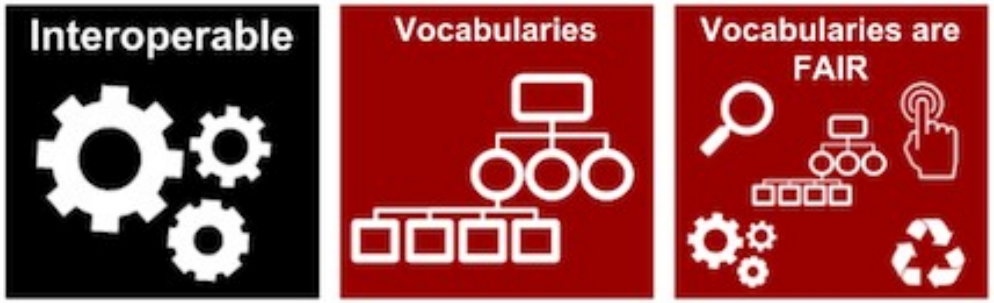
Metadata Standards & Ontologies

Federico Bianchini
Data Steward
University of Oslo/ELIXIR Norway

Controlled vocabularies

Vocabulary a collection of preferred terms used to annotate and retrieve content. Predefined terms are mandated to make each entry unambiguous and consistent.



The controlled vocabulary used to describe datasets needs to be <u>resolvable</u> using globally unique and persistent identifiers.

In practice, why do we need controlled vocabularies?

How many ways can you say “female”?

How many ways can you say “female”?

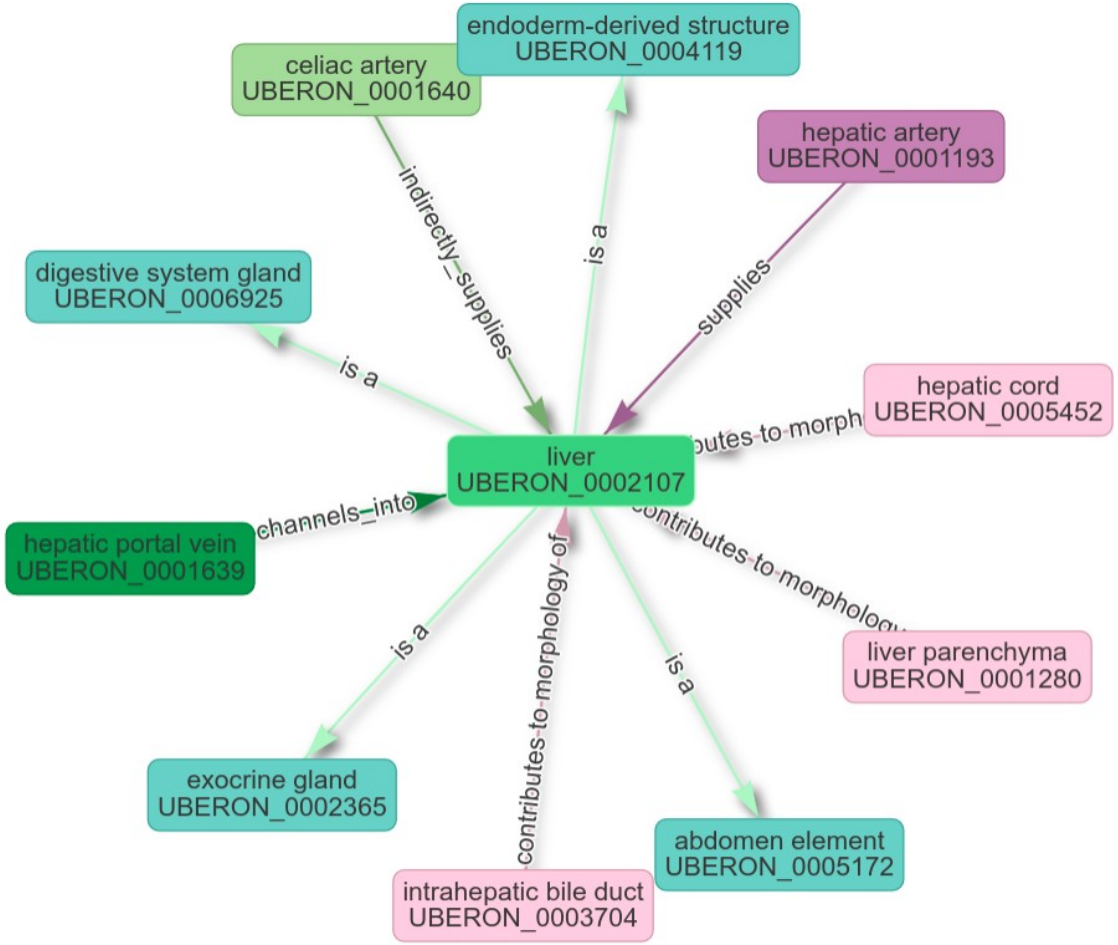
18-day pregnant females
2 yr old female
400 yr. old female
adult female
asexual female
castrate female
cf.female
cystocarpic female
dikaryon
dioecious female
diploid female
f
famale
femail
female
female - worker
female (alate sexual)
female (calf)
hen
female (lactating)
female (pregnant)
female (outbred)
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
probably female (based on morphology)
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)

