



Metadata standards, controlled vocabularies, and ontologies

Federico Bianchini

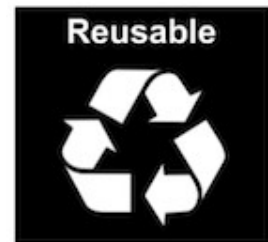
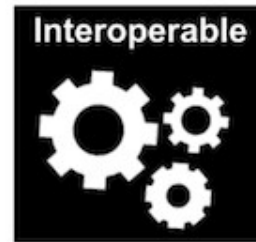
Department of Biosciences, University of Oslo

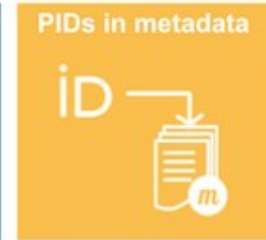
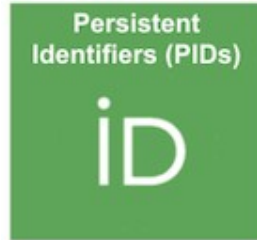


Life Science Data Management: Planning workshop 18th March 2025

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?

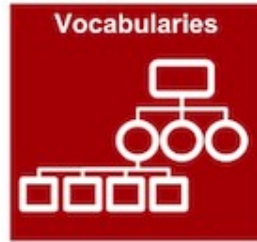




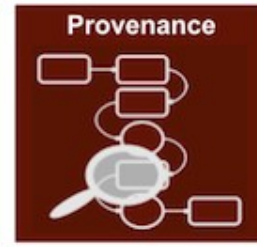
- The metadata requires a persistent identifier
- Metadata standard compliant with repositories checklist
- Metadata fields used to identify/retrieve the data



- 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



MINSEQE
MIAME

...



HUPO-PSI
TraML
MIAPE

...



SRA-XML



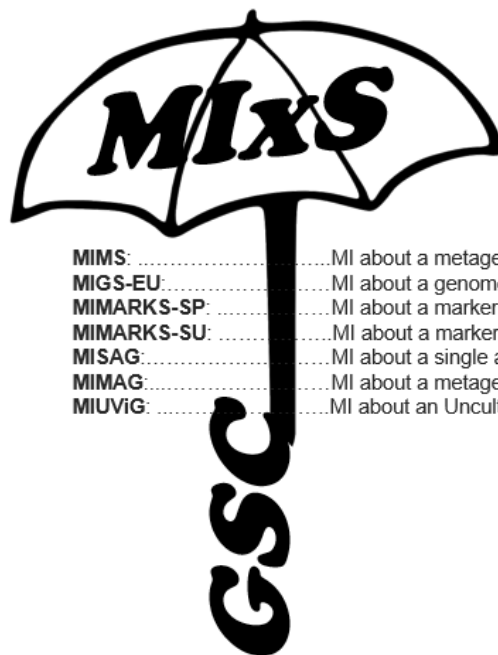
Sample Checklists

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.



Sample Checklists

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



MIMS:MI about a metagenome sequence.
MIGS-EU:MI about a genome sequence
MIMARKS-SP:MI about a marker gene sequence obtained directly from the environment
MIMARKS-SU:MI about a marker gene sequence from cultured or voucher-identifiable specimens
MISAG:MI about a single amplified genome sequence.
MIMAG:MI about a metagenome-assembled genome sequence.
MIUViG:MI about an Uncultivated Virus Genome

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

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

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



ERC000014 GSC MixS human associated

number of replicons



restricted text

regular expression 

[+]?[0-9]+

smoker



text choice

ex-smoker

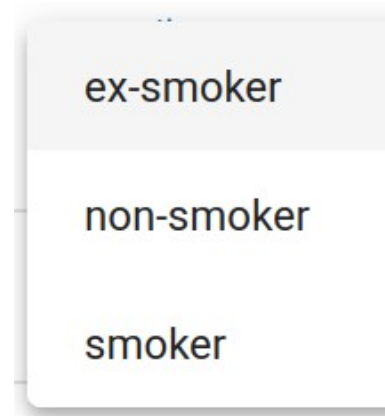
non-smoker

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

How many ways can you say “female”?

