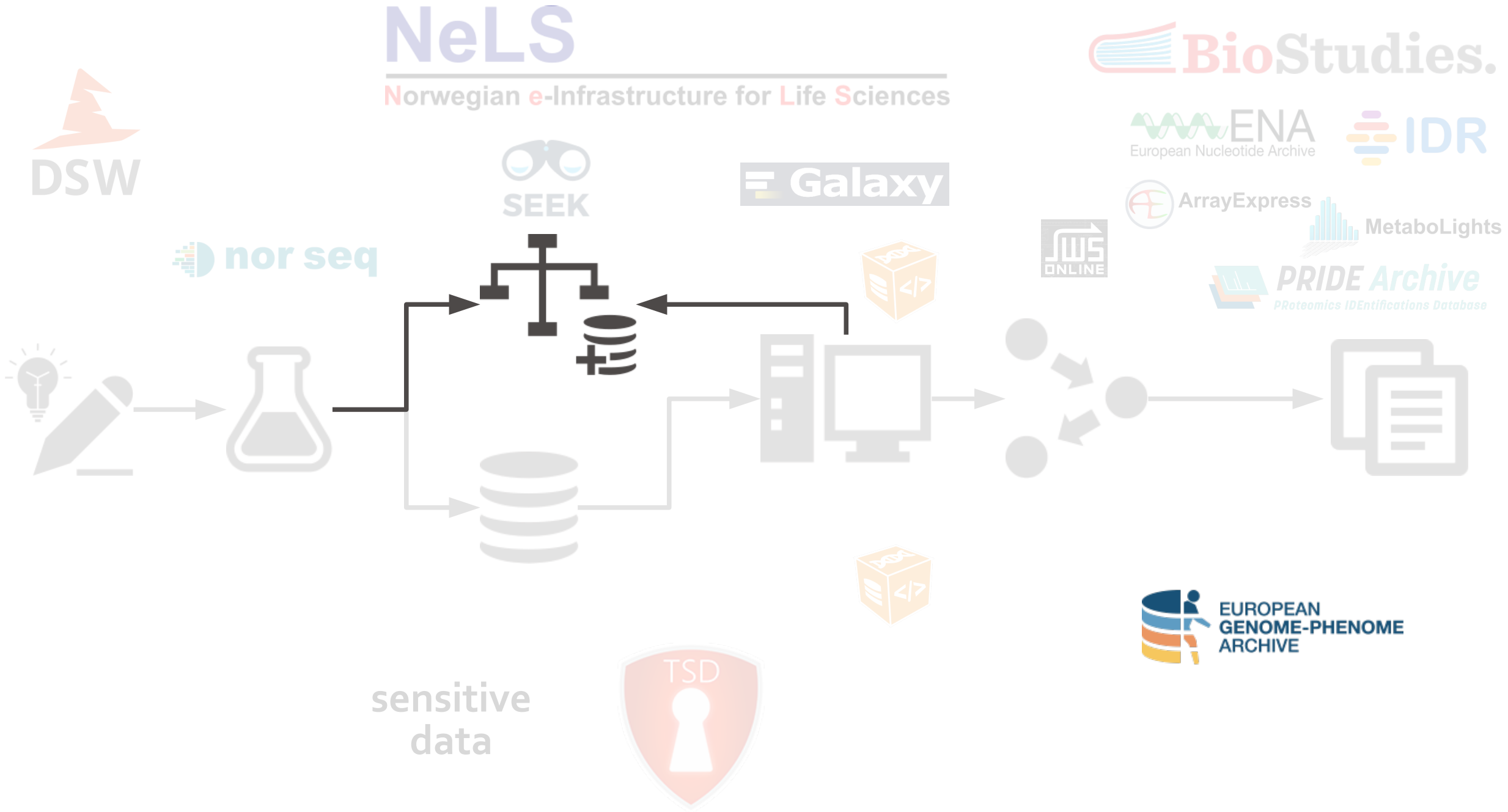
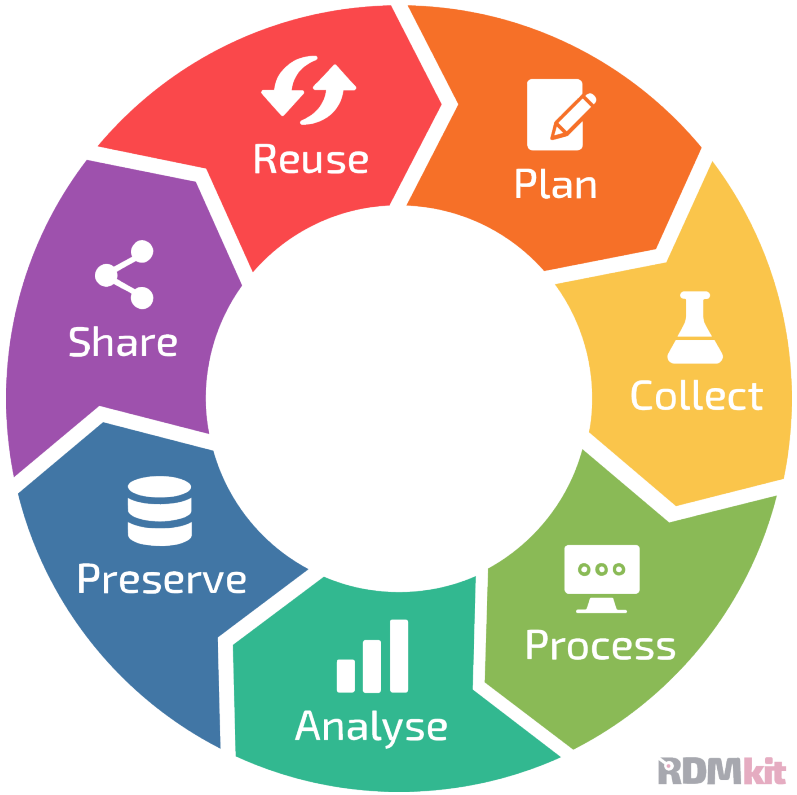
Metadata, Persistent identifiers & Ontologies

Korbinian Bösl
Data management coordinator
ELIXIR Norway/Digital Life Norway


NeLS
Norwegian e-Infrastructure for Life Sciences
BioStudies.
ENA
European Nucleotide Archive
IDR
DSW
SEEK
Galaxy
ArrayExpress
MetaboLights
nor seq
ONLINE
PRIDE Archive
PRoteomics IDEntifications Database
EUROPEAN
GENOME-PHENOME
ARCHIVE
sensitive
data
TSD

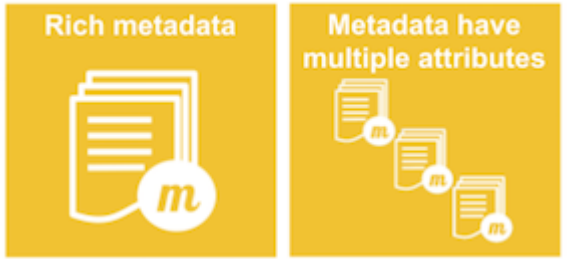
- Data life cycle +
- Your role +
- Your domain +
- Your problem -
  - Compliance monitoring
  - Data analysis
  - Data management plan
  - Data organisation
  - Data protection
  - Data publication
  - Data quality
  - Data storage
  - Data transfer
  - Identifiers
  - Licensing
  - **Documentation and metadata**
  - Sensitive data
- All tools and resources
- Tool assembly +



RDMkit

Link to RDMkit: https://rdmkit.elixir-europe.org/

What is metadata?



Experimental design

Outcome = Treatment effect + Biological effect + Technical effects + Error

Lazic, S.E. (2016). Cambridge University Press.

Experimental design

Outcome = Treatment effect + Biological effect + Technical effects + Error

Environment
Compound
Infection
Inhibitor
siRNA
sgRNA
Dose
Time

Lazic, S.E. (2016). Cambridge University Press.

Experimental design

Outcome = Treatment effect + Biological effect + Technical effects + Error

Environment	Sex
Compound	Age
Infection	Weight
Inhibitor	Litter
siRNA	Genotype
sgRNA	Species
Dose	Cell line
Time	

Lazic, S.E. (2016). Cambridge University Press.

Experimental design

Outcome = Treatment effect + Biological effect + Technical effects + Error

Environment	Sex	Operator
Compound	Age	Batch
Infection	Weight	Plate
Inhibitor	Litter	Cage
siRNA	Genotype	Array
sgRNA	Species	Flowcell
Dose	Cell line	Instrument
Time		Day
		Order
		Source

Lazic, S.E. (2016). Cambridge University Press.

Experimental design

Outcome = Treatment effect + Biological effect + Technical effects + Error

Treatment effect	Biological effect	Technical effects	Error
Environment	Sex	Operator	Experimental
Compound	Age	Batch	Treatment
Infection	Weight	Plate	Sampling
Inhibitor	Litter	Cage	Measurement
siRNA	Genotype	Array	
sgRNA	Species	Flowcell	
Dose	Cell line	Instrument	
Time		Day	
		Order	
		Source	

Lazic, S.E. (2016). Cambridge University Press.

Experimental design

“Data”

“Metadata”

Outcome = Treatment effect + Biological effect + Technical effects + Error

Treatment effect	Biological effect	Technical effects	Error
Environment	Sex	Operator	Experimental
Compound	Age	Batch	Treatment
Infection	Weight	Plate	Sampling
Inhibitor	Litter	Cage	Measurement
siRNA	Genotype	Array	
sgRNA	Species	Flowcell	
Dose	Cell line	Instrument	
Time		Day	
		Order	
		Source	

Lazic, S.E. (2016). Cambridge University Press.

What is a PID and why do we need it?

Science
AAAS

REPORT

One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds

Erik A. Schultes, David P. Bartel*

Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142, USA.

