulations, processes and research outcomes. It preserves associations between them, along with information about the people and organisations involved.

- National users (via Digital Life):

1000100 1010011
DIGITAL SALMON

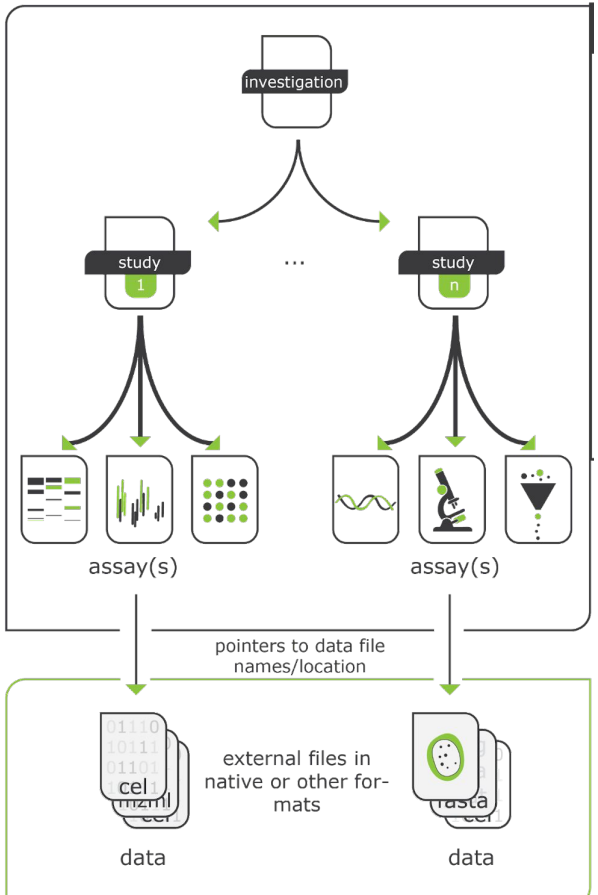
dCod 1.0
decoding the systems toxicology of atlantic cod

- Target new users (via ELIXIR)

nor seq

NAPI

The ISA model



Investigation

- Persons
- Organizations
- Publications

Study(s)

- Design
- Factor
- Protocol

Assay(s)

- Measurement
- Technology
- Materials
- Data

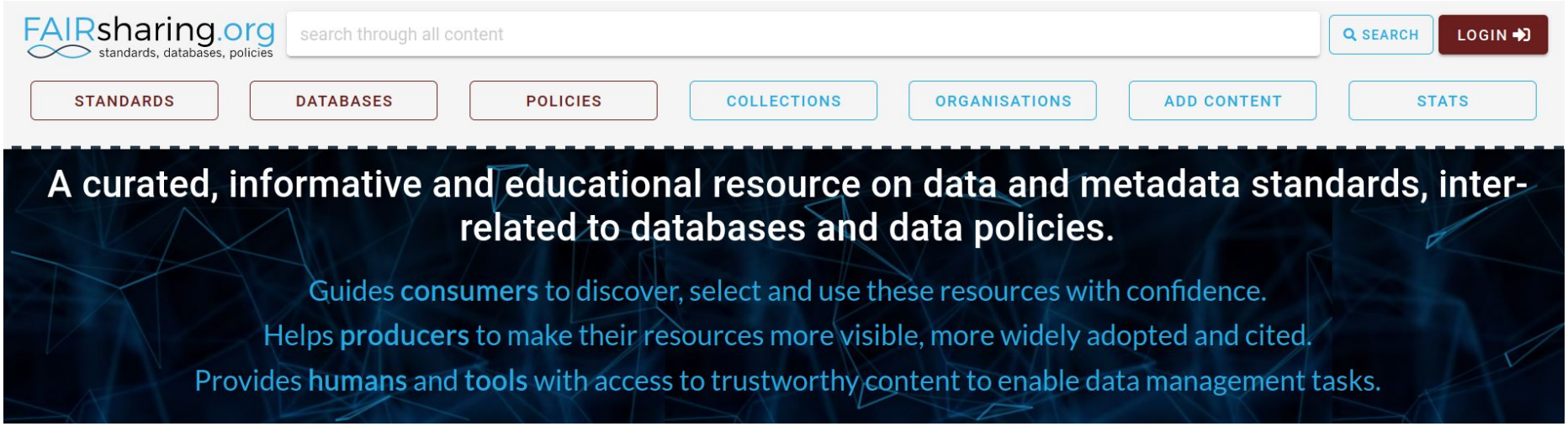
Domain-agnostic standard with typical metadata fields for a project in the life sciences

Possibility of semantic annotations

elixir Recommended Interoperability Resources

Standards and repositories - where to find them

FAIRsharing.org
standards, databases, policies



Live demonstration of search capabilities

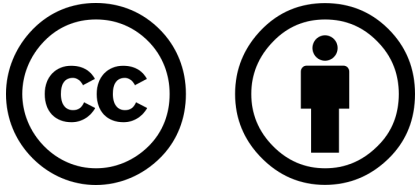
https://fairsharing.org/

Thank you!

elixir-norway.org

@elixirnorway

support@elixir.no



Except where otherwise noted, this work is licensed under a

Creative Commons Attribution 4.0 International License

https://creativecommons.org/licenses/by/4.0/