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

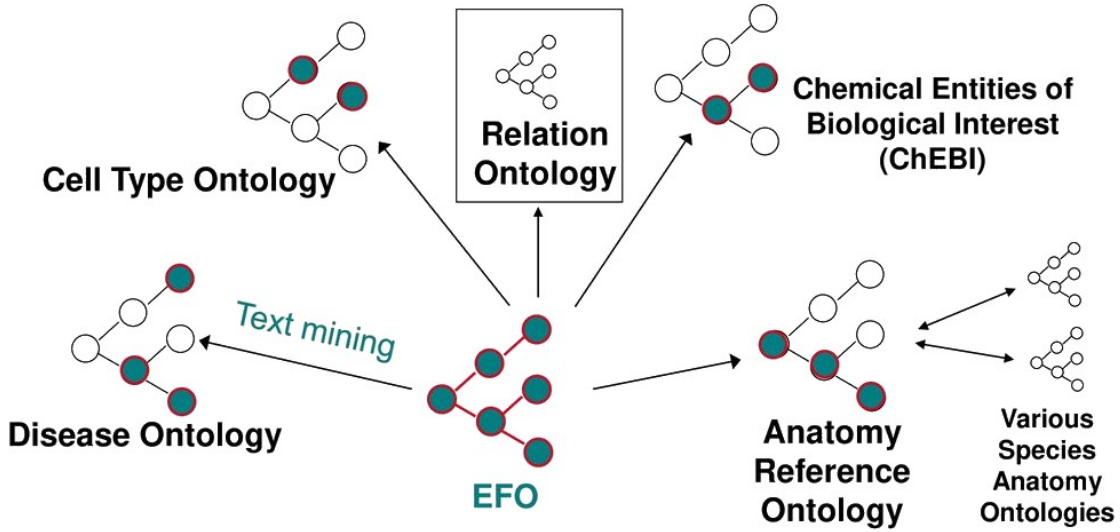
Ontologies

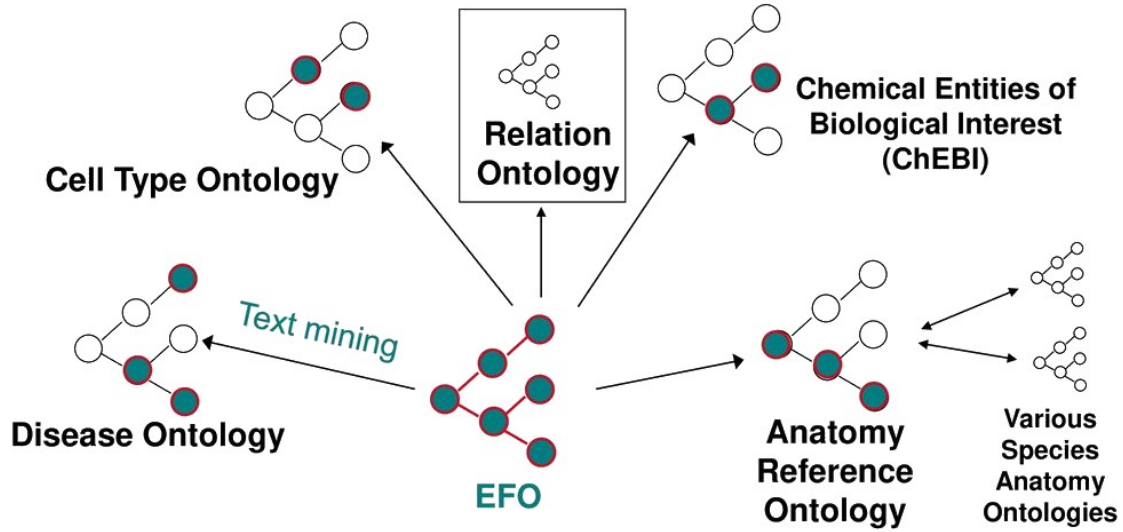
Ontology a set of concepts and categories in a subject area or domain that shows their properties and the relations between them.



Relationship	Color	Visibility
Extended nodes (*)		-
is a		☑
part of		☐
develops from		☐
contributes to morphology of		☑
drains		☐
supplies		☑

ifo


Cell Type Ontology
Relation Ontology
Chemical Entities of Biological Interest (ChEBI)
Text mining
Disease Ontology
EFO
Anatomy Reference Ontology
Various Species Anatomy Ontologies



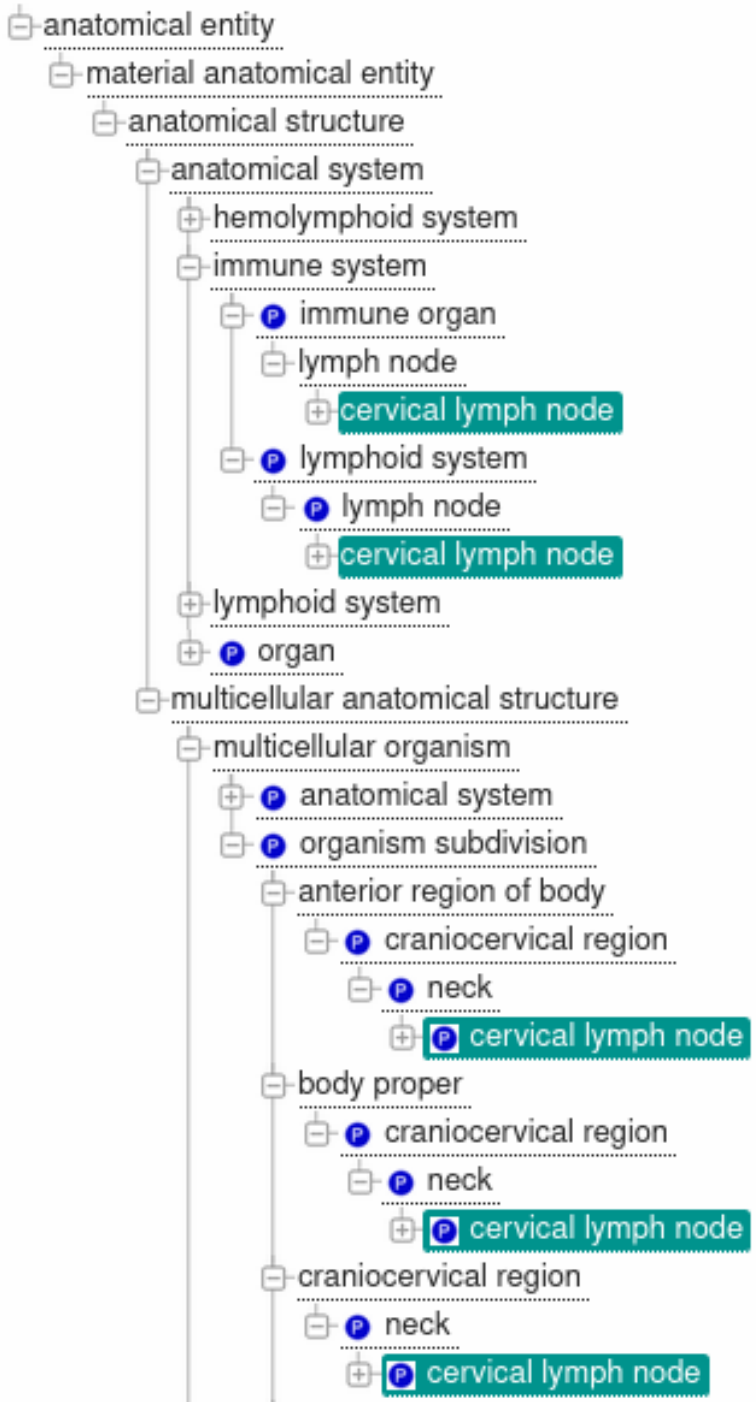
- experimental factor
  - material property
    - quality
      - sex
        - female

