



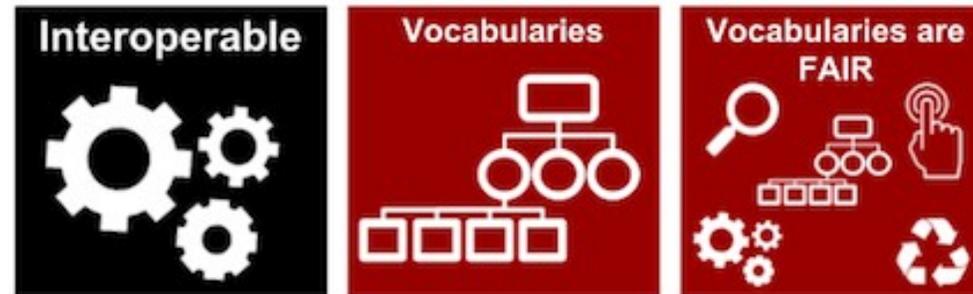
Metadata Standards & Ontologies



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 resolvable 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”?

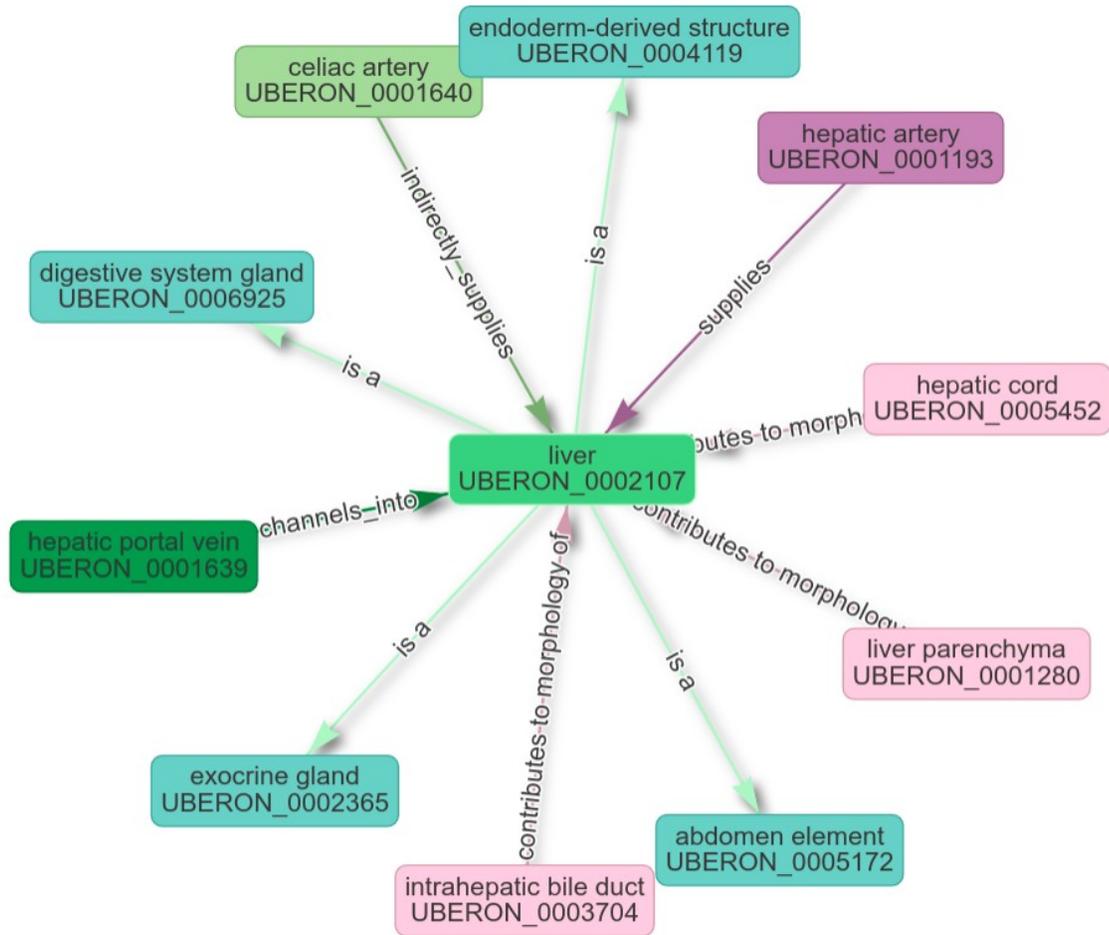
18-day pregnant females	female (lactating)	individual female	worker caste (female)
2 yr old female	female (pregnant)	lgb*cc females	sex: female
400 yr. old female	female (outbred)	mare	female, other
adult female	female parent	female (worker)	female child
asexual female	female plant	monosex female	femal
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(gynoecious)	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)	gynoecious	thelytoky	hexaploid female
female (calf)	healthy female	female (gynoecious)	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