– Hide authors and affiliations

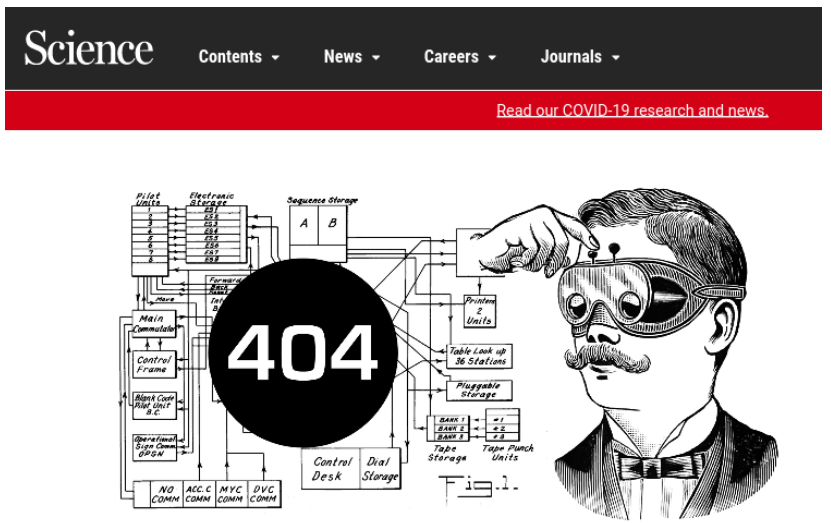
Science 21 Jul 2000:
Vol. 289, Issue 5478, pp. 448-452

https://doi.org/10.1126/science.289.5478.448

25. Supplemental data showing the predicted secondary structures of each construct (Fig. 3) and explaining the ligation activity of truncated ribozymes (Fig. 2B) are available at *Science* Online at www.sciencemag.org/feature/data/1050240.shl.

25. Supplemental data showing the predicted secondary structures of each construct (Fig. 3) and explaining the ligation activity of truncated ribozymes (Fig. 2B) are available at *Science* Online at www.sciencemag.org/feature/data/1050240.shl.

25. Supplemental data showing the predicted secondary structures of each construct (Fig. 3) and explaining the ligation activity of truncated ribozymes (Fig. 2B) are available at *Science* Online at www.sciencemag.org/feature/data/1050240.shl.



A PID consists of 2 components:

a unique identifier

a service that locates the resource over time even when it's location changes

Examples for digital objects

Digital Object Identifiers

Handels

Archival Resource Keys (ARK)

Persistent Uniform Resource Locator (URL)

Identifiers.org

PIDs exits also for

- Persons ORCID

PIDs exits also for

- Persons ORCID
- $ Funding bodies Crossref

PIDs exits also for

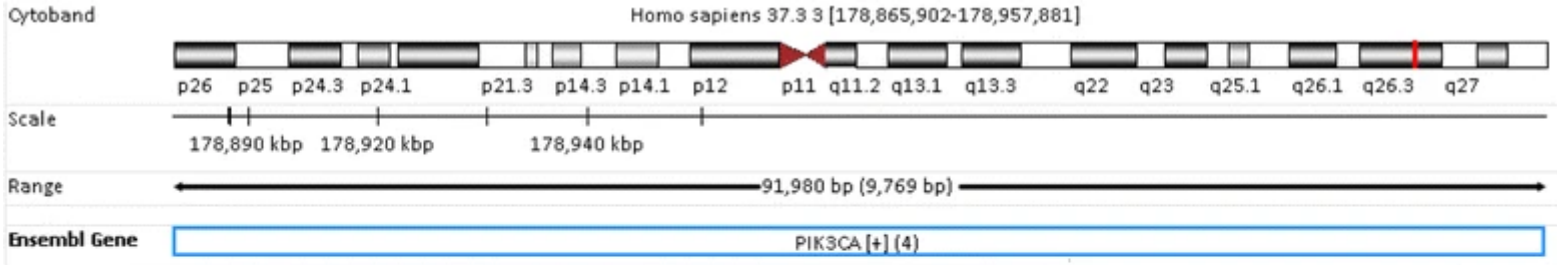
- Persons ORCID
- $ Funding bodies Crossref
- Institutions

PIDs exits also for

- Persons ORCID
- $ Funding bodies Crossref
- Institutions
- Instruments (soon)

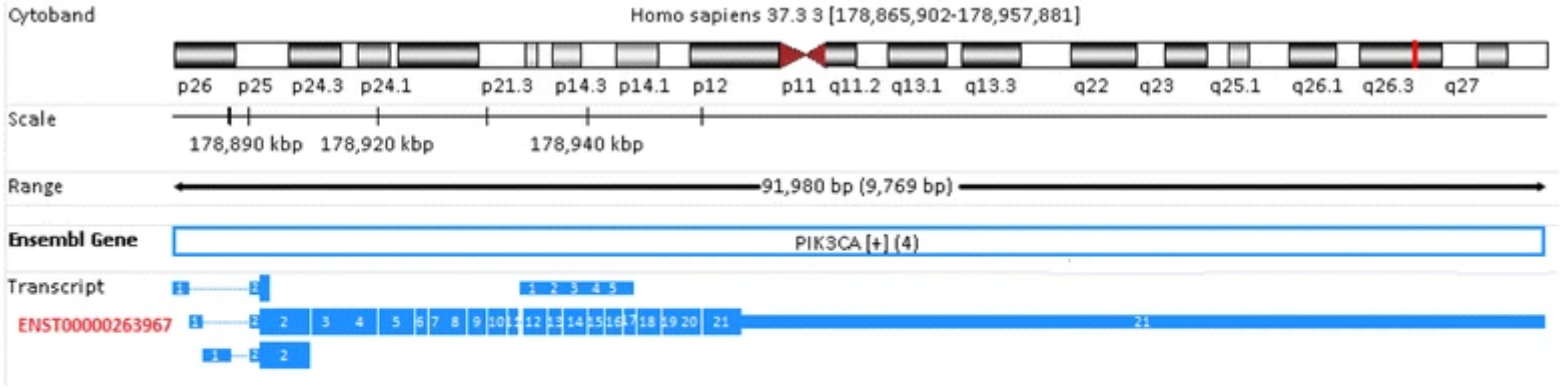
Sequence identifiers:

- Gene: PIK3CA



Sequence identifiers:

- Gene: PIK3CA



Sequence identifiers:

- Gene: PIK3CA

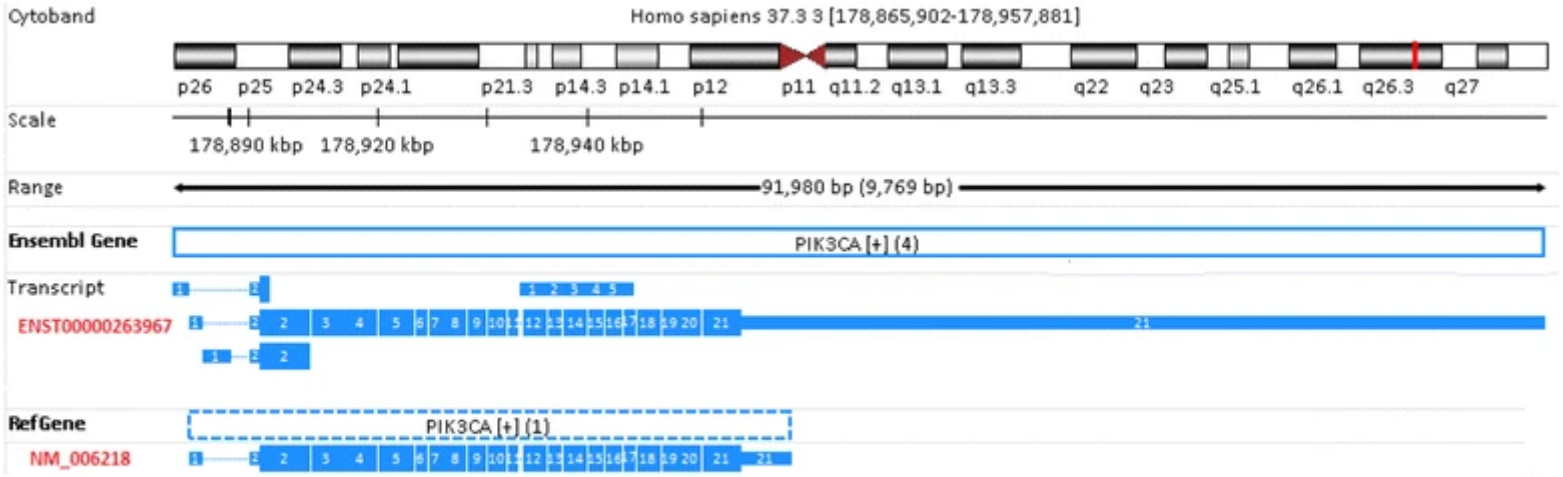


Figure 6 from Zhao, S., and Zhang, B. (2015). BMC Genomics 16 licensed under Creative Commons Attribution 4.0 International License

Sequence identifiers:

- Gene: PIK3CA

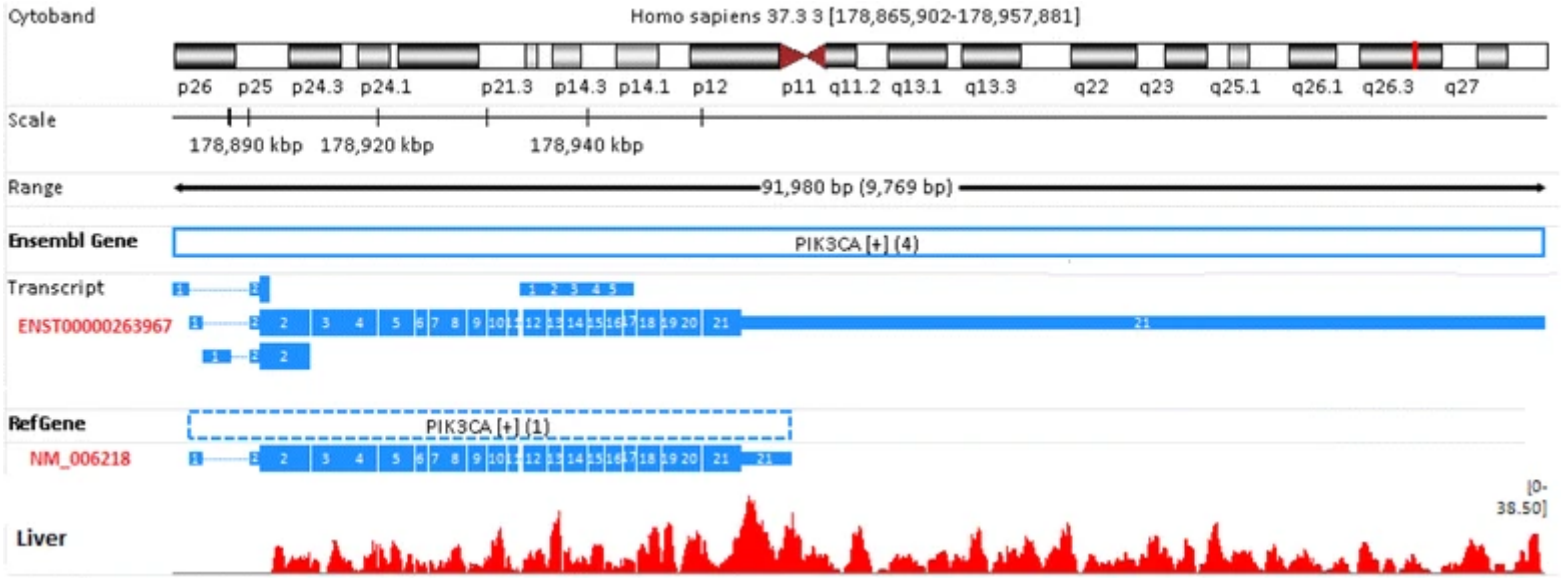
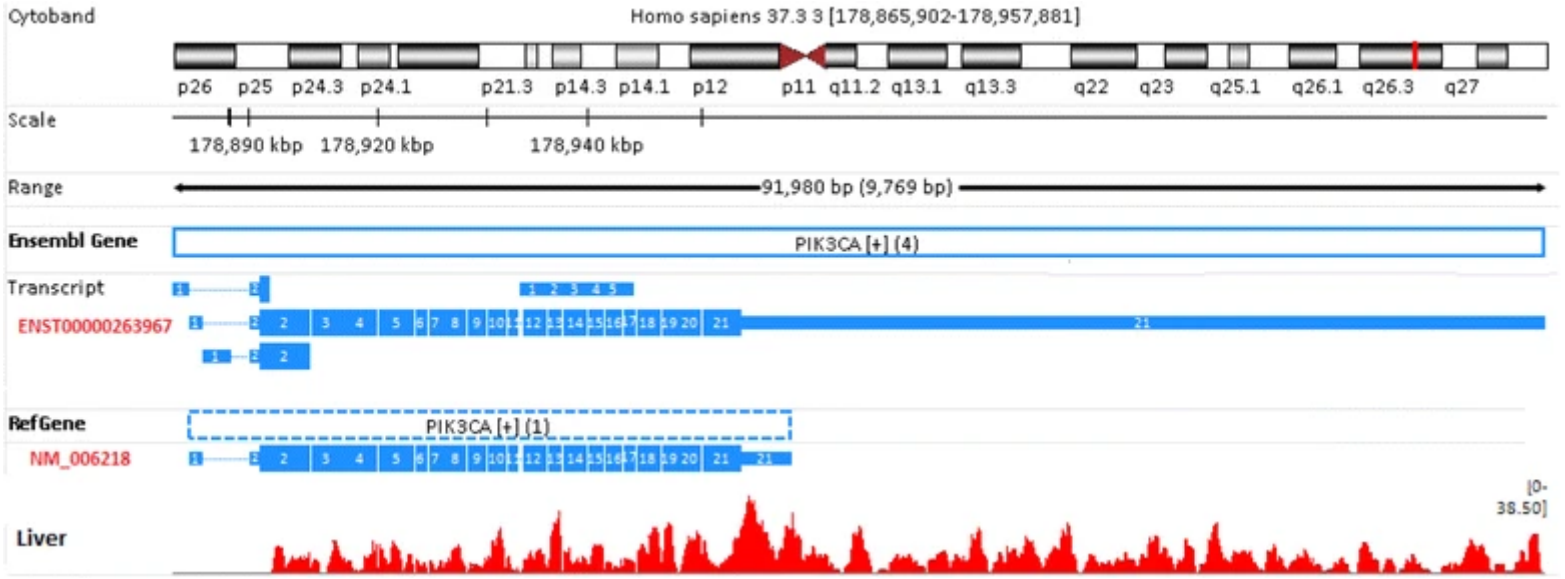


Figure 6 from Zhao, S., and Zhang, B. (2015). BMC Genomics 16 licensed under Creative Commons Attribution 4.0 International License

Sequence identifiers:

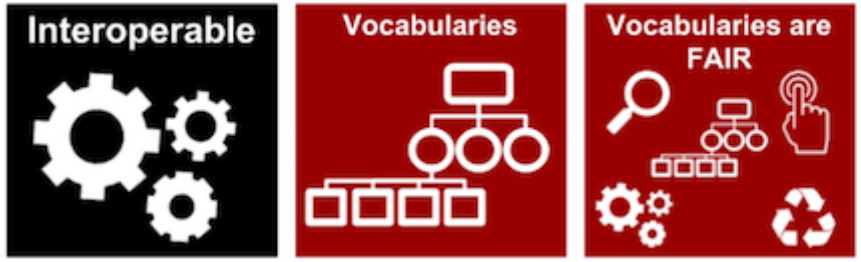
- Gene: PIK3CA



ENST00000263967.2
3
4

Figure 6 from Zhao, S., and Zhang, B. (2015). BMC Genomics 16 licensed under Creative Commons Attribution 4.0 International License

Why do we need standard vocabularies?



How many way can you say “female”?

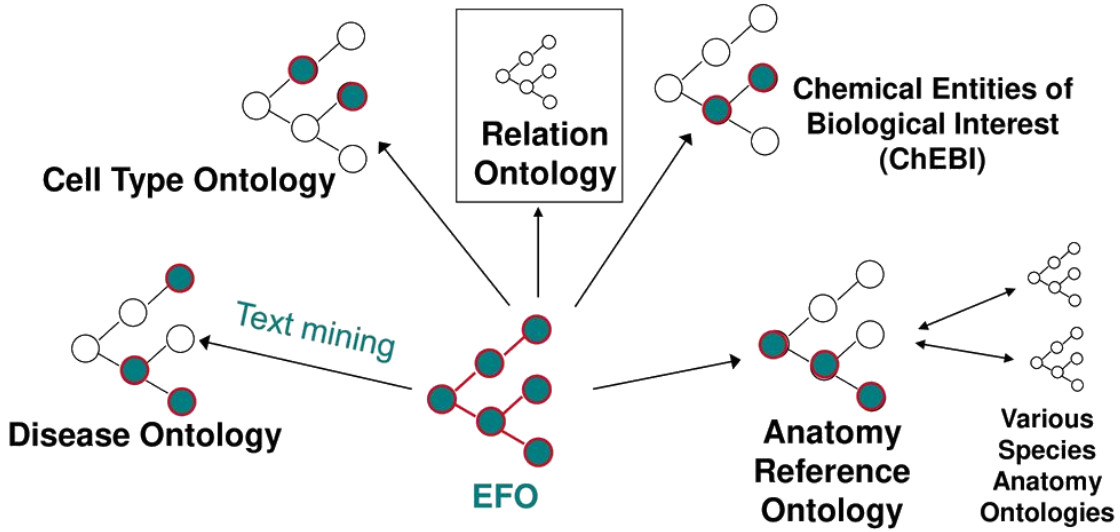
How many way can you say “female”?