database cross reference

- MSH:D005260
- MO:506
- NCIt:C16576
- SNOMEDCT:248152002
- CARO:0000028
- PATO:0000383

Uberon

Onthologies enable hierarchical searches



Controlled vocabulary & Ontologies

Metadata standards – controlled vocabulary for

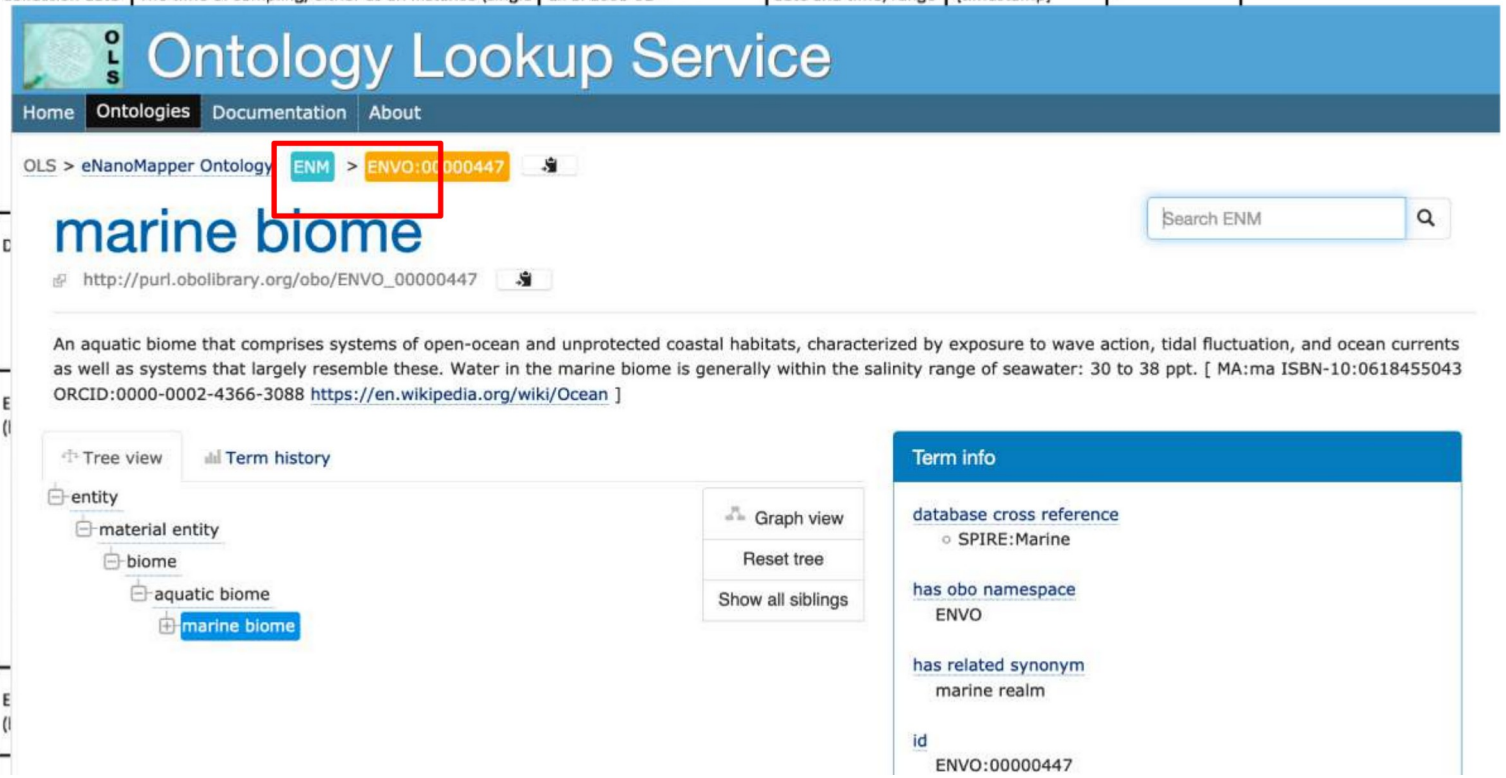
Structured comment name	Item	Description	Examples	Expected value	Value syntax	Preferred units / suffix
alt_elev	Geographic location (altitude/elevation)	Sample taken at given elevation above sea level, defined in meters(m) as a positive floating number with two decimals.	Ex 1: 3.06 Ex 2: 1.80-2.15	-	{float} or {range}	meters (m)
collection_date	Collection date	The time of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated.	Ex 1: 2008-01-23T19:23:10+00:00 Ex 2: 2011-11-10 Ex 3: 2001-12 Ex 7: 2015 Ex 4: 2003--2006 Ex 5: 2010-01--2011-03 Ex 6: 2011-05-28--2011-08-10	date and time, range	{timestamp}	-
depth	Depth	Please refer to the definitions of depth in the environmental packages. Water: Sample taken at given depth below sea level, defined in meters(m) as a positive floating number or as a range, both with two decimals.	Ex 1: 355.20 Ex 2: 2.00-5.00	-		meters (m)
env_biome	Environment (biome)	In environmental biome level are the major classes of ecologically similar communities of plants, animals, and other organisms. Biomes are defined based on factors such as plant structures, leaf types, plant spacing, and other factors like climate. Examples include: desert, taiga, deciduous woodland, or coral reef. EnvO (v1.53) terms listed under environmental biome can be found from the link:(http://www.environmentontology.org/Browse-EnvO)	Ex 1: coral reef Ex 2: tropical	EnvO	{free text}	-
env_biome_ENVO	Environment (biome_id)	Corresponding ENVO identifier related to the term name of Environment (biome).	Ex 1: ENVO:00000150 Ex 2: ENVO:01000204	EnvO	{accession}	-

Not collected	->	missing
250 M	->	250
Not applicable	->	NA
Superficial	->	missing
-1 m	->	1
-2 m	->	2
-2901.0	->	2901
0 m.	->	0
1912 ft	->	582.80
40 mm from surface	->	0.04
0.75 m above seafloor	->	missing
700meters	->	700
Intracellular	->	missing
Surface water of 0 meter	->	0
Zero	->	0
Below surface	->	Missing

Controlled vocabulary & Ontologi

Ontology Lookup Service (OLS) is a resource for biomedical ontologies

Structured comment name	Item	Description	Examples	Expected value	Value syntax	Preferred units / suffix
alt_elev	Geographic location (altitude/elevation)	Sample taken at given elevation above sea level, defined in meters(m) as a positive floating number with two decimals.	Ex 1: 3.06 Ex 2: 1.80-2.15	-	{float} or {range}	meters (m)
collection_date	Collection date	The time of sampling, either as an instance (single	Ex 1: 2008-01-	date and time, range	{timestamp}	-
depth	D					
env_biome	E (					
env_biome_ENVO	E (					