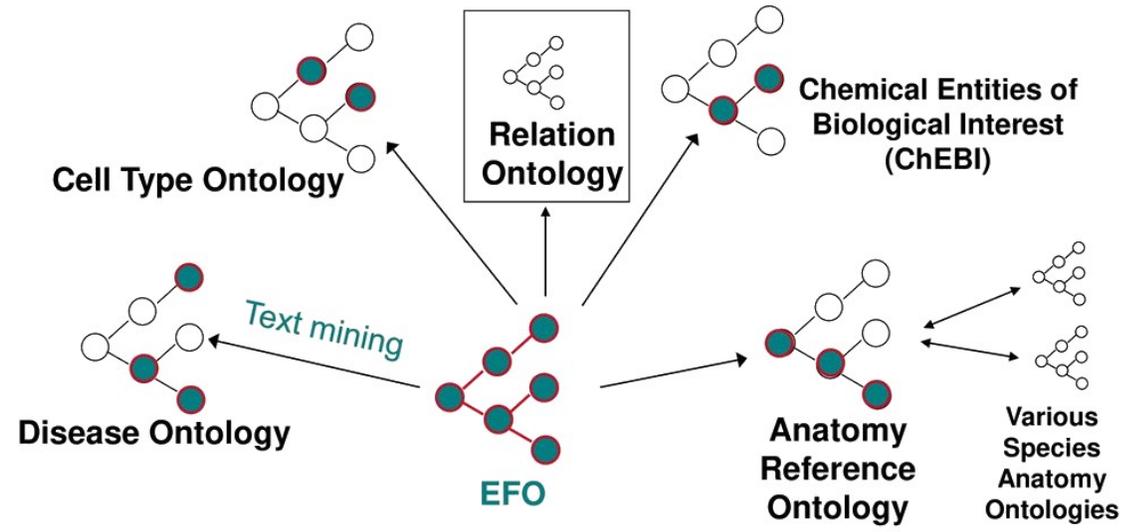
Ontologies

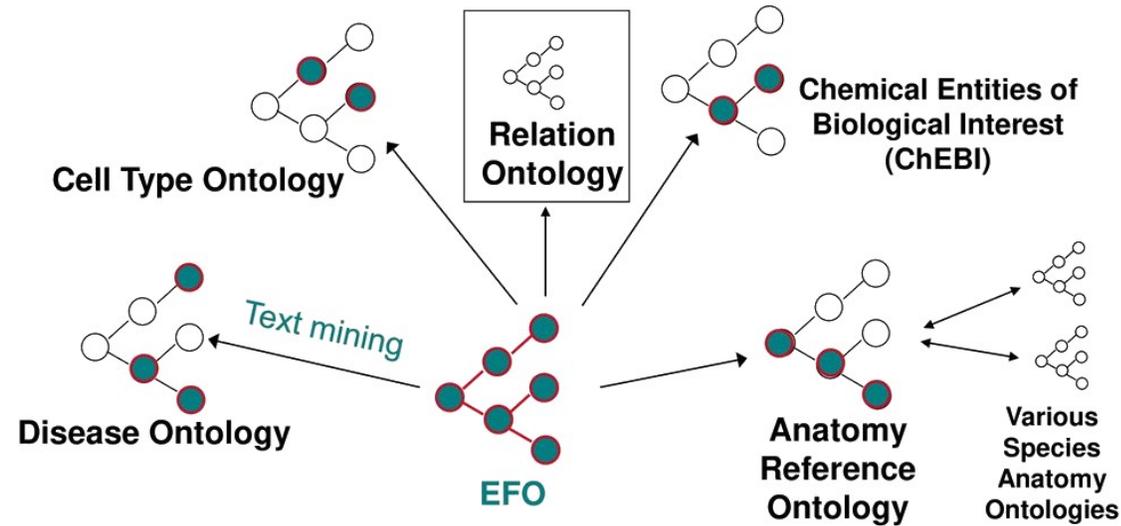
Ontology a set of concepts and categories in a subject area or domain that shows their properties and the relations between them.



Relationship	Color	Visibility
Extended nodes (*)	Red	-
is a	Teal	<input checked="" type="checkbox"/>
part of	Yellow	<input type="checkbox"/>
develops from	Light Green	<input type="checkbox"/>
contributes to morphology of	Pink	<input checked="" type="checkbox"/>
drains	Grey	<input type="checkbox"/>
supplies	Purple	<input checked="" type="checkbox"/>