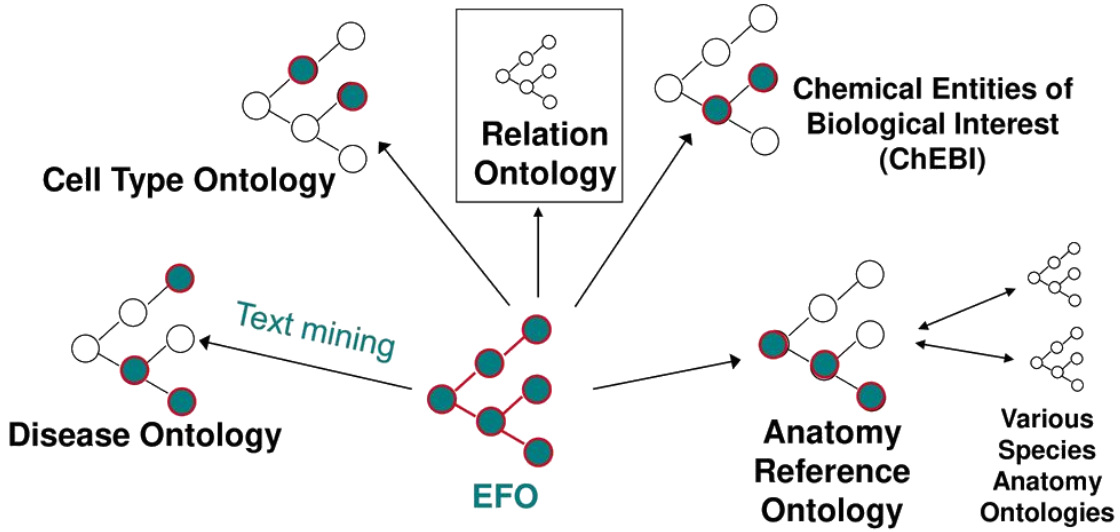
18-day pregnant females	female (lactating)	individual female	worker caste (female)
2 yr old female	female (pregnant)	lgb*cc females	sex: female
400 yr. old female	female (outbred)	mare	female, other
adult female	female parent	female (worker)	female child
asexual female	female plant	monosex female	femal
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	pseudohermaprhoditic female	femlale
diploid female	female(gynoecious)	remale	metafemale
f	femele	semi-engorged female	sterile female
famale	female, pooled	sexual oviparous female	normal female
femail	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


efo


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



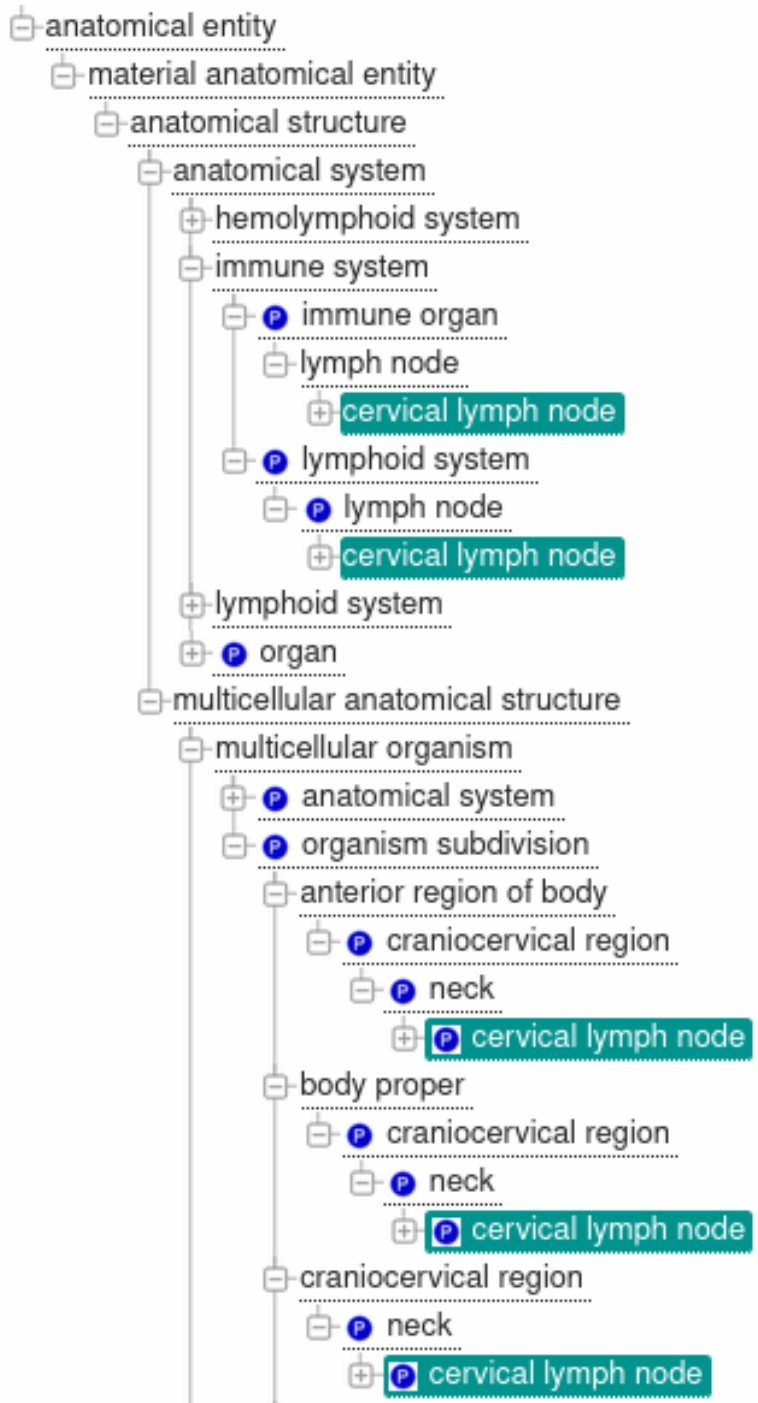
- experimental factor
  - material property
    - quality
      - sex
        - female

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

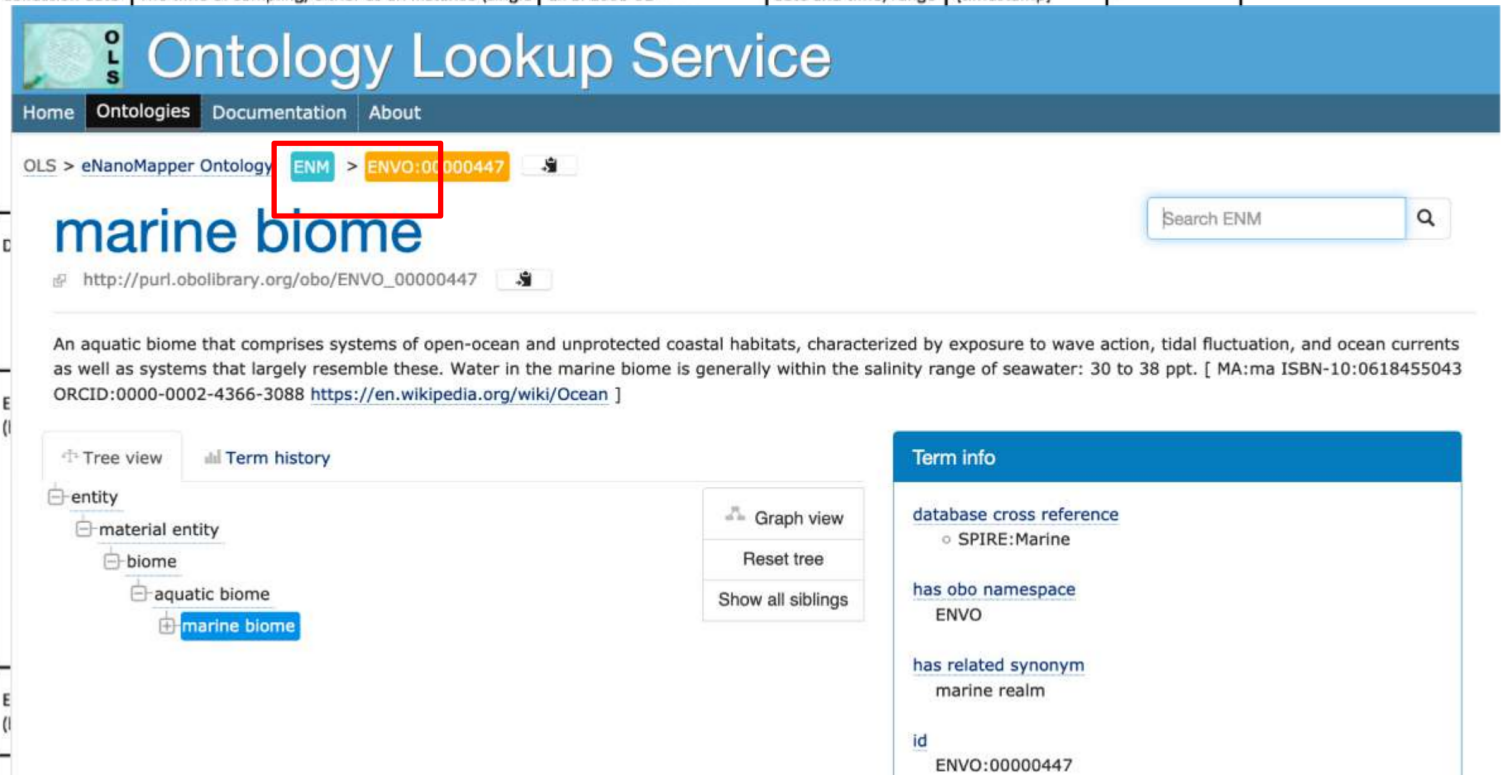
Structured comment name	Item	Description	Examples	Expected value	Value syntax	Preferred units / suffix
alt_elev	Geographic location (altitude/elevation)	Sample taken at given elevation above sea level, defined in meters(m) as a positive floating number with two decimals.	Ex 1: 3.06 Ex 2: 1.80-2.15	-	{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 & Ontologies