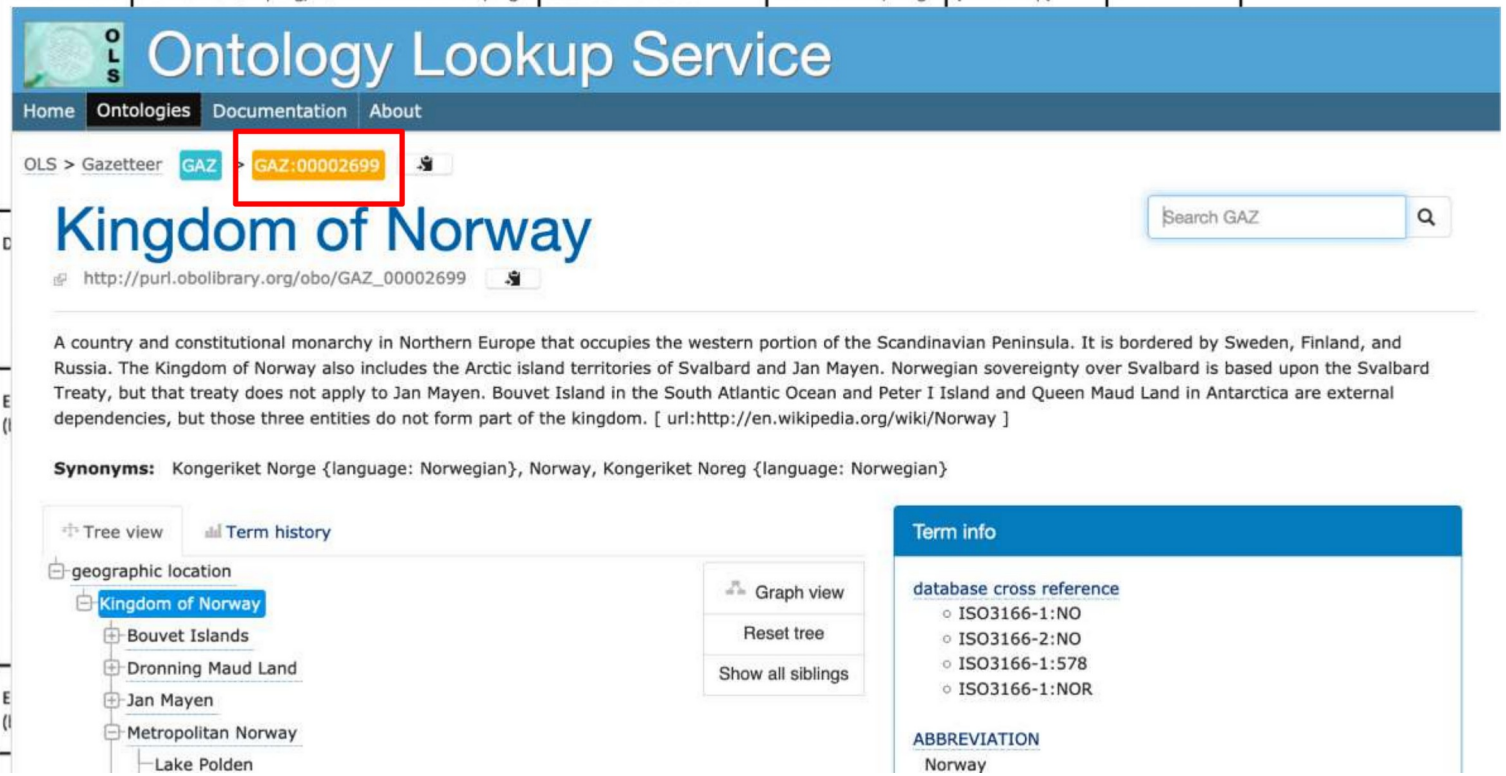


The ENVO ontology describes the environment of the sampling

Controlled vocabulary & Ontologi

Ontology Lookup Service (OLS) is a resource for biomedical ontologies

Structured comment name	Item	Description	Examples	Expected value	Value syntax	Preferred units / suffix
alt_elev	Geographic location (altitude/elevation)	Sample taken at given elevation above sea level, defined in meters(m) as a positive floating number with two decimals.	Ex 1: 3.06 Ex 2: 1.80-2.15	-	{float} or {range}	meters (m)
collection_date	Collection date	The time of sampling, either as an instance (single	Ex 1: 2008-01-	date and time, range	{timestamp}	-
depth	D					
env_biome	E (					
env_biome_ENVO	E (					



The GAZ ontology describes the geographical location of the sampling

ONTOLOGY SEARCH

Home | Ontologies | Documentation | About

Welcome to the EMBL-EBI Ontology Lookup Service

Search OLS... Search

Examples: diabetes, GO:0098743

Looking for a particular ontology?

About OLS

The Ontology Lookup Service (OLS) is a repository for biomedical ontologies that aims to provide a single point of access to the latest ontology versions. You can browse the ontologies through the website as well as programmatically via the OLS API. OLS is developed and maintained by the Samples, Phenotypes and Ontologies Team (SPOT) at EMBL-EBI.

Related Tools

In addition to OLS the SPOT team also provides the OxO, Zooma and Webulous services. OxO provides cross-ontology mappings between terms from different ontologies. Zooma is a service to assist in mapping data to ontologies in OLS and Webulous is a tool for building ontologies from spreadsheets.

Report an Issue

For feedback, enquiries or suggestion about OLS or to request a new ontology please use our GitHub issue tracker. For announcements relating to OLS, such as new releases and new features sign up to the OLS announce mailing list

Data Content

Updated 28 May 2021 08:03

- 264 ontologies
- 6,460,093 terms
- 32,279 properties
- 497,528 individuals

Tweets by @EBIOLS

EBISPOT OLS
@EBIOLS

Are you interested in deploying OLS, Zooma and OxO in your own environment? If so, please checkout our documentation in this regard github.com/EBISPOT/ontoto….
Many thanks to @_jmcl and @NicoMatentzoglu for their work on this. Great job!

EBISPOT/
ontotools-docker

Configuration to deploy ontotools using docker compose

3 Contributors 2 Issues 1 Stars 5 Forks

https://www.ebi.ac.uk/ols/index

The Ontology Lookup Service is part of the ELIXIR infrastructure

OLS is an Elixir interoperability service Learn more ›

What is a metadata standard?

but often following the same concept:

Investigation

Study(s)

Assay(s)

FAIRsharing.org
standards, databases, policies

Technology & domain specific

but often following the same concept:

Investigation

Persons
Organizations
Publications

Study(s)

Assay(s)

FAIRsharing.org
standards, databases, policies

Technology & domain specific

but often following the same concept:

Investigation

Persons
Organizations
Publications

Study(s)

Design
Factor
Protocol

Assay(s)

FAIRsharing.org
standards, databases, policies

Technology & domain specific

but often following the same concept:

Investigation

Persons
Organizations
Publications

Study(s)

Design
Factor
Protocol

Assay(s)

Measurement
Technology
Materials
Data

FAIRsharing.org
standards, databases, policies

Technology & domain specific

but often following the same concept:

Investigation

Persons
Organizations
Publications

Study(s)

Design
Factor
Protocol

Assay(s)

Measurement
Technology
Materials
Data

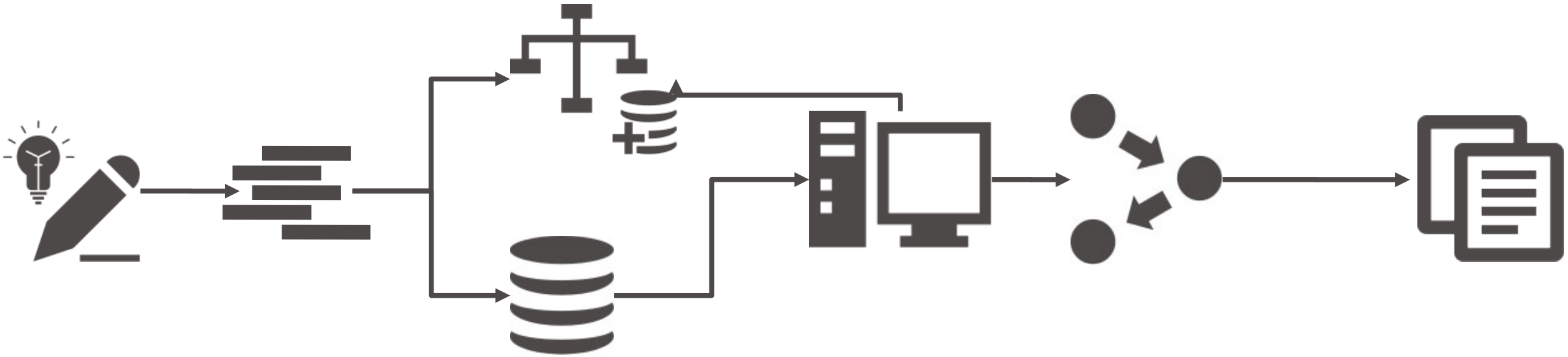
Controlled Vocabularies
Ontologies
Standards