☐ experimental factor

☐ material property

☐ quality

☐ sex

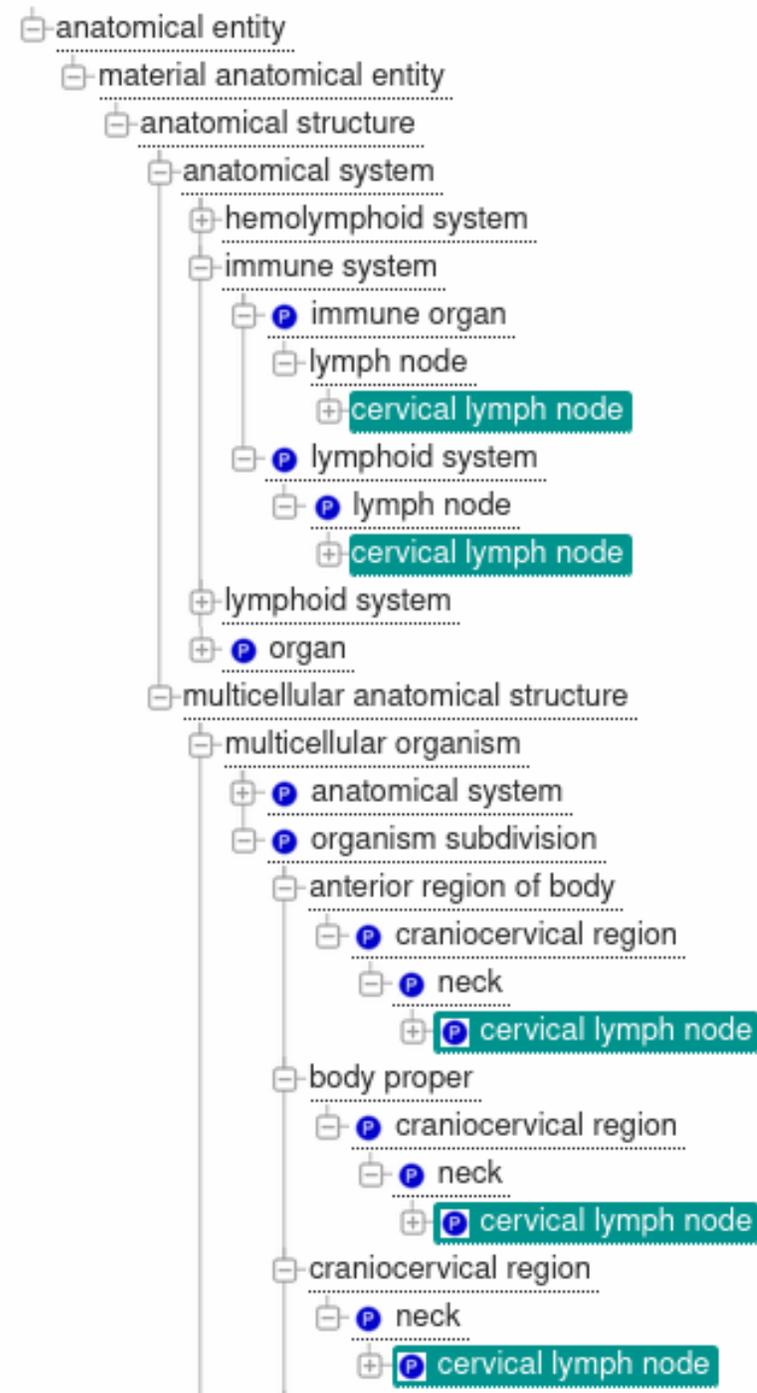
female

database cross reference

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

• ù Uberon

**Ontologies
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}	-

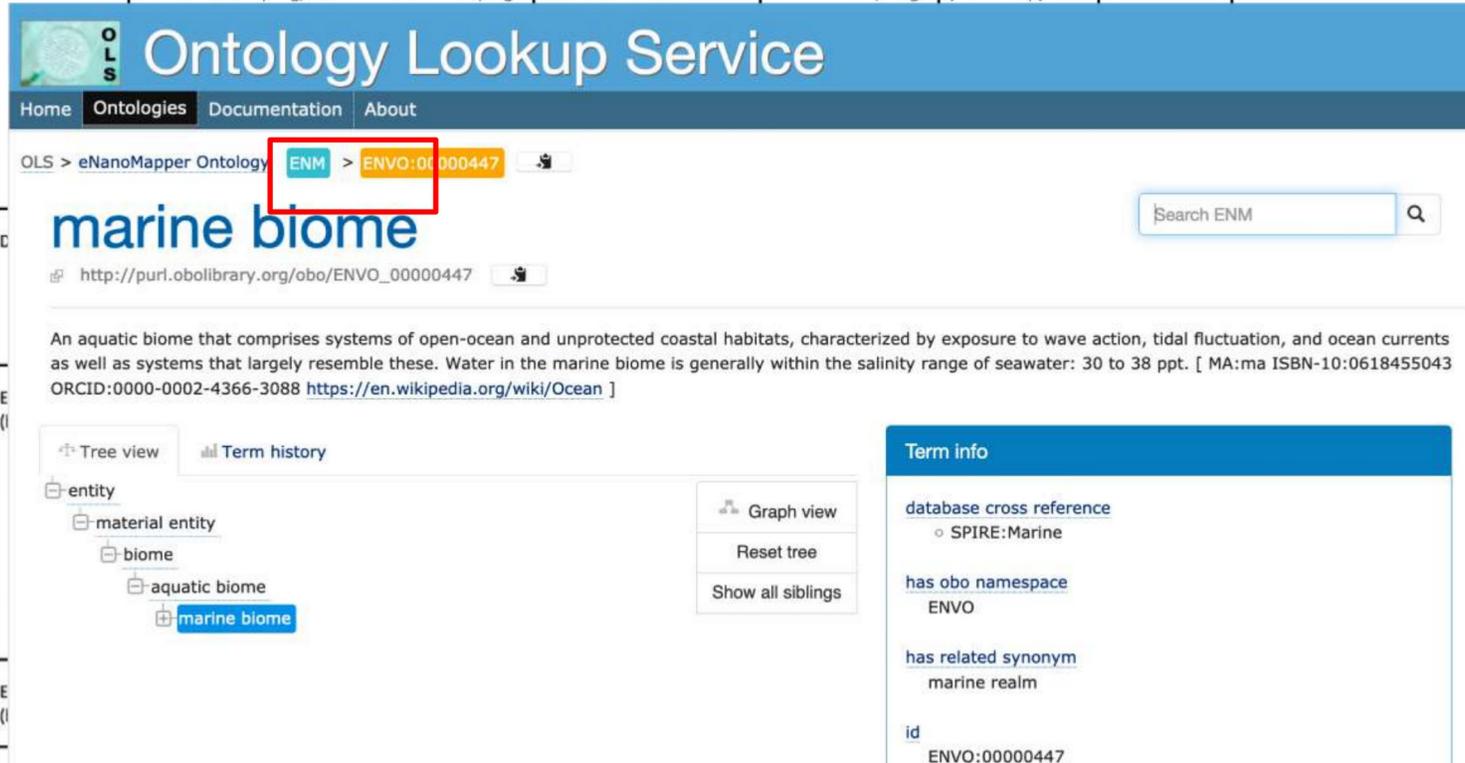
- Not collected -> missing
- 250 M -> 250
- Not applicable -> NA
- Superficial -> missing
- 1 m -> 1
- 2 m -> 2
- 2901.0 -> 2901
- 0 m. -> 0
- 1912 ft -> 582.80
- 40 mm from surface -> 0.04
- 0.75 m above seafloor -> missing
- 700meters -> 700
- Intracellular -> missing
- Surface water of 0 meter ->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

Controlled vocabulary & Ontology

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}	-



The screenshot shows the OLS interface for the 'marine biome' term. The breadcrumb path is 'OLS > eNanoMapper Ontology > ENM > ENVO:00000447'. The term 'marine biome' is highlighted in blue. Below the term, there is a search bar and a description: 'An aquatic biome that comprises systems of open-ocean and unprotected coastal habitats, characterized by exposure to wave action, tidal fluctuation, and ocean currents as well as systems that largely resemble these. Water in the marine biome is generally within the salinity range of seawater: 30 to 38 ppt. [MA:ma ISBN-10:0618455043 ORCID:0000-0002-4366-3088 https://en.wikipedia.org/wiki/Ocean]'. A tree view on the left shows the hierarchy: entity > material entity > biome > aquatic biome > marine biome. A 'Term info' panel on the right lists: database cross reference (SPIRE:Marine), has obo namespace (ENVO), has related synonym (marine realm), and id (ENVO:00000447).

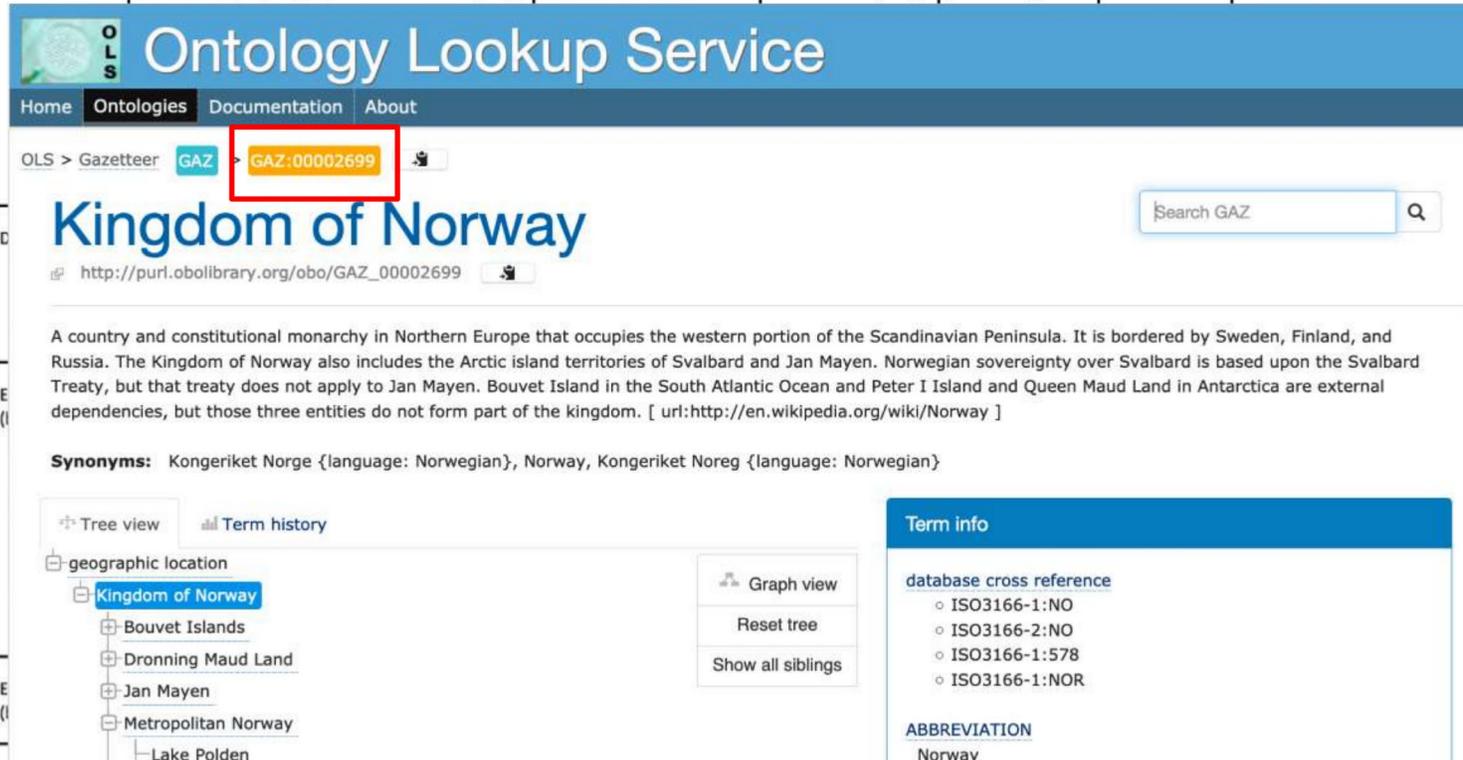
The ENVO ontology describes the environment of the sampling