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					

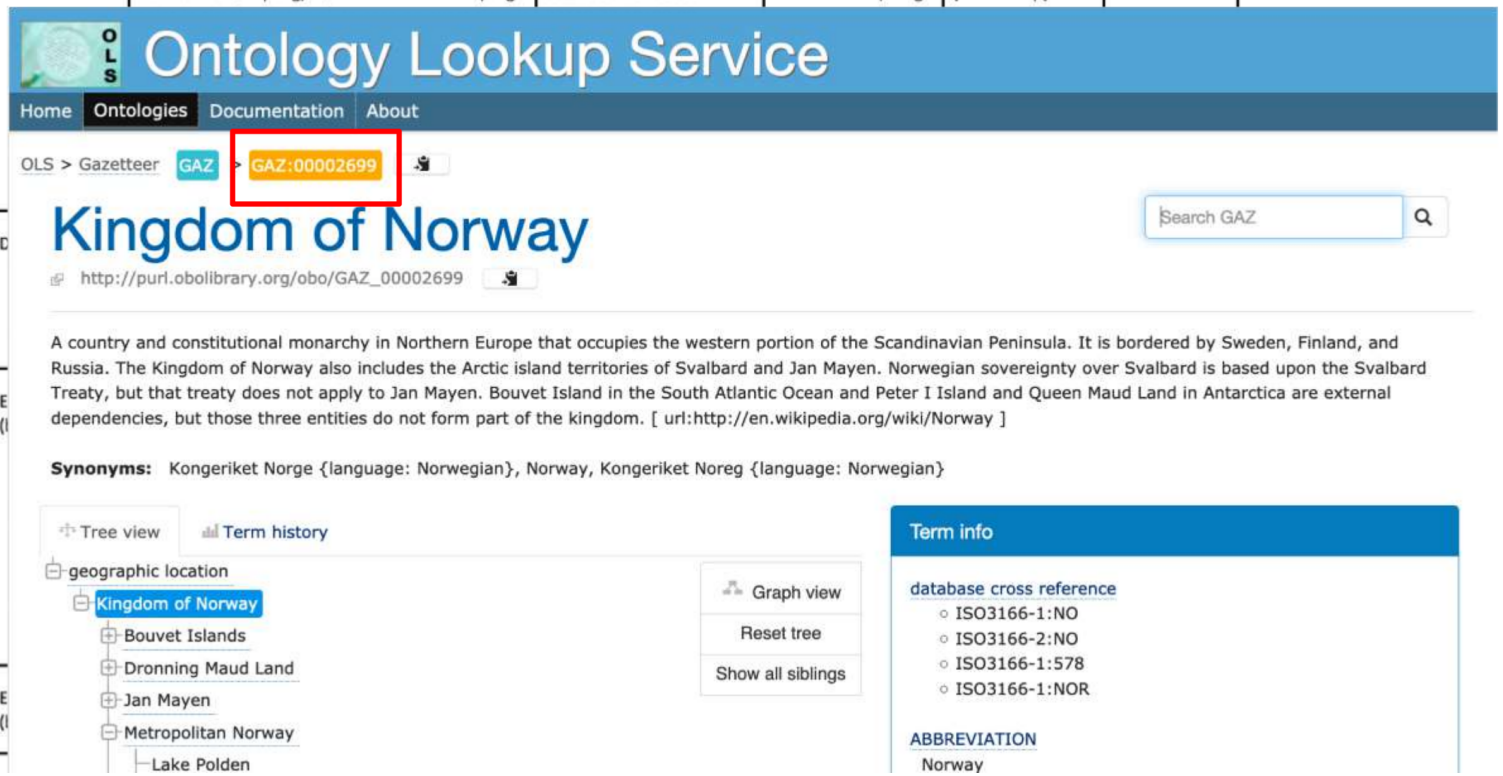


The ENVO ontology describes the environment of the sampling

Controlled vocabulary & Ontologies

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

What is a metadata standard?

but often following the same concept:

Investigation

Study(s)

Assay(s)

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)

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

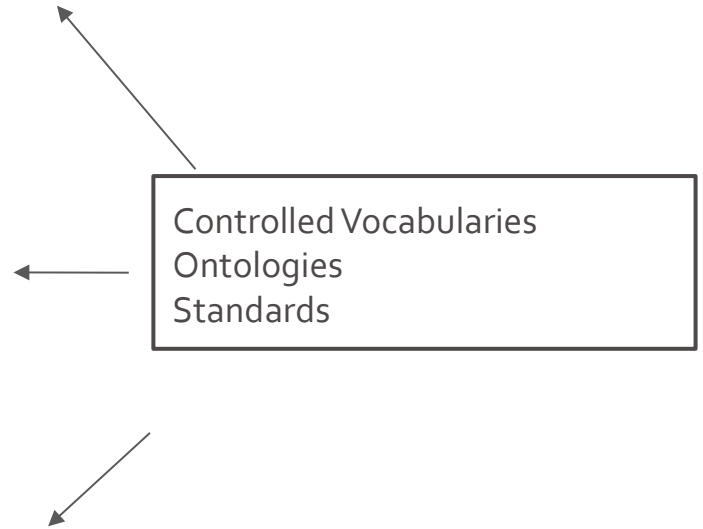
Persons
Organizations
Publications

Study(s)

Design
Factor
Protocol

Assay(s)

Measurement
Technology
Materials
Data

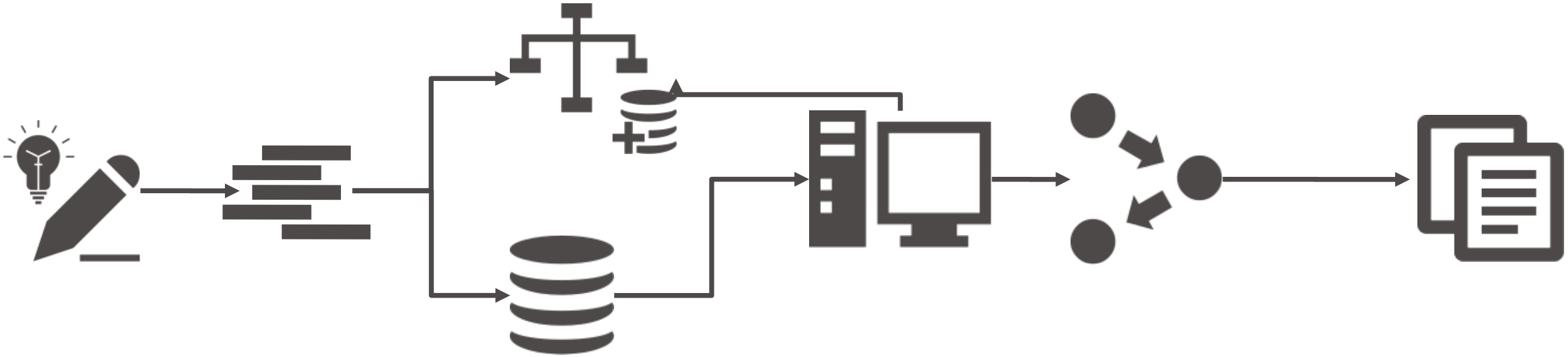


FAIRsharing.org
standards, databases, policies

Technology & domain specific

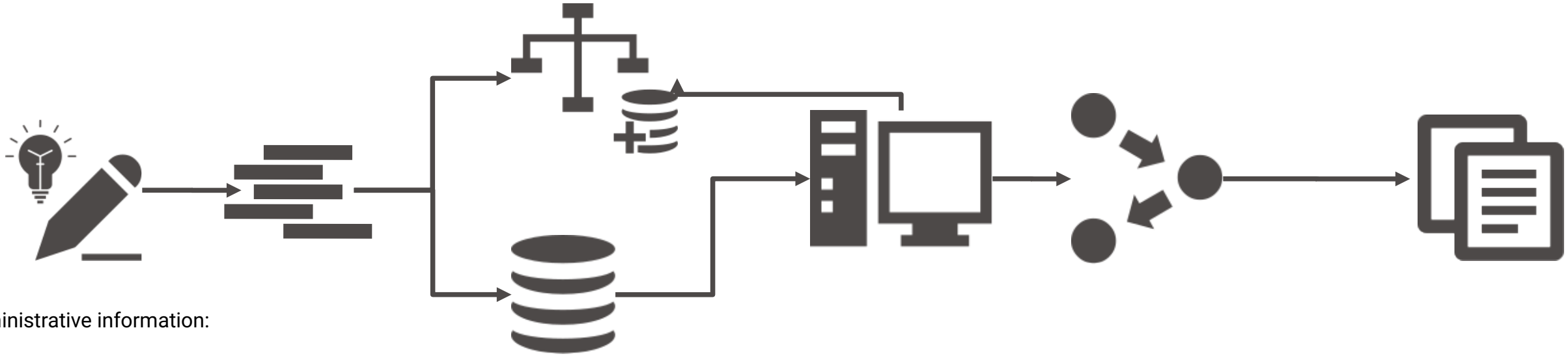
MINSEQE

ArrayExpress



MINSEQE

ArrayExpress



Administrative information:

Persons
Organizations
Publications

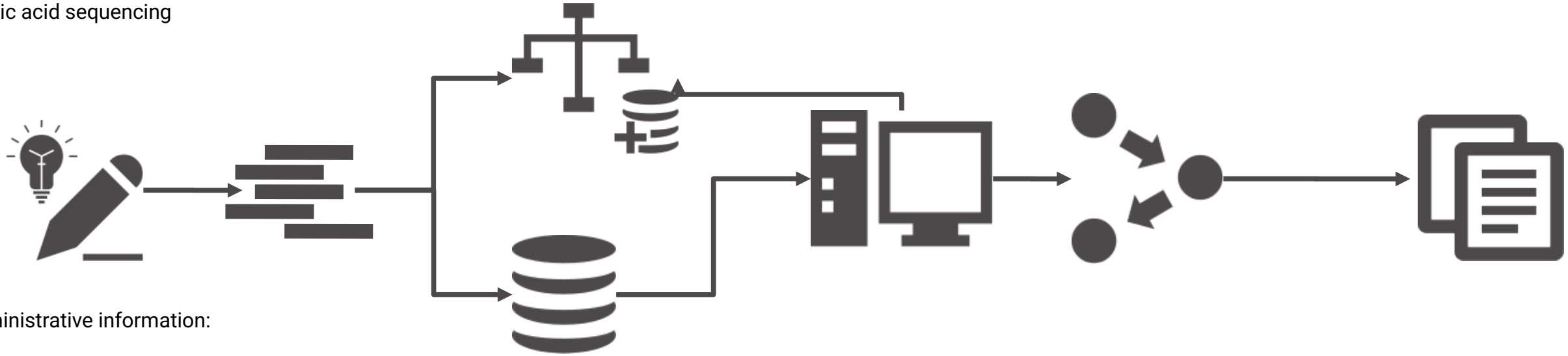
Experimental conditions/design

protocols:

MINSEQE

ArrayExpress

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



Administrative information:

Persons
Organizations
Publications

MINSEQE

Experimental conditions/design

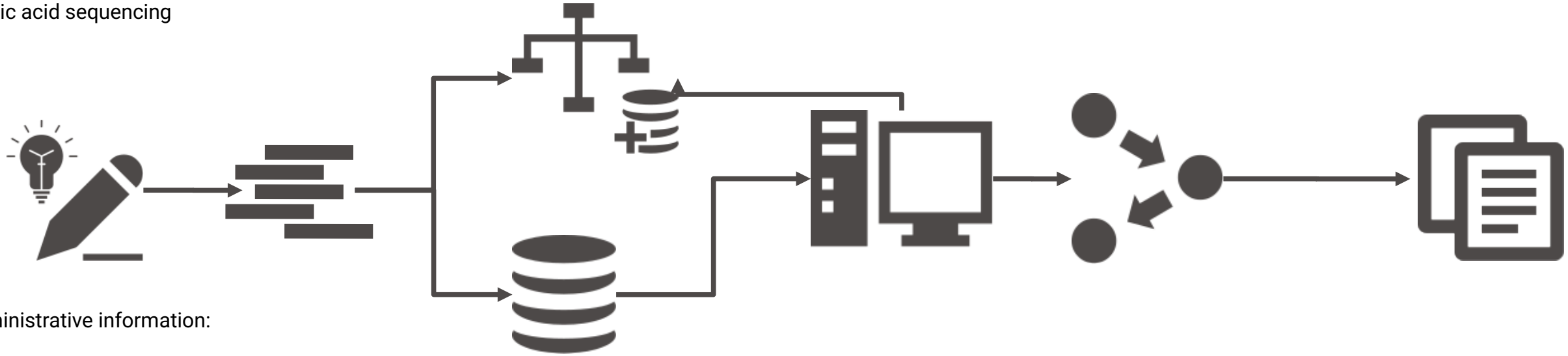
protocols:

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

ArrayExpress

protocols:

high throughput sequence alignment
normalization data transformation



Administrative information:

Persons
Organizations
Publications

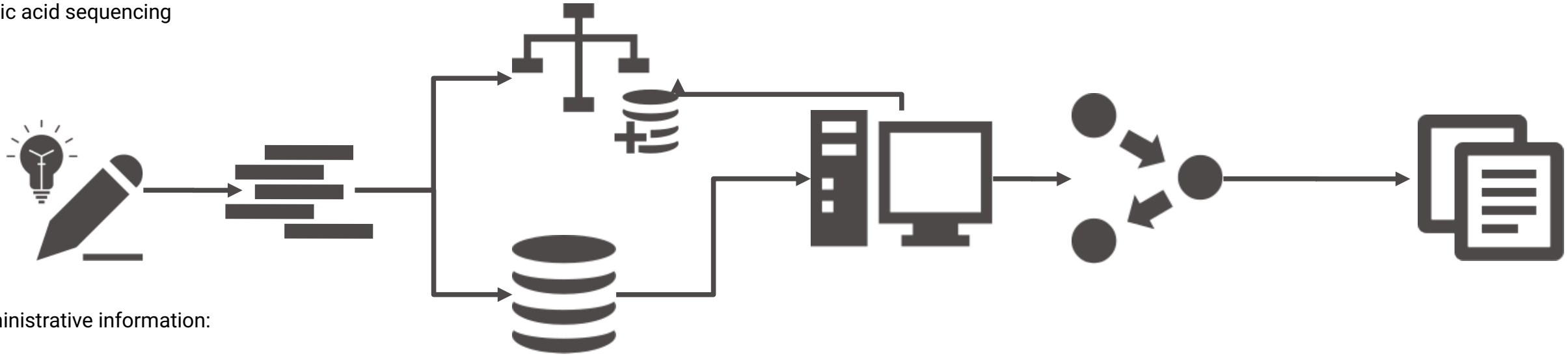
fastq
bam
csv/tsv

MINSEQE

Experimental conditions/design

protocols:

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



Administrative information:

Persons
Organizations
Publications

ArrayExpress

protocols:

high throughput sequence alignment
normalization data transformation

MINSEQE

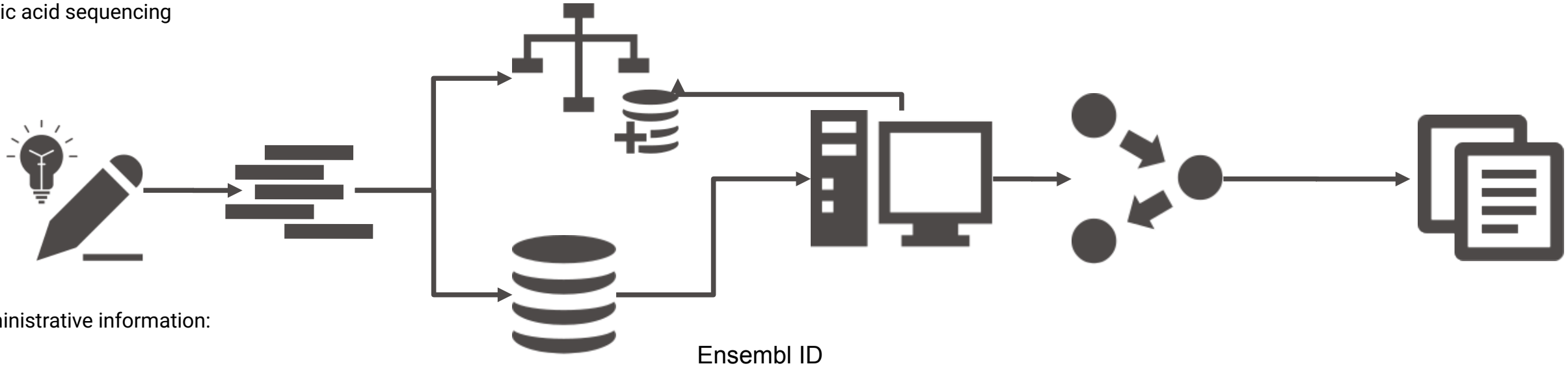
Experimental conditions/design

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

fastq
bam
csv/tsv

Ensembl ID
RefSeq ID

ArrayExpress

MINSEQE

Experimental conditions/design

protocols:

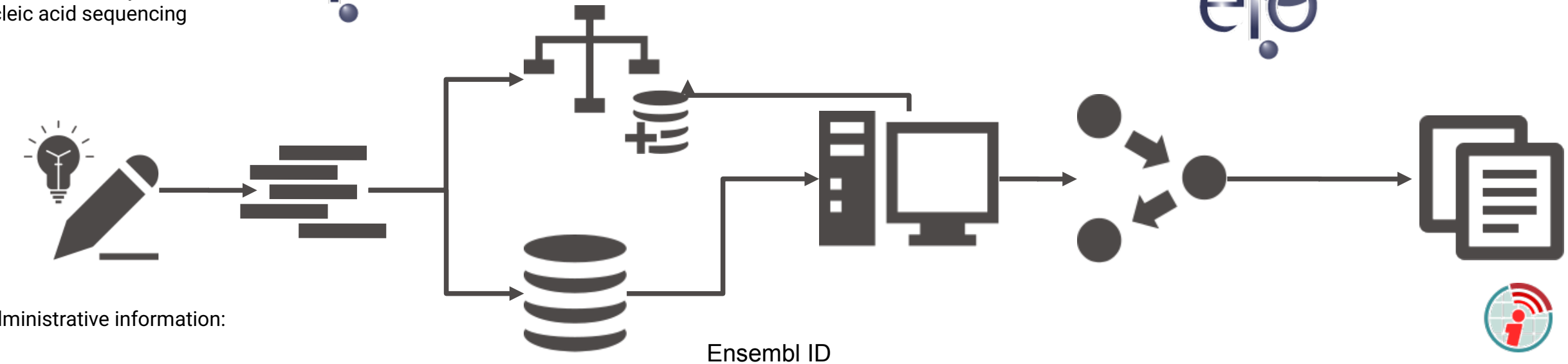
treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library construction
nucleic acid sequencing

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:

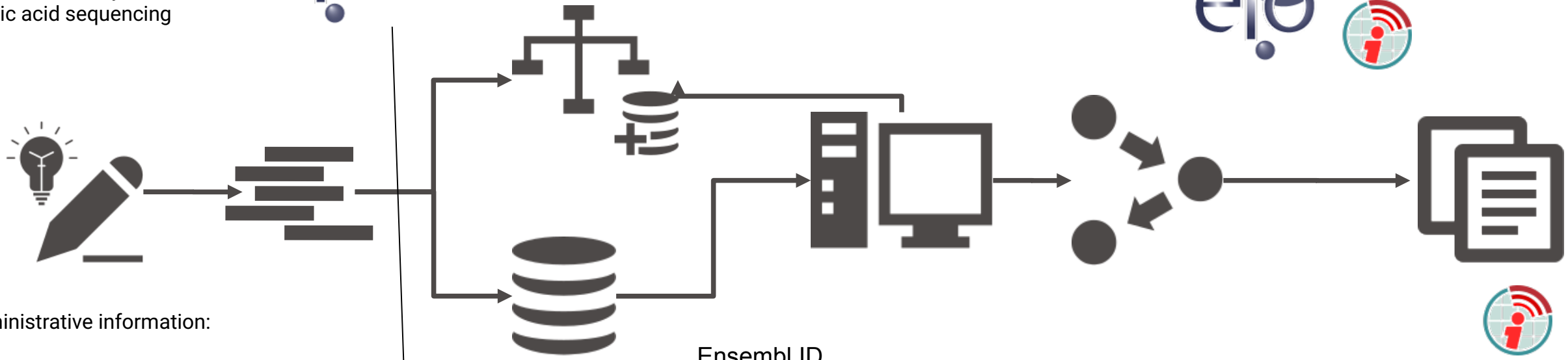
treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library construction
nucleic acid sequencing