FAIRsharing.org
standards, databases, policies

Technology & domain specific

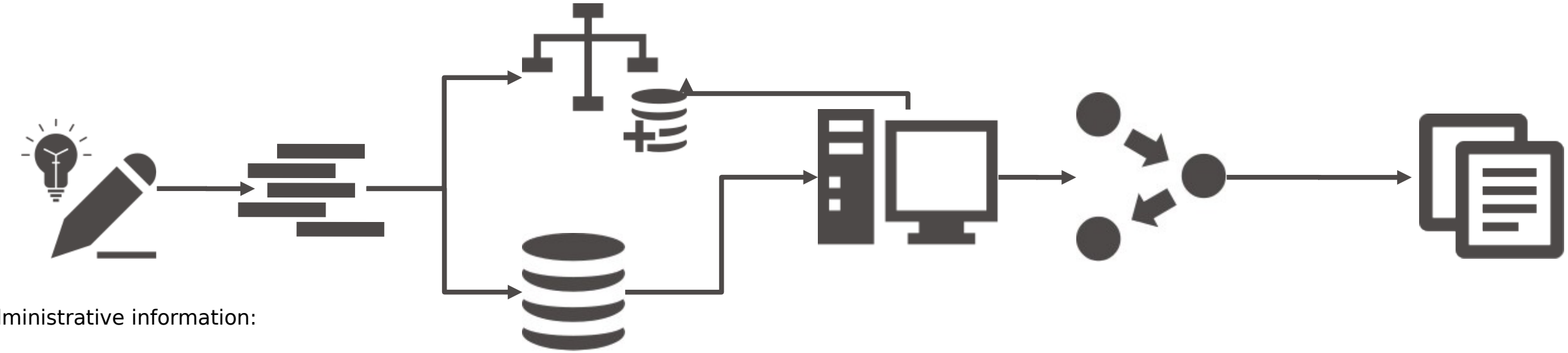
MINSEQE

ArrayExpress



MINSEQE

ArrayExpress



Administrative information:

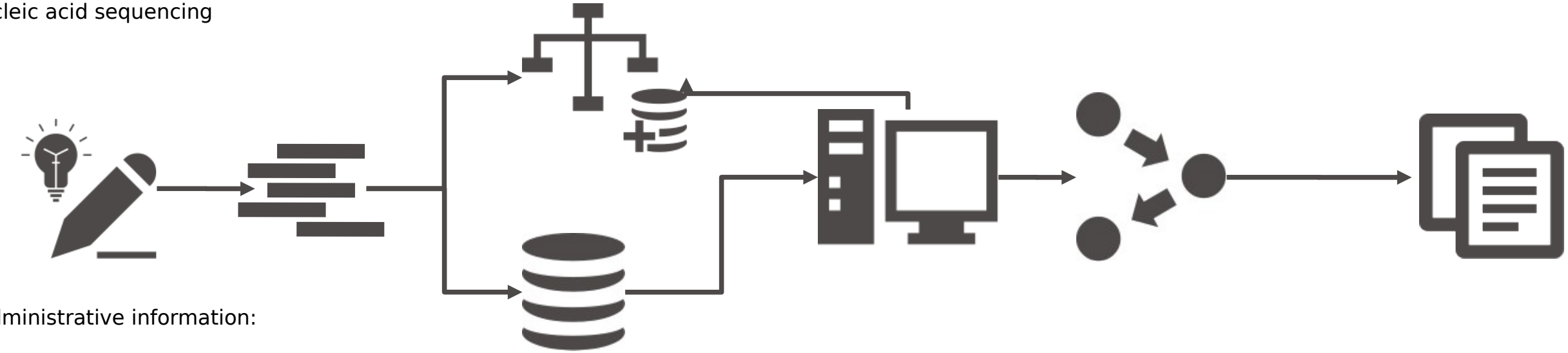
Persons
Organizations
Publications

MINSEQE

Experimental conditions/design

protocols:

treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library construction
nucleic acid sequencing



Administrative information:

Persons
Organizations
Publications

MINSEQE

ArrayExpress

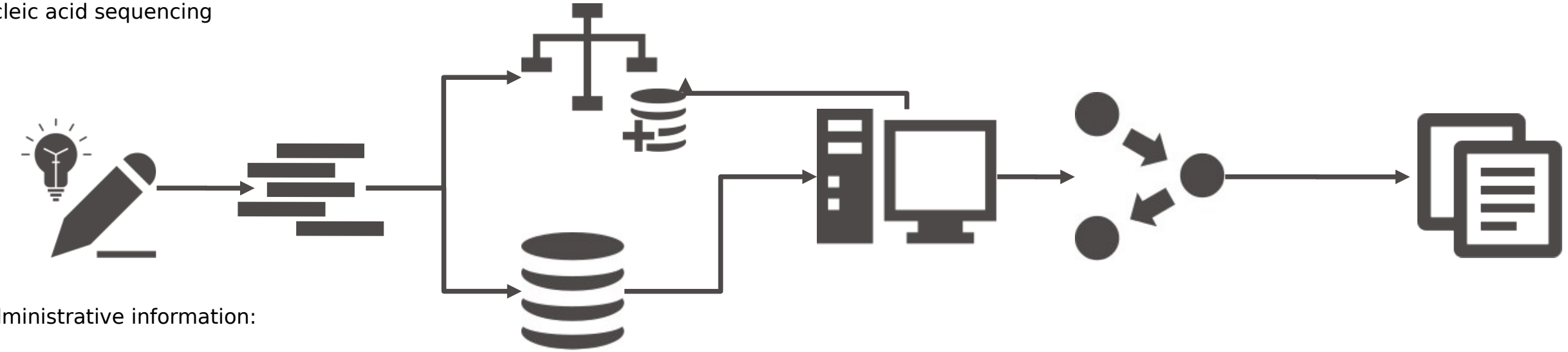
Experimental conditions/design

protocols:

treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library construction
nucleic acid sequencing

protocols:

high throughput sequence alignment
normalization data transformation



Administrative information:

Persons
Organizations
Publications

fastq
bam
csv/tsv

MINSEQE

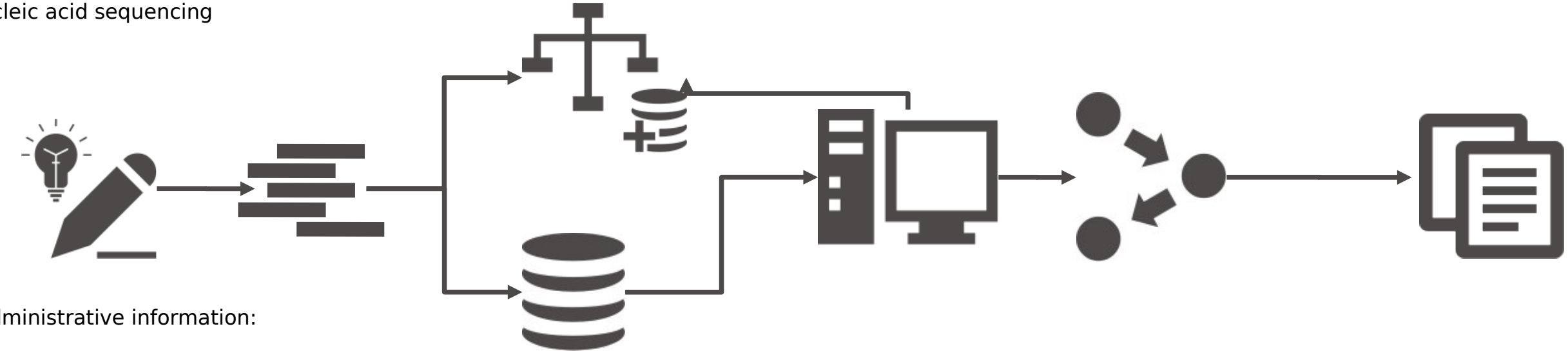
Experimental conditions/design

protocols:

treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library construction
nucleic acid sequencing

protocols:

high throughput sequence alignment
normalization data transformation



Administrative information:

Persons
Organizations
Publications

MINSEQE

ArrayExpress

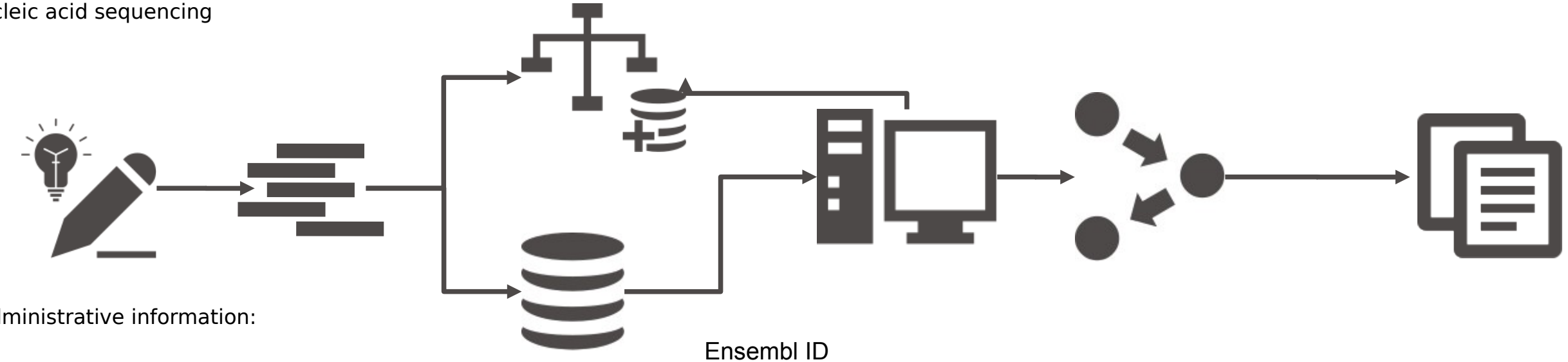
Experimental conditions/design

protocols:

treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library construction
nucleic acid sequencing

protocols:

high throughput sequence alignment
normalization data transformation



Administrative information:

Persons
Organizations
Publications

fastq
bam
csv/tsv

Ensembl ID
RefSeq ID

MINSEQE

Experimental conditions/design

protocols:

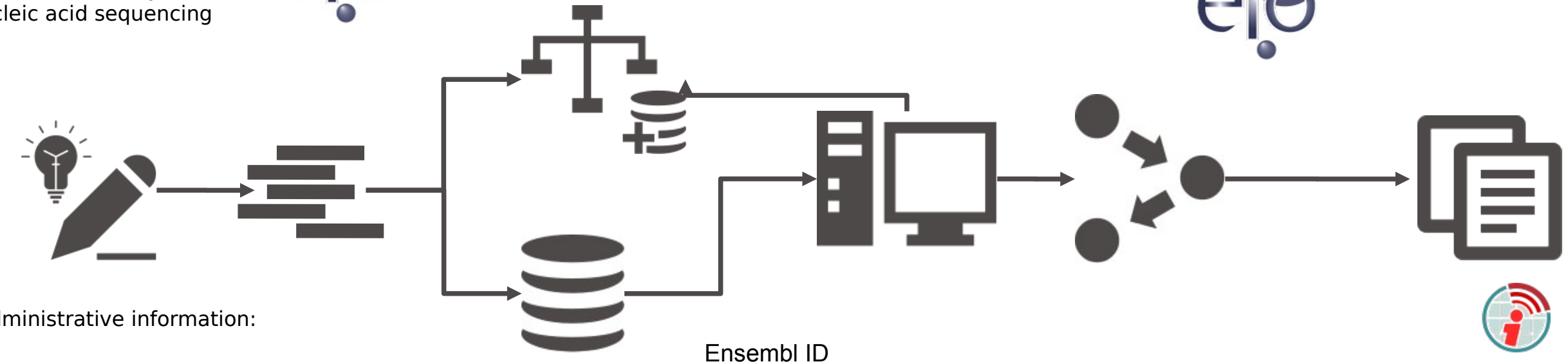
treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library constructi
nucleic acid sequencing

Taxonomy

protocols:

high throughput sequence alignment
normalization data transformation

ArrayExpress



Administrative information:

Persons
Organizations
Publications

fastq
bam
csv/tsv

Ensembl ID
RefSeq ID

MINSEQE

Experimental conditions/design

protocols:

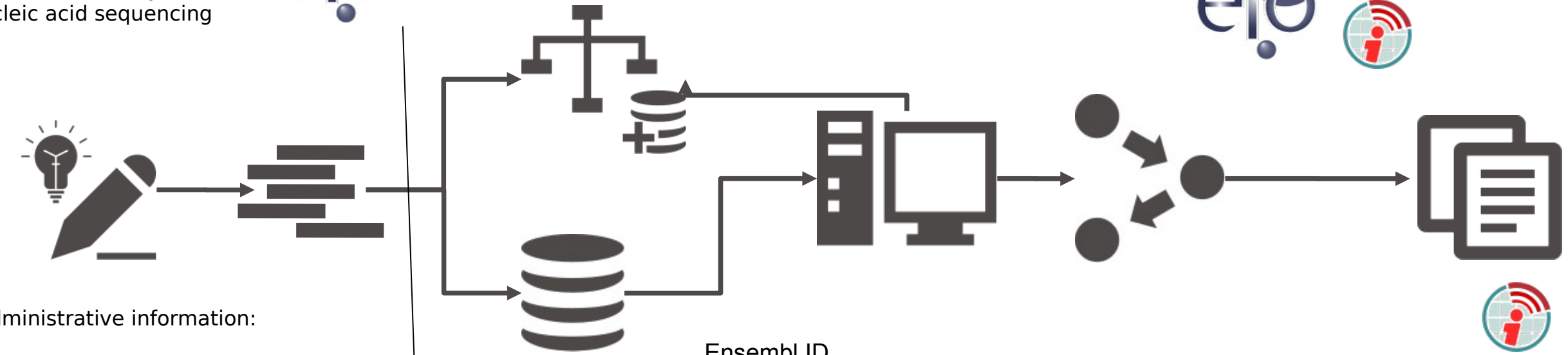
treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library constructi
nucleic acid sequencing

Taxonomy

protocols:

high throughput sequence alignment
normalization data transformation

ArrayExpress



Administrative information:

Persons
Organizations
Publications

fastq
bam
csv/tsv

Ensembl ID
RefSeq ID

Interlinking with other resources

MINSEQE

Experimental conditions/design

protocols:

treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library constructi
nucleic acid sequencing

Taxonomy

protocols:

high throughput sequence alignment
normalization data transformation

ArrayExpress

Annotare

https://www.ebi.ac.uk/fg/annotare

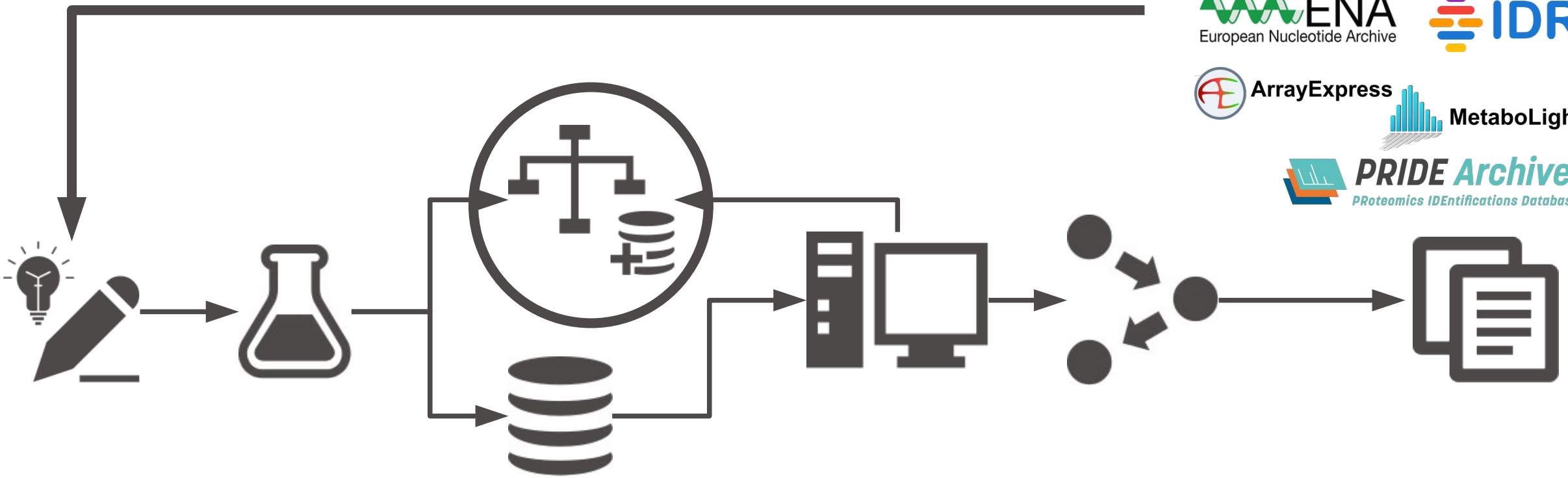
Administrative information:

Persons
Organizations
Publications

fastq
bam
csv/tsv

Ensembl ID
RefSeq ID

Interlinking with other resources



BioStudies.

ENA
European Nucleotide Archive

IDR

ArrayExpress

MetaboLights

PRIDE Archive
PRoteomics IDEntifications Database

EUROPEAN
GENOME-PHENOME
ARCHIVE

Meta data standards

ArrayExpress

MINSEQE
MIAME
...

Meta data standards

ArrayExpress

MINSEQE
MIAME
...

PRIDE Archive
PRoteomics IDEntifications Database

HUPO-PSI
TraML
MIAPE
...

Meta data standards

ArrayExpress

MINSEQE
MIAME
...

ENA
European Nucleotide Archive

SRA-XML

PRIDE Archive
PRoteomics IDEntifications Database

HUPO-PSI
TraML
MIAPE
...

Which metadata standard?

FAIRsharing

Demo

Data format standards

Common formats

Non-proprietary formats (accessible with open source tools)

Avoiding binary data formats (data corruption)

Examples: FASTQ, TIFF, mzML,...

Data format standards

ArrayExpress

FASTQ
MAGE-ML
...

PRIDE Archive
PRoteomics IDEntifications Database

mzML
mzQuantML
...

ENA
European Nucleotide Archive

FASTA
FASTQ
...

Metadata tracking platforms

Domain specific:

- **COPO for plant sciences**
- **MOLGENIS for biobanking**
- **...**

MOLGENIS

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

Metadata tracking platforms

Domain specific:

- **COPO for plant sciences**
- **MOLGENIS for biobanking**
- **...**

Adaptable (configuration requires domain knowledge):