Controlled vocabulary & Ontology

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}	-



The screenshot shows the OLS interface with the 'Ontology Lookup Service' header and navigation tabs for Home, Ontologies, Documentation, and About. The 'Ontologies' tab is active, showing a search for 'GAZ:00002699' which is highlighted with a red box. Below the search, the 'Kingdom of Norway' entry is displayed, including its URL, a description, synonyms, and a hierarchical tree view. The tree view shows 'geographic location' as the parent term, with 'Kingdom of Norway' as a child. Other children include 'Bouvet Islands', 'Dronning Maud Land', 'Jan Mayen', 'Metropolitan Norway', and 'Lake Polden'. A 'Term info' box on the right provides database cross references and an abbreviation.

The GAZ ontology describes the geographical location of the sampling



ONTOLOGY SEARCH

[Home](#)[Ontologies](#)[Documentation](#)[About](#)

Welcome to the EMBL-EBI Ontology Lookup Service

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

[Looking for a particular ontology?](#)

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 Star 5 Forks

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](#)

<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)

but often following the same concept:

Investigation

Persons
Organizations
Publications

Study(s)

Assay(s)



Technology & domain specific

but often following the same concept:

Investigation

Persons
Organizations
Publications

Study(s)

Design
Factor
Protocol

Assay(s)



Technology & domain specific

but often following the same concept:

Investigation

Persons
Organizations
Publications

Study(s)

Design
Factor
Protocol

Assay(s)

Measurement
Technology
Materials
Data



Technology & domain specific

but often following the same concept:

Investigation

Persons
Organizations
Publications

Study(s)

Design
Factor
Protocol

Controlled Vocabularies
Ontologies
Standards

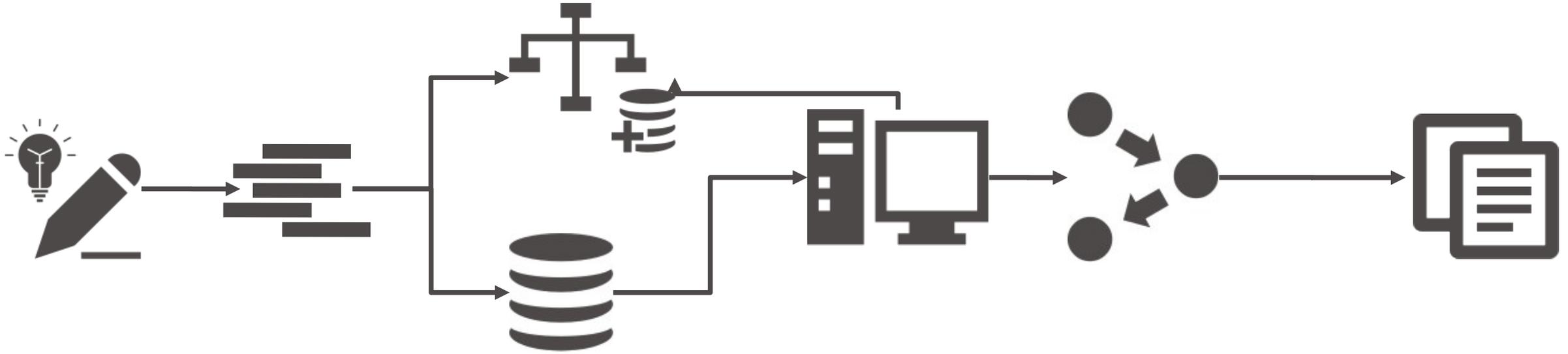
Assay(s)

Measurement
Technology
Materials
Data

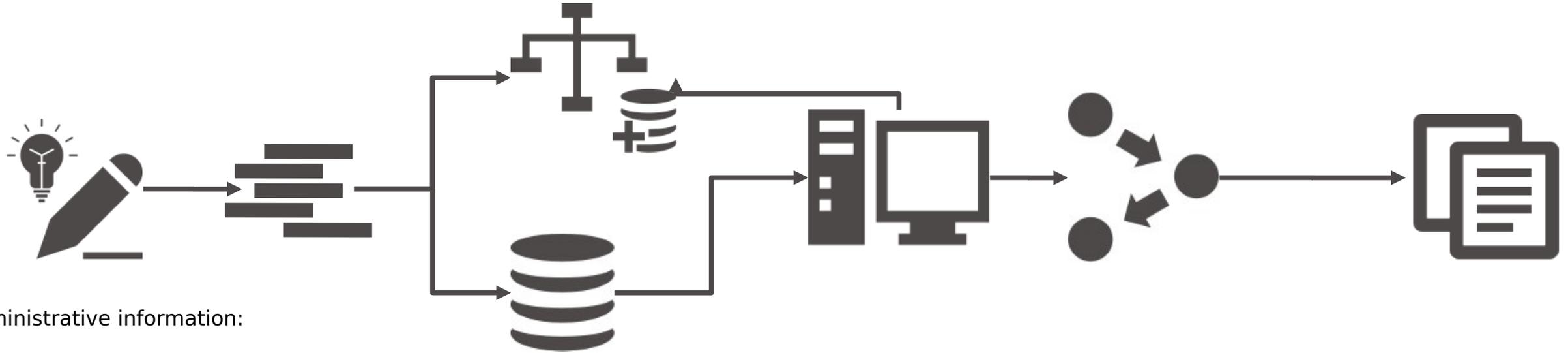


Technology & domain specific

MINSEQE



MINSEQE



Administrative information:

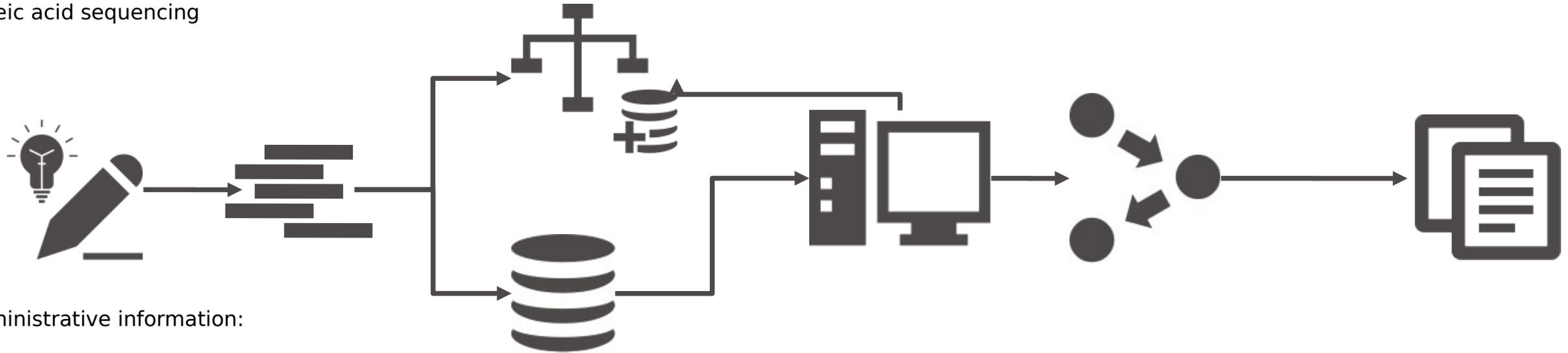
- Persons
- Organizations
- Publications

Experimental conditions/design

protocols:

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

MINSEQE



Administrative information:

Persons
Organizations
Publications

Experimental conditions/design

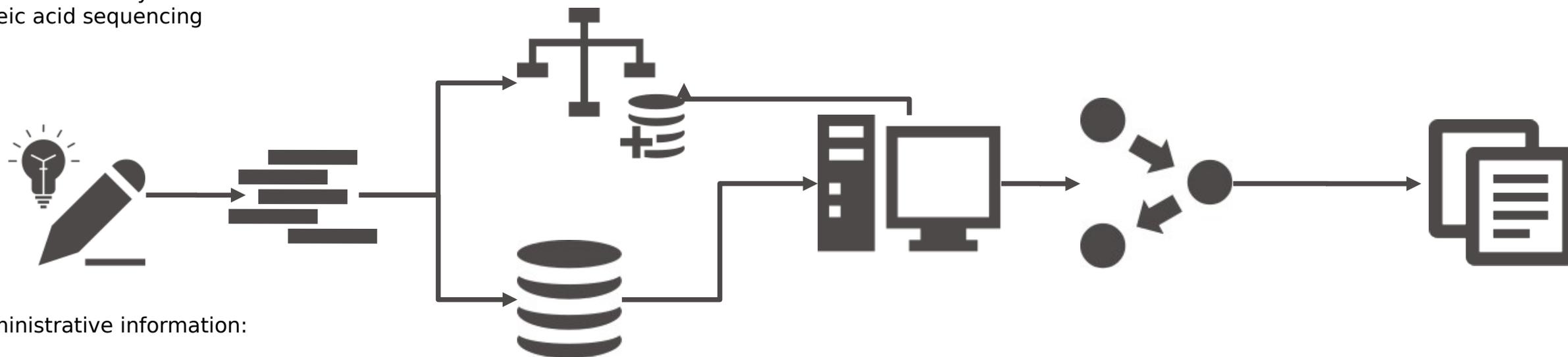
protocols:

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

MINSEQE

protocols:

high throughput sequence alignment
normalization data transformation



Administrative information:

Persons
Organizations
Publications

fastq
bam
csv/tsv

Experimental conditions/design

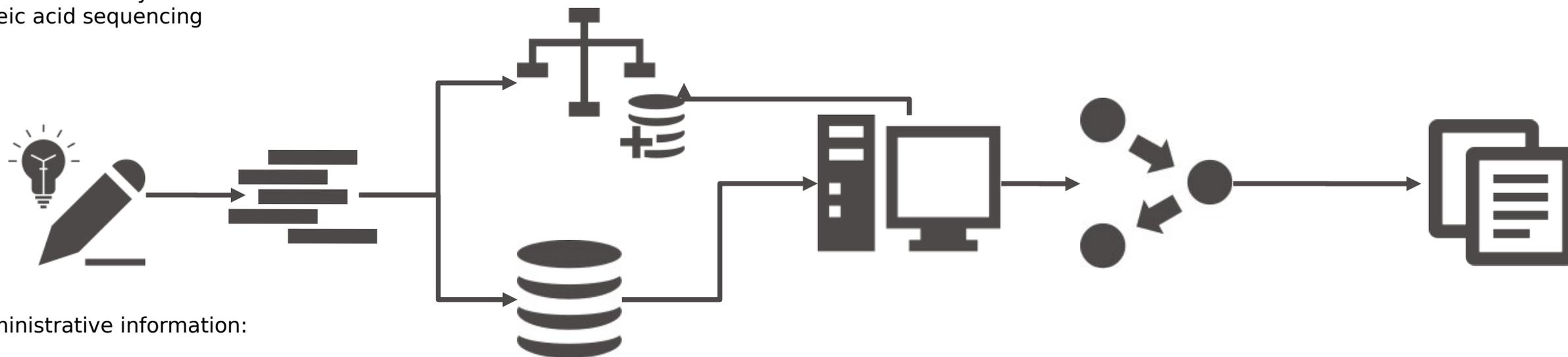
protocols:

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

MINSEQE

protocols:

high throughput sequence alignment
normalization data transformation



Administrative information:

Persons
Organizations
Publications

Experimental conditions/design

protocols:

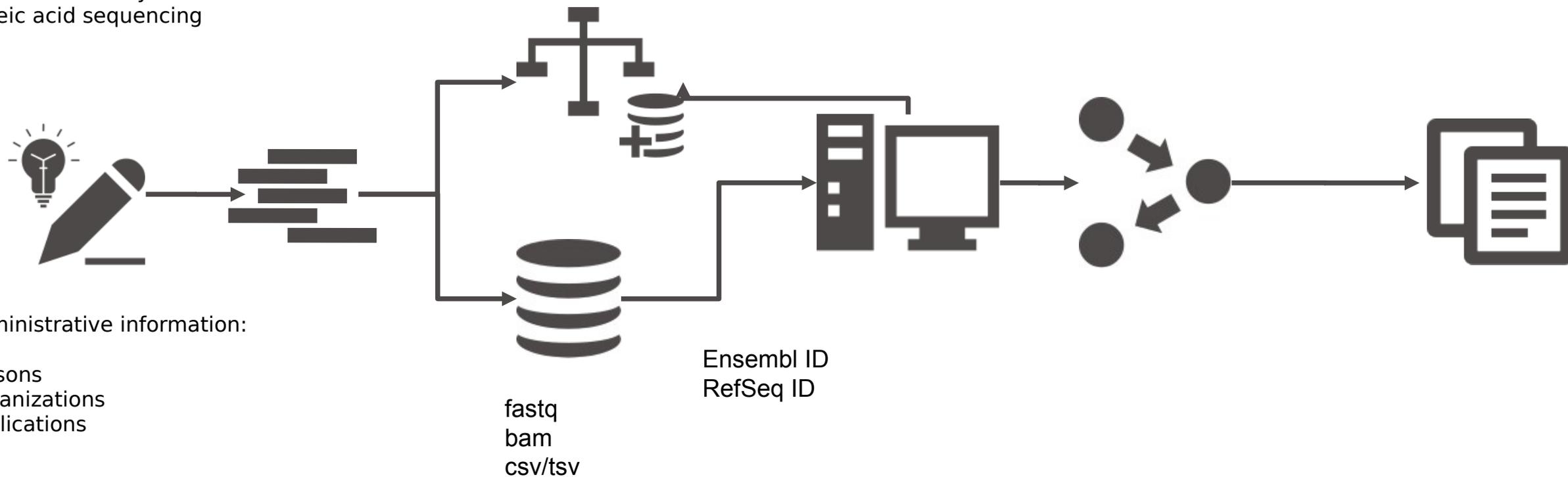
- treatment
- sample collection
- growth
- nucleic acid extraction
- conversion
- nucleic acid library construction
- nucleic acid sequencing

MINSEQE



protocols:

- high throughput sequence alignment
- normalization data transformation



Administrative information:

- Persons
- Organizations
- Publications

Experimental conditions/design

protocols:

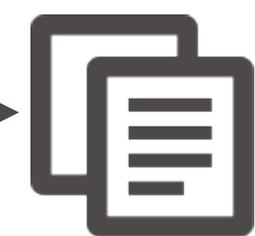
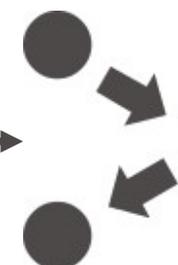
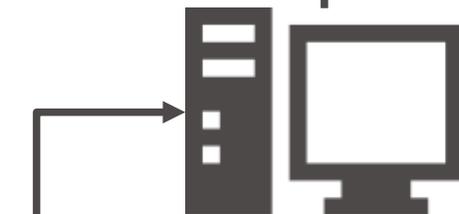
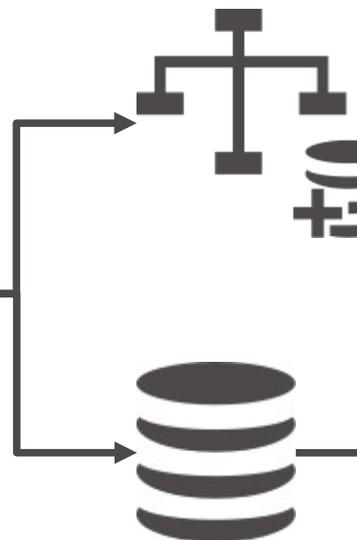
- treatment
- sample collection
- growth
- nucleic acid extraction
- conversion
- nucleic acid library constructi
- nucleic acid sequencing



MINSEQE

protocols:

- high throughput sequence alignment
- normalization data transformation



Administrative information:

- Persons
- Organizations
- Publications

fastq
bam
csv/tsv

Ensembl ID
RefSeq ID

Experimental conditions/design

protocols:

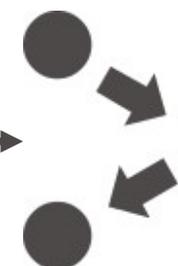
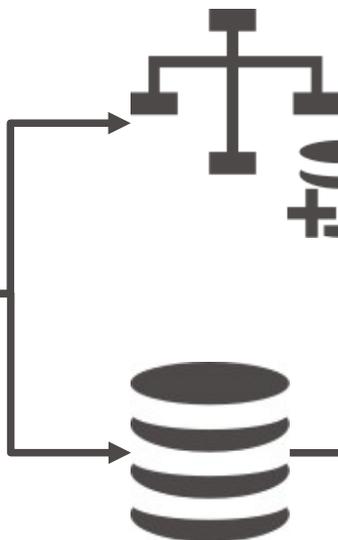
- treatment
- sample collection
- growth
- nucleic acid extraction
- conversion
- nucleic acid library constructi
- nucleic acid sequencing



MINSEQE

protocols:

- high throughput sequence alignment
- normalization data transformation



Administrative information:

- Persons
- Organizations
- Publications

fastq
bam
csv/tsv

Ensembl ID
RefSeq ID

Interlinking with other resources

Experimental conditions/design

protocols:

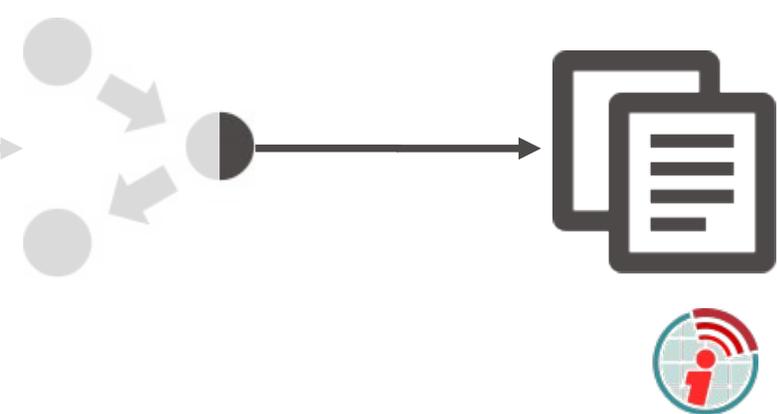
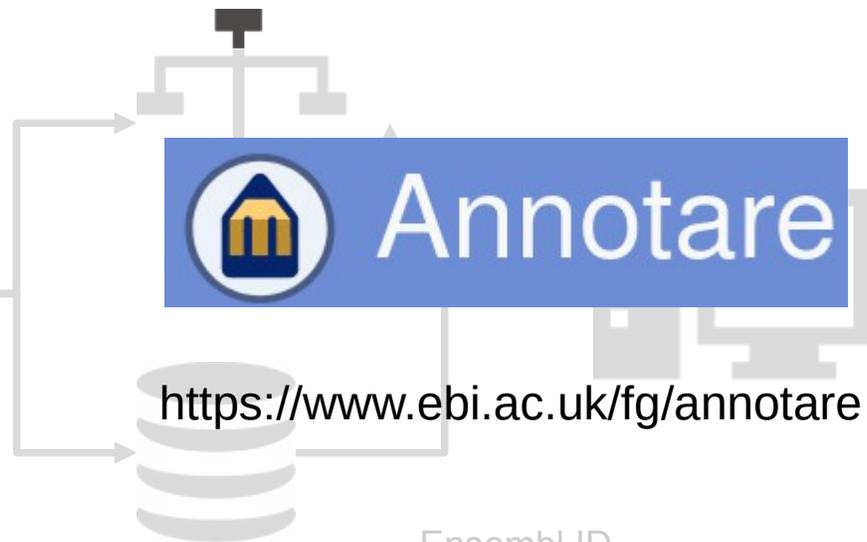
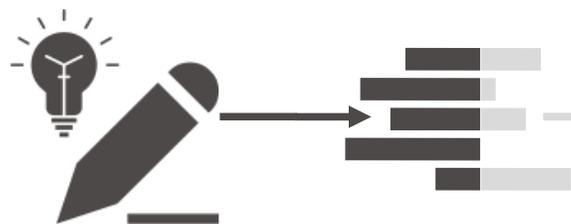
- treatment
- sample collection
- growth
- nucleic acid extraction
- conversion
- nucleic acid library construction
- nucleic acid sequencing



MINSEQE

protocols:

- high throughput sequence alignment
- normalization data transformation



Administrative information:

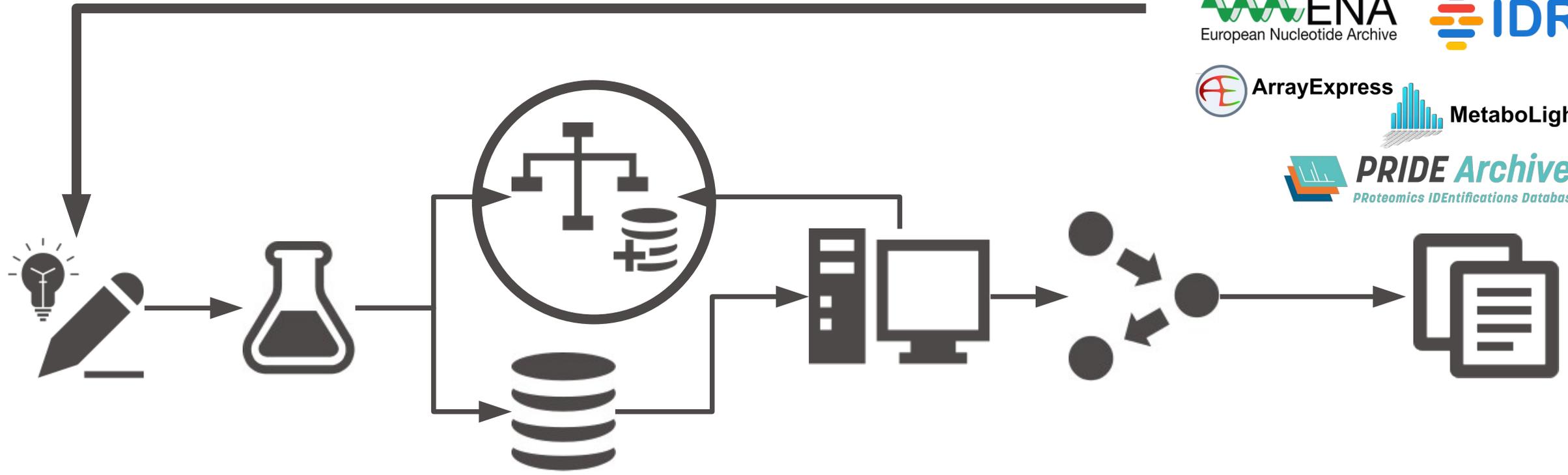
- Persons
- Organizations
- Publications

fastq
bam
csv/tsv

Ensembl ID
RefSeq ID



Interlinking with other resources



Meta data standards



ArrayExpress

MINSEQE

MIAME

...

Meta data standards



ArrayExpress

MINSEQE
MIAME

...



HUPO-PSI
TraML
MIAPE

...

Meta data standards



MINSEQE
MIAME

...



HUPO-PSI
TraML
MIAPE

...



SRA-XML

Which metadata standard?

FAIRsharing

The logo for FAIRsharing, featuring a stylized infinity symbol or figure-eight shape below the text. The left loop of the symbol is blue, and the right loop is black.

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



FASTQ
MAGE-
ML
...



mzML
mzQuantML
...



FASTA
FASTQ
...

Metadata tracking platforms

Domain specific:

COPO for plant sciences

MOLGENIS for biobanking

...



MOLGENIS

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



MOLGENIS



RightField



SEEK



File Edit Sheet Help

	A	B	C	D	E	F
1	# This is an excel template...					
2	# Use this template for ...					
3	# Click the Metadata Ex...					
4	# Field names (in blu...					
5	# CLICK HERE for the F...					
6						
7	SERIES					
8	# This section describes ...					
9						
10	title					
11	summary					
12	summary					
13	overall design					
14	contributor					
15	contributor (SEEK ID)					
16	SEEK Project	Project				
17	Experiment Class (a...	transcriptomics				
18	Experiment Design t...	ExperimentDesignT...				
19	Technology type	microarray				
20	quality control type	QualityControlDesc...				
21						
22	SAMPLES					
23	# The Sample name...					
24	# CLICK HERE to find t...					
25						
26	Sample name	title	CEL file	source name	organism	characteristics... char
27	SAMPLE 1				organism	
28	SAMPLE 2				organism	
29	SAMPLE 3				organism	
30	SAMPLE 4				organism	
31	SAMPLE 5				organism	
32	SAMPLE 6				organism	
33	SAMPLE 7				organism	
34	SAMPLE 8				organism	
35	SAMPLE 9				organism	
36	SAMPLE X				organism	
37						
38						
39	PROTOCOLS					
40	# This section includes pr...					
41	# Protocols which are ap...					
42						
43	growth protocol					
44	treatment protocol					
45	extract protocol					
46	label protocol					