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

Experimental conditions/design

protocols:

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

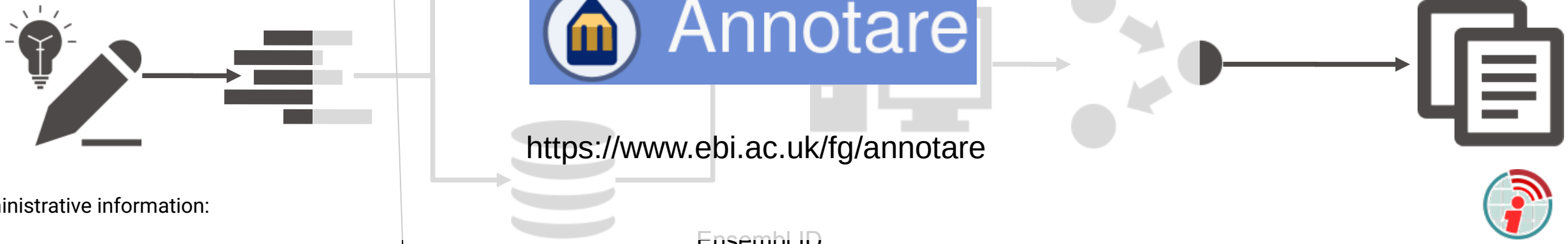
Taxonomy

MINSEQE

ArrayExpress

protocols:

high throughput sequence alignment
normalization data transformation



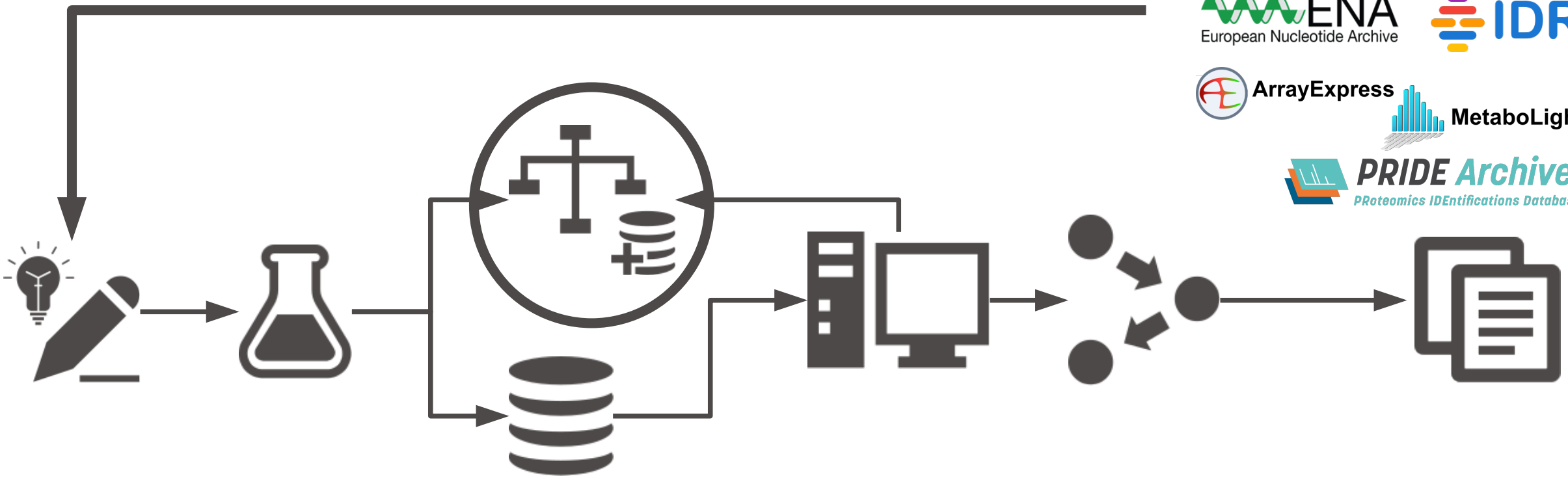
Administrative information:

Persons
Organizations
Publications

fastq
bam
csv/tsv

Ensembl ID
RefSeq ID

Interlinking with other resources



BioStudies.

ENA European Nucleotide Archive

IDR

ArrayExpress

MetaboLights

PRIDE Archive
PRoteomics IDEntifications Database

EUROPEAN GENOME-PHENOME ARCHIVE

Meta data standards

MINSEQE
MIAME
...

Meta data standards

MINSEQE
MIAME
...

HUPO-PSI TraML
MIAPE
...

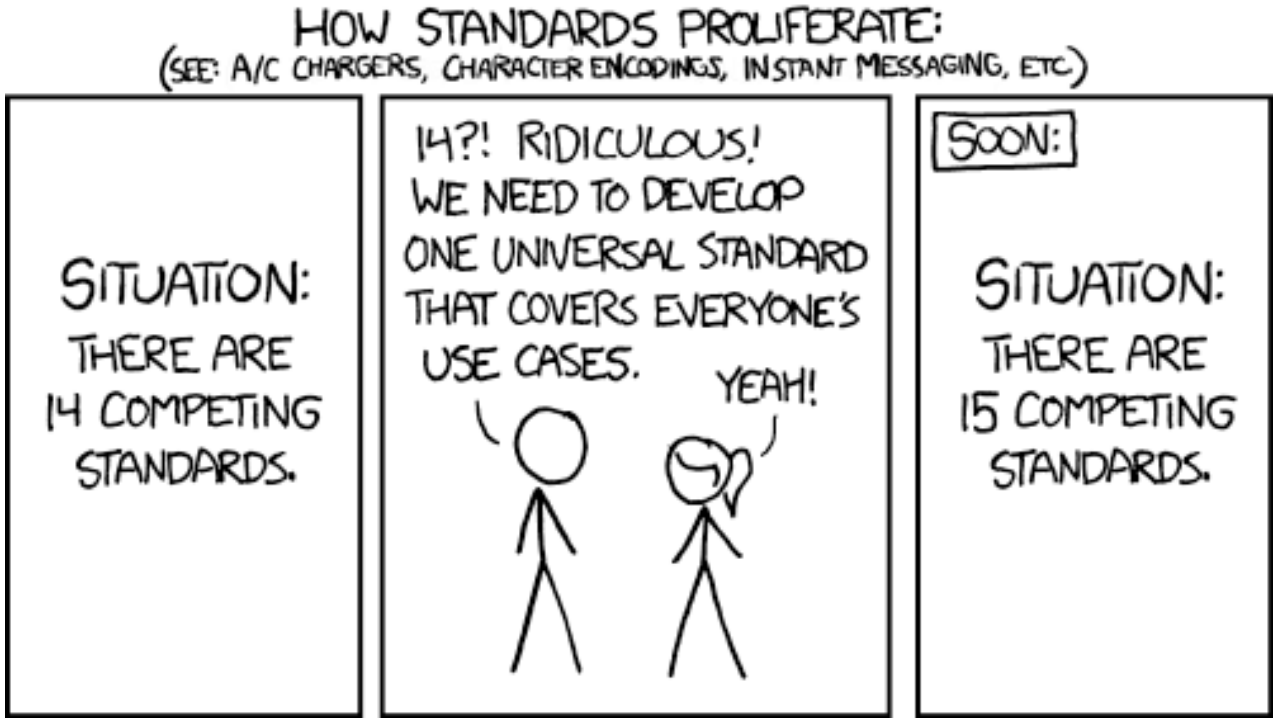
Meta data standards

MINSEQE
MIAME
...

SRA-XML

HUPO-PSI TraML
MIAPE
...

Community standards vs. formal ISO standards



https://xkcd.com/927/
https://elixir-europe.org/events/elixir-guest-webinar-two-universes-one-world-community-standards-vs-formal-iso-standards-life

Which metadatastandard?

Demo

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 ›

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
…

FASTA
FASTQ
…

mzML
mzQuantML
…

Metadata tracking platforms

Domain specific:

- COPO for plant sciences
- MOLGENIS for biobanking
- …

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

MOLGENIS

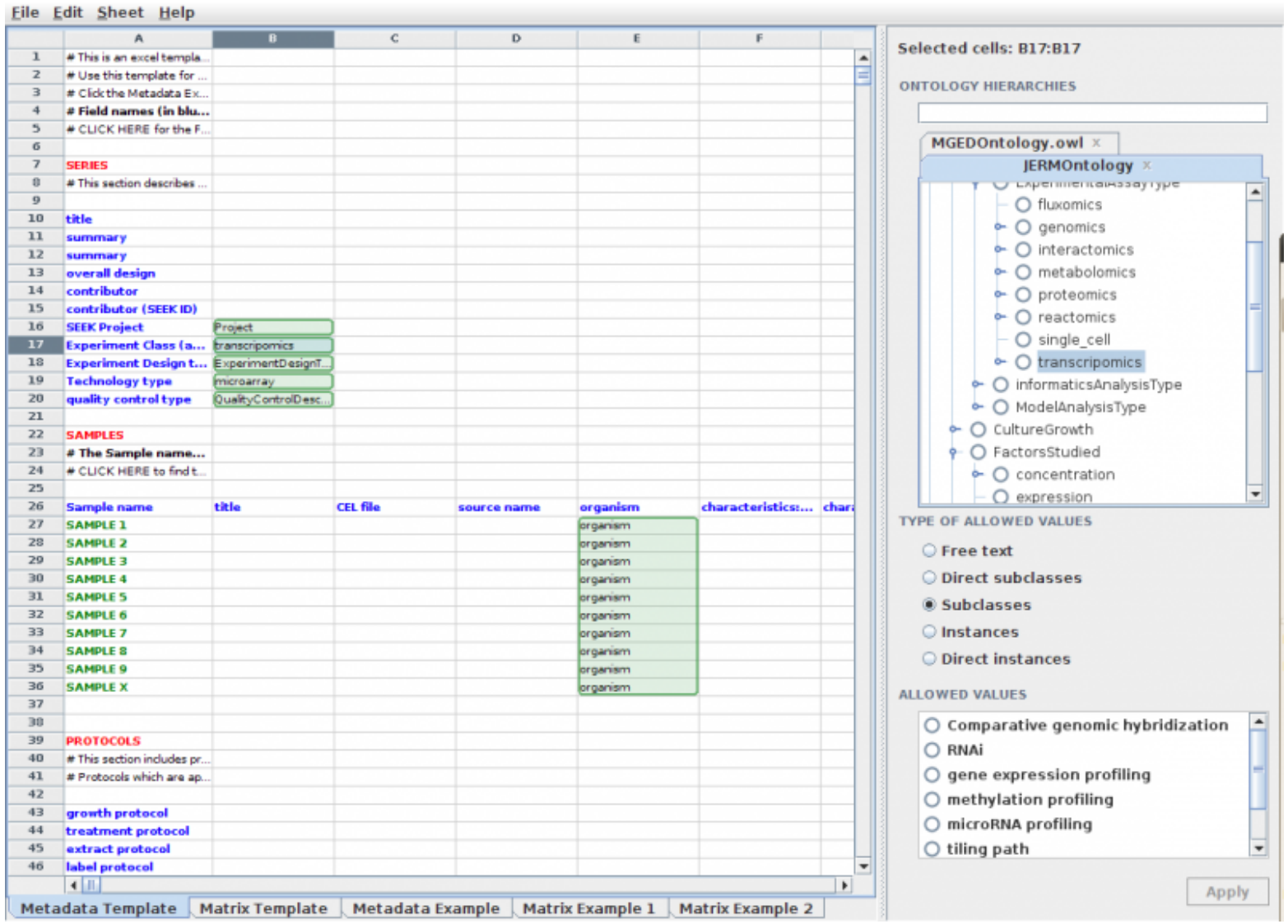
openBIS

SEEK

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

RightField

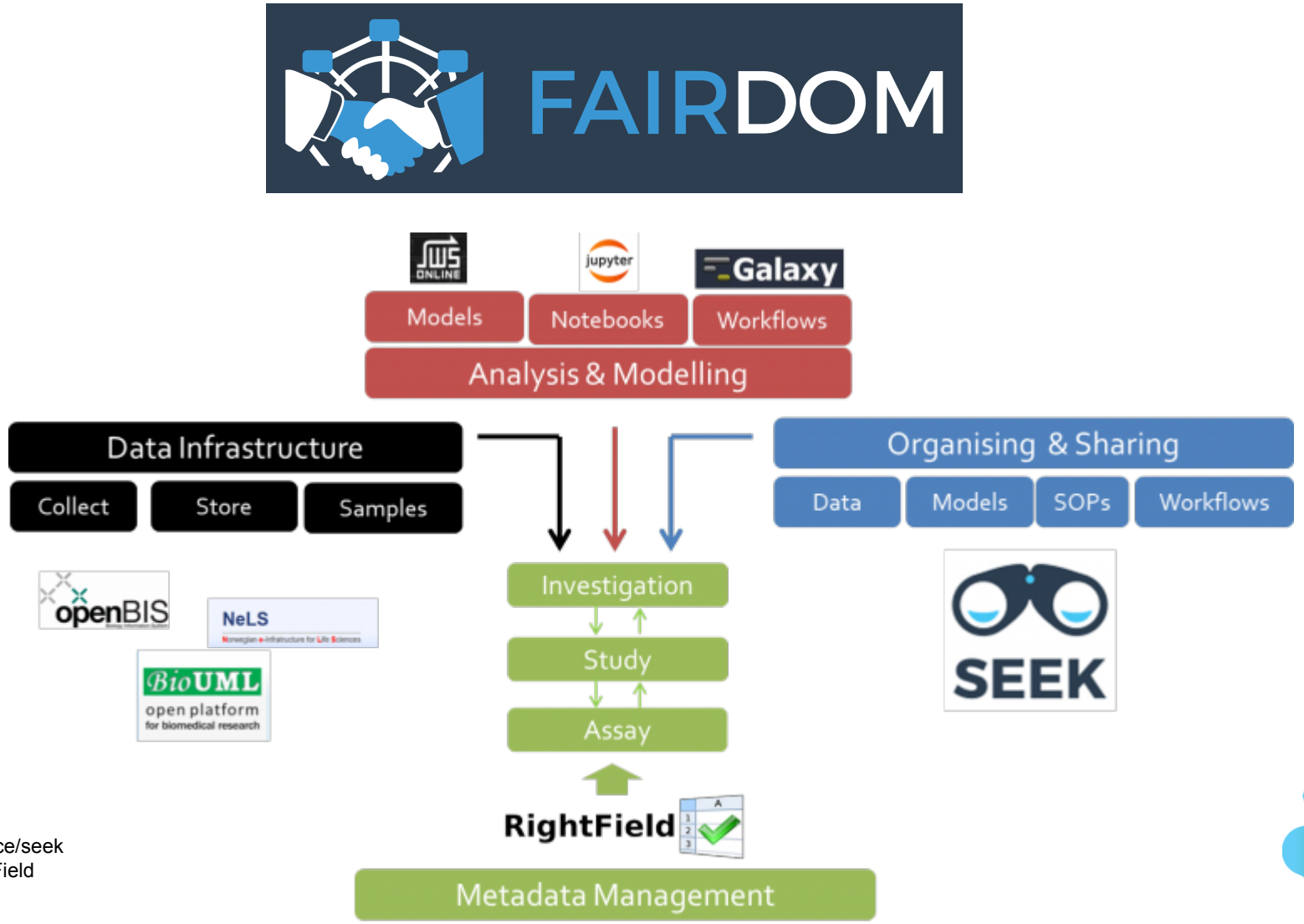
SEEK



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

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

FAIRDOM integration



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

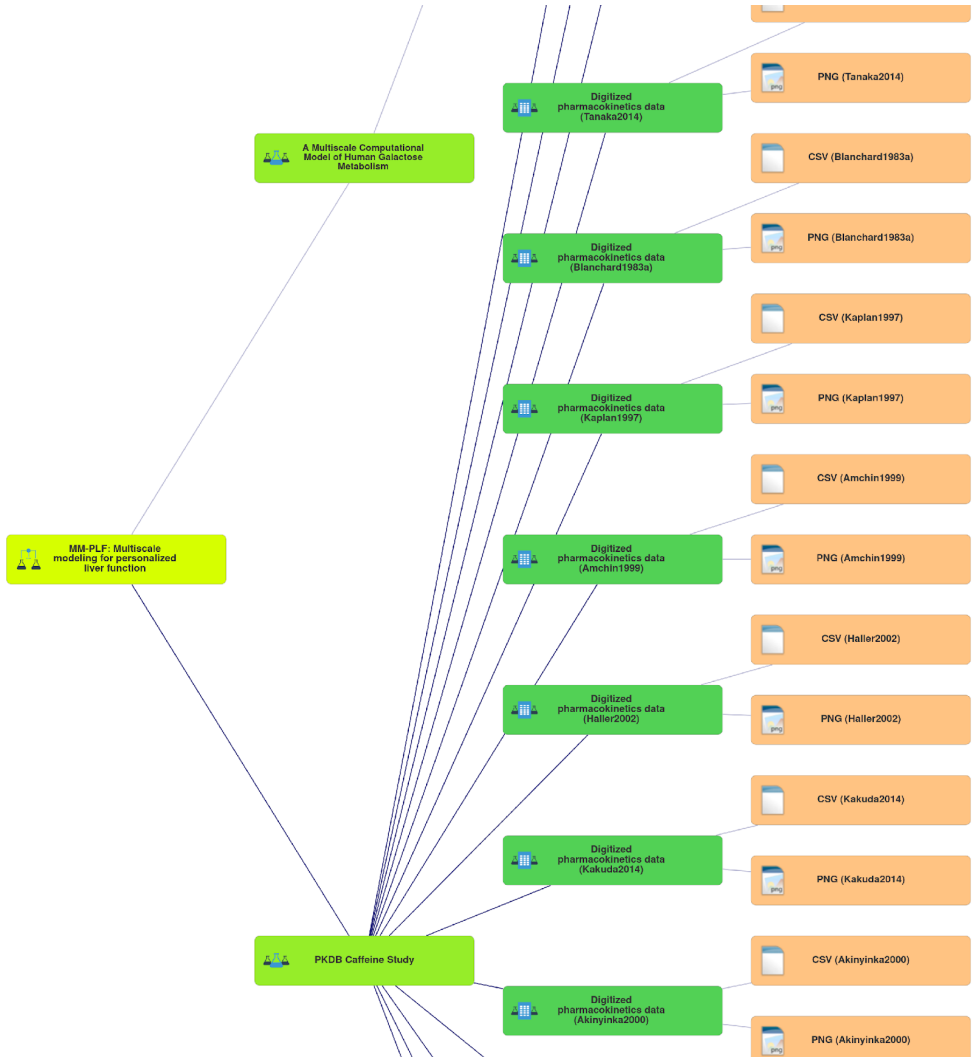
CENTRE FOR DIGITAL LIFE NORWAY

SEEK

Investigation

Study

Assay



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