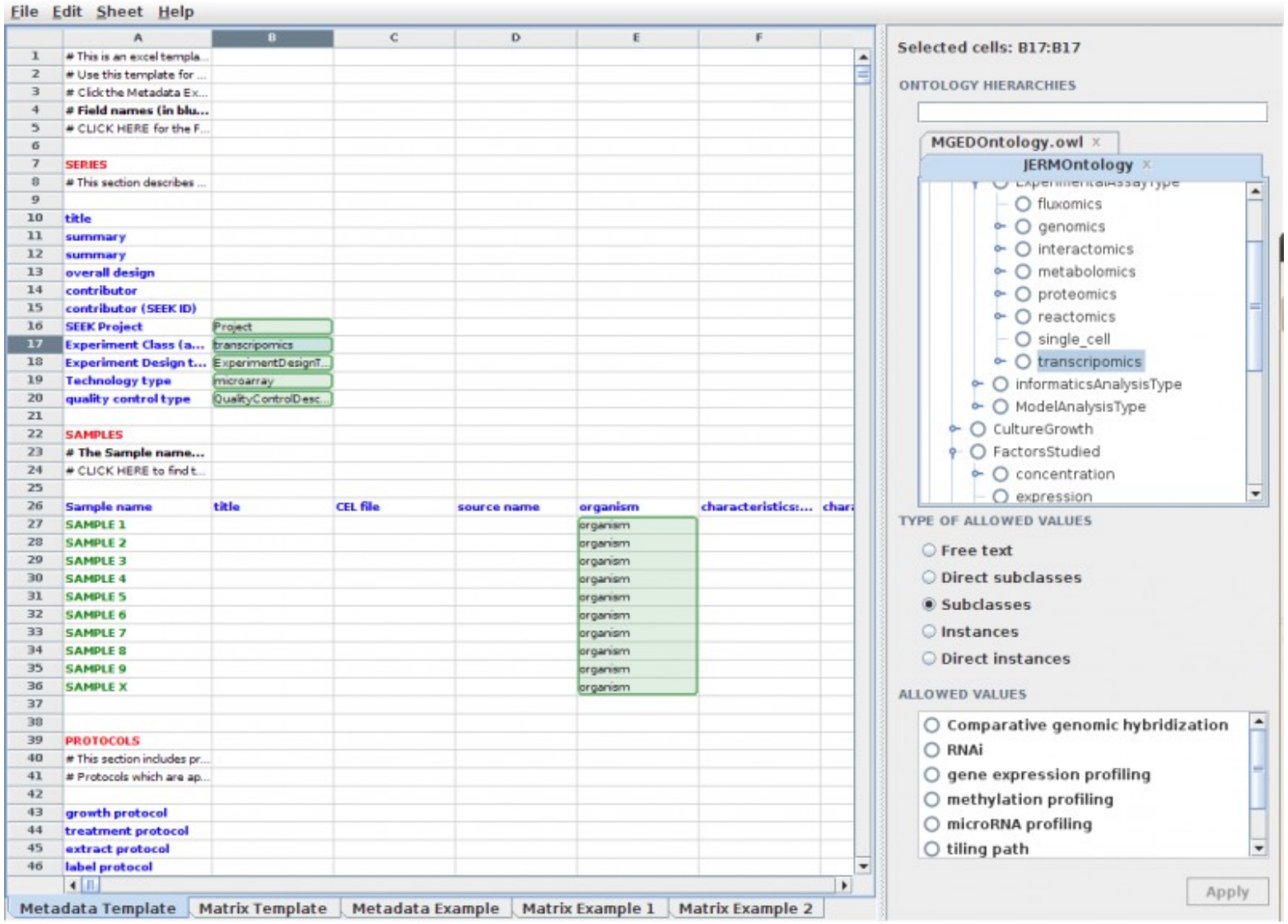
- **Proprietary ELNs/LIMS - often poor support for ontologies**
- **openBIS - open source ELN/LIMS**
- **SEEK**

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

RightField

SEEK

SEEK



https://seek4science.org/
https://rightfield.org.uk/

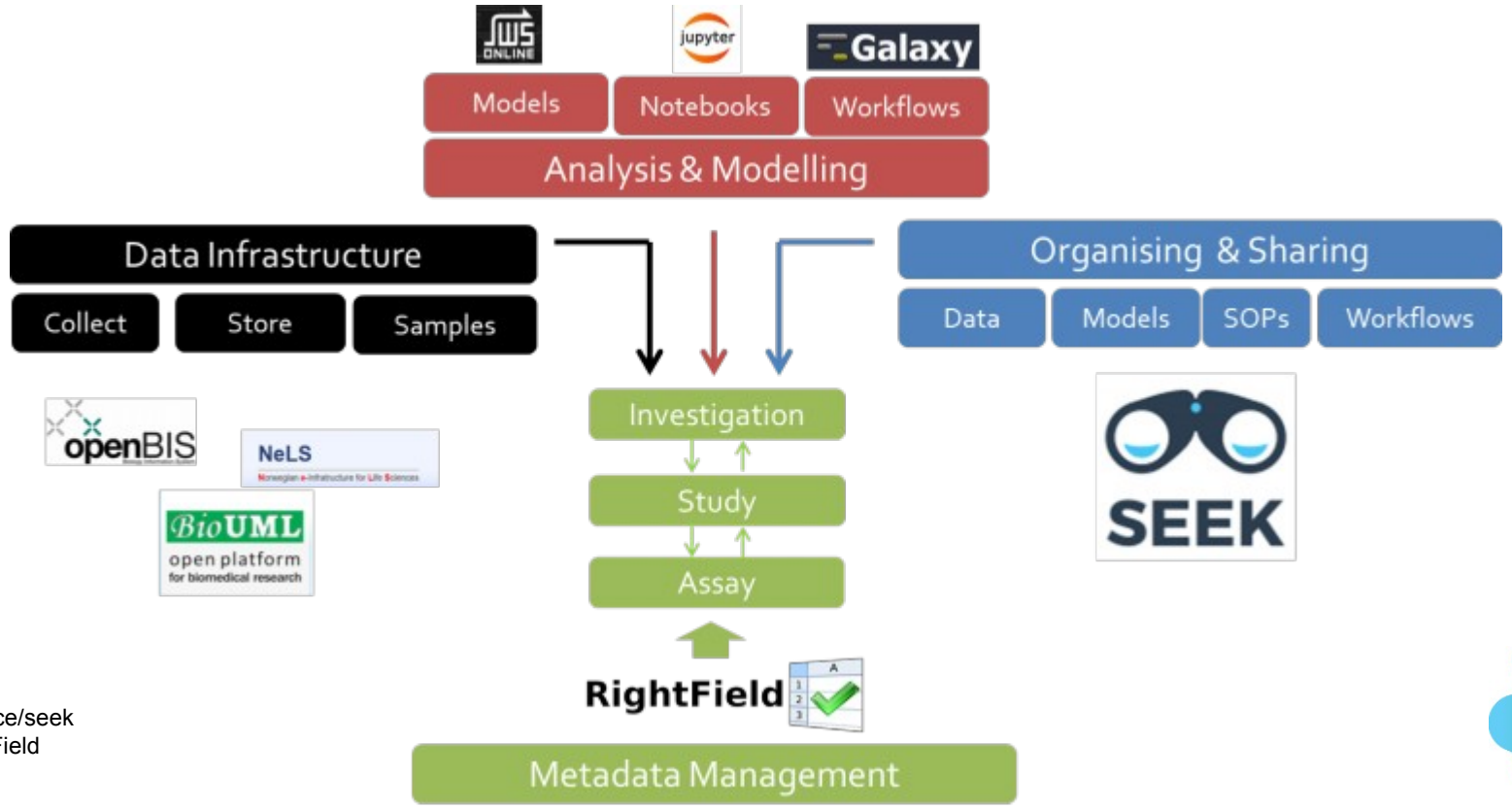
FAIRDOM

SEEK

FAIRDOM HUB

https://seek4science.org/
https://fairdomhub.org/

FAIRDOM integration



fair-dom.org
seek4science.org - github.com/seek4science/seek
rightfield.org.uk - github.com/myGrid/RightField

CENTRE FOR DIGITAL LIFE NORWAY