Metadata Template Matrix Template Metadata Example Matrix Example 1 Matrix Example 2

Selected cells: B17:B17

ONTOLOGY HIERARCHIES

MGEDOntology.owl x

JERMOntology x

- ExperimentAnalysisType
 - fluxomics
 - genomics
 - interactomics
 - metabolomics
 - proteomics
 - reactomics
 - single_cell
 - transcriptomics
- informaticsAnalysisType
- ModelAnalysisType
- CultureGrowth
- FactorsStudied
 - concentration
 - expression

TYPE OF ALLOWED VALUES

Free text

Direct subclasses

Subclasses

Instances

Direct instances

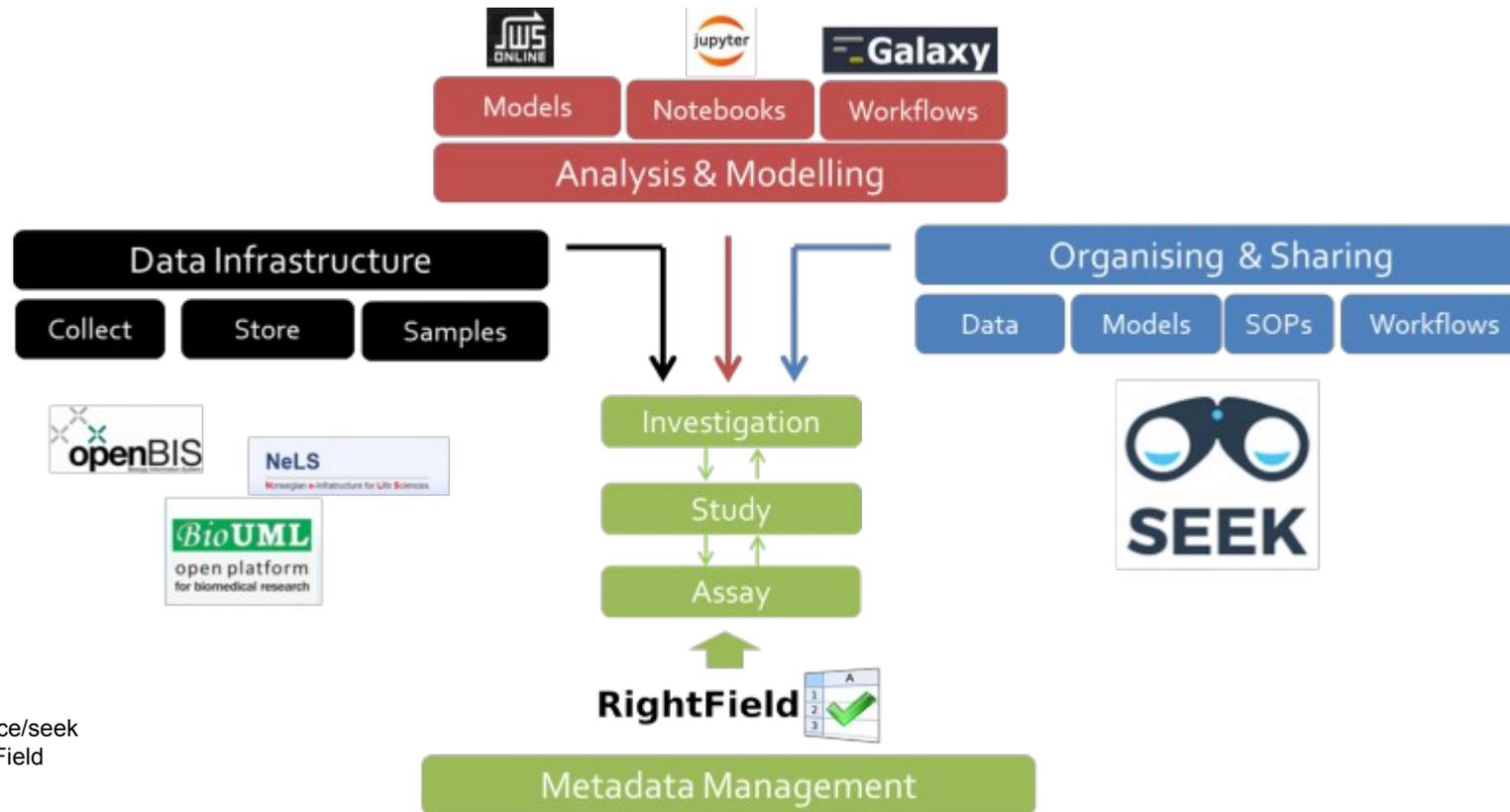
ALLOWED VALUES

- Comparative genomic hybridization
- RNAi
- gene expression profiling
- methylation profiling
- microRNA profiling
- tiling path

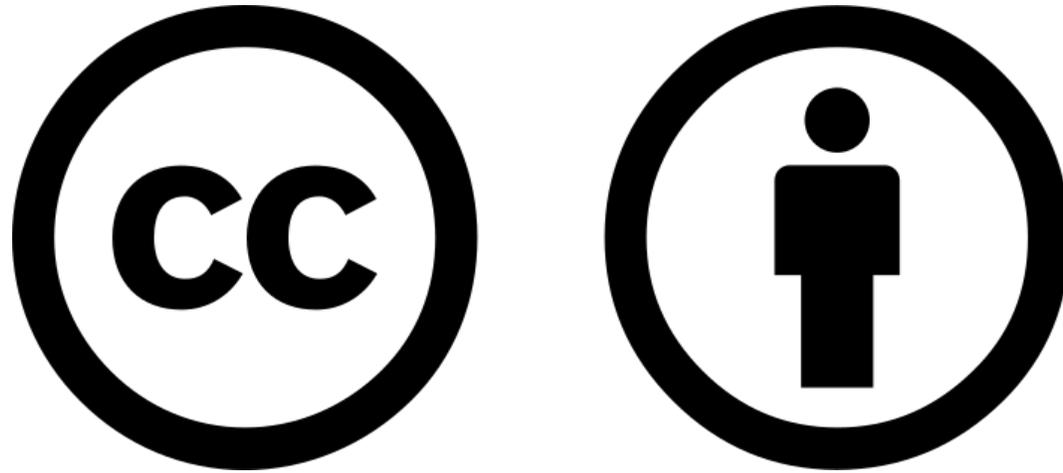
Apply



FAIRDOM integration



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



Except where otherwise noted, this work is licensed under:

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