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

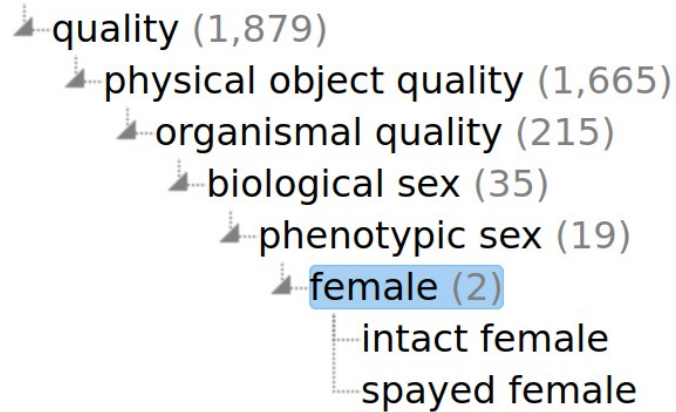
While **controlled vocabularies** offer more control over free text boxes, this solution is not ideal as different data sources could use different options with e.g. a different granularity.

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

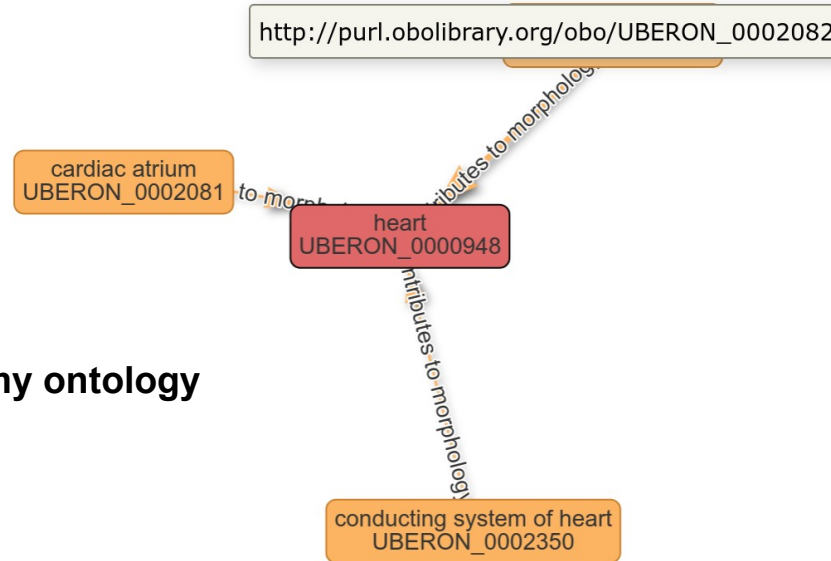
https://www.ebi.ac.uk/ols4/ontologies/pato/classes/http%253A%252F%252Fpurl.obolibrary.org%252Fobo%252FPATO_0000383





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		<input type="checkbox"/>
in right side of		<input type="checkbox"/>
in left side of		<input type="checkbox"/>
contributes to		<input checked="" type="checkbox"/>
morphology of		
has part		<input type="checkbox"/>
continuous		<input type="checkbox"/>
with		
channels_from		<input type="checkbox"/>
results in		<input type="checkbox"/>
formation of		
results in		<input type="checkbox"/>
growth of		
has soma		<input type="checkbox"/>
location		
develops from		<input type="checkbox"/>

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.

heart

 http://purl.obolibrary.org/obo/UBERON_0000948  Copy

Also appears in

OVAE

OPMI

CPONT

ZP

AISM

EFO

CLO

OBIB

MONDO

TXPO

OHPI

XPO

FOVT

PCL

ENVO

WBPHENOTYPE

OBI

RBO

NBO

HP

GAZ

OBA

ONE

ECTO

EUPATH

HCAO

FLOPO

GENEPIO

ICO

FOODON

BCGO

CCF

OAE

MICRO

MAXO

MP

ECOCORE

GO

MS

CL

DOID


PATO


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.


heart


http://purl.obolibrary.org/obo/UBERON_0000948  Copy

A myogenic muscular circulatory organ found in the vertebrate cardiovascular system composed of chambers of cardiac muscle. It is the primary circulatory organ. 

Synonym

branchial heart 

cardium 

chambered heart 

vertebrate heart

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

[Home](#)[Ontologies](#)[Help](#)[About](#)[Downloads](#)

Welcome to the EMBL-EBI Ontology Lookup Service

[Search](#)

☐ Exact match ☐ Include obsolete terms ☒ Include imported terms

Examples: [diabetes](#), [GO:0098743](#)

[Looking for a particular ontology?](#)

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

Ontologies - where to find them 2



Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies

Search for a class

Enter a class, e.g. Melanoma



[Advanced search](#)

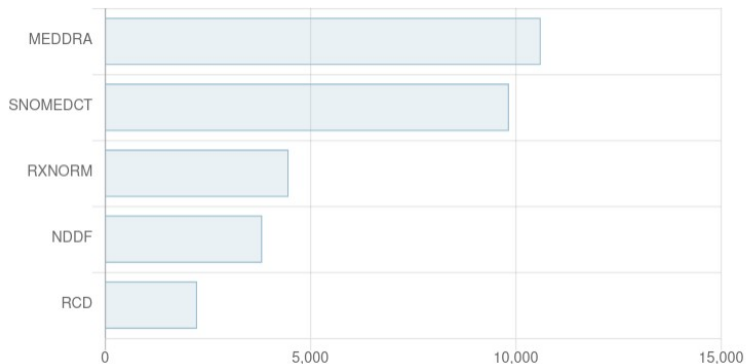
Find an ontology

Start typing ontology name, then choose from list



[Browse ontologies](#)

Ontology visits (February 2025)



Statistics

Ontologies

1,185

Classes

15,335,435

Properties

36,286

Mappings

102,816,942

Ontologies - where to find them 3

STANDARDS

DATABASES

POLICIES

COLLECTIONS

ORGANISATIONS

ADD CONTENT

STATS

A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies.



1861 Standards

<u>Terminology Artifact</u>	812
Model/Format	641
Reporting Guideline	317
Identifier Schema	68

[VIEW ALL](#)



2322 Databases

Repositories	1222
Knowledgebases	927
Knowledgebase/Repositories	173
Institutional Repositories	96

[VIEW ALL](#)



344 Policies

Journal	193
Institution	53
Funder	40
Journal publisher	35
Society	17
Project	6

[VIEW ALL](#)

Not for browsing ontologies, but to discover relations between ontologies, standards and repositories

https://fairsharing.org/search?recordType=terminology_artefact



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 Experimental 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/>

Sample type was successfully created. ×

Editing Sample Type

Title *

Apple Pie ⊞

Description

baking an apple pie

Projects ·

The following projects are associated with this sample type:

Default Project [\[remove\]](#)

Select Project ... ⌵

Tags

Attributes

Re-arrange attributes by clicking and dragging the button on the left-hand side of each row.

Order	Name	Required?	Title?	Type	Unit	
1	Bake ID	*	*	String		Remove
2	Date of bake	□	□	String		Remove
3	Cooking temperature	□	□	String		Remove
4	Cooking time	□	□	String		Remove
5	Type of Apple	□	□	String		Remove

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.

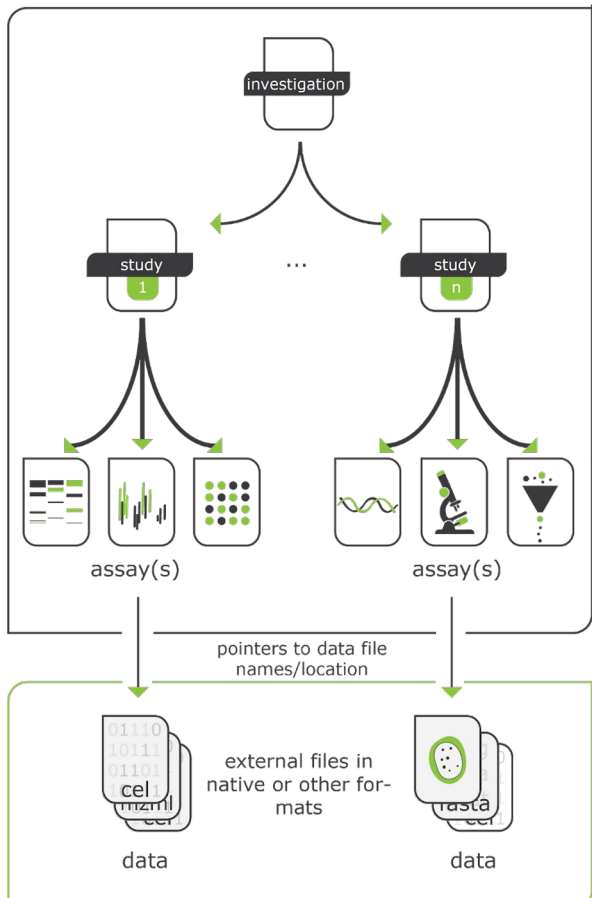
- National users (via Digital Life):



- Target new users (via ELIXIR)



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



Standards and repositories - where to find them



search through all content

Q SEARCH

LOGIN ➔

STANDARDS

DATABASES

POLICIES

COLLECTIONS

ORGANISATIONS

ADD CONTENT

STATS

A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies.

Guides consumers to discover, select and use these resources with confidence.

Helps producers to make their resources more visible, more widely adopted and cited.

Provides humans and tools with access to trustworthy content to enable data management tasks.

**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/)

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