Data Collection: Metadata Management

Speakers: Federico Bianchini (Centre for Bioinformatics UiO)
Moderator: Michal Torma / Rukaya Johaadein (UiO)

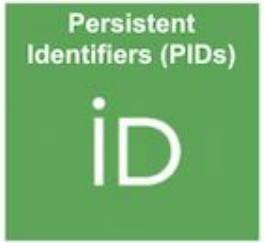
elixir NORWAY

BioMedData

14 June 2022

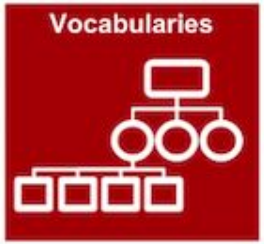
What is a metadata standard?

- 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

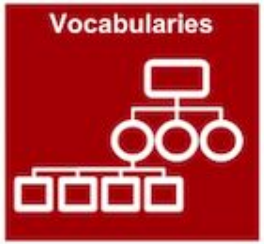


- A persistent identifier is required in the metadata
- Metadata standard compliant with repositories checklist
- Metadata fields used to identify/retrieve the data

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



- Interoperability fully rely 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



- Interoperability fully rely 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

How many ways can you say “female”?

18-day pregnant females
2 yr old female
400 yr. old female
adult female
asexual female

female (lactating)
female (pregnant)
female (outbred)
female parent
female plant

individual female
lgb*cc females
mare
female (worker)
monosex female

worker caste (female)
sex: female
female, other
female child
femal

How many ways can you say “female”?

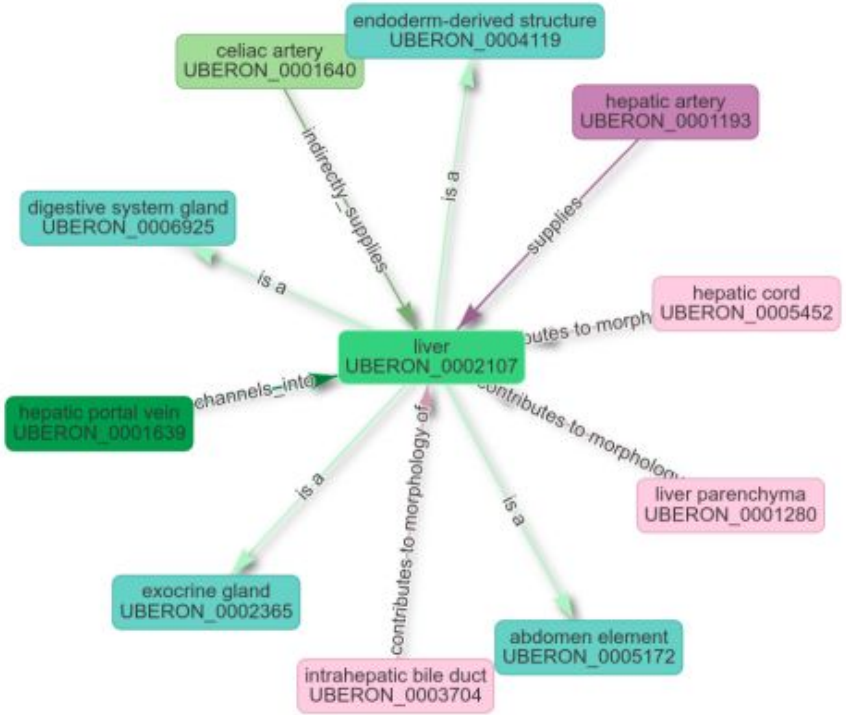
18-day pregnant females
2 yr old female
400 yr. old female
adult female
asexual female
castrate female
cf.female
cystocarpic female
dikaryon
dioecious female
diploid female
f
famale
femail
female
female - worker
female (alate sexual)
female (calf)
hen
female (lactating)
female (pregnant)
female (outbred)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

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		☑

- 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

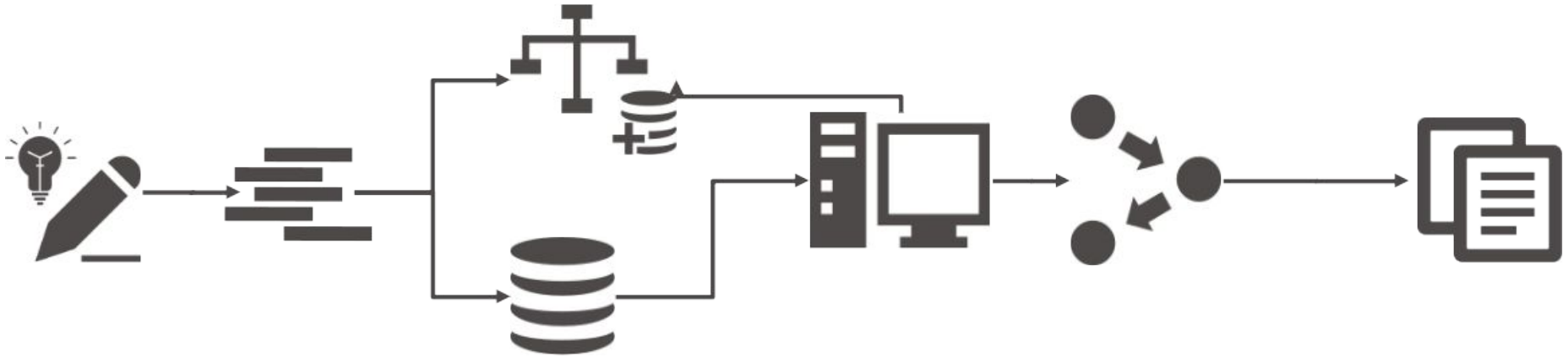
HUPO-PSI
TraML
MIAPE
...

ENA
European Nucleotide Archive

SRA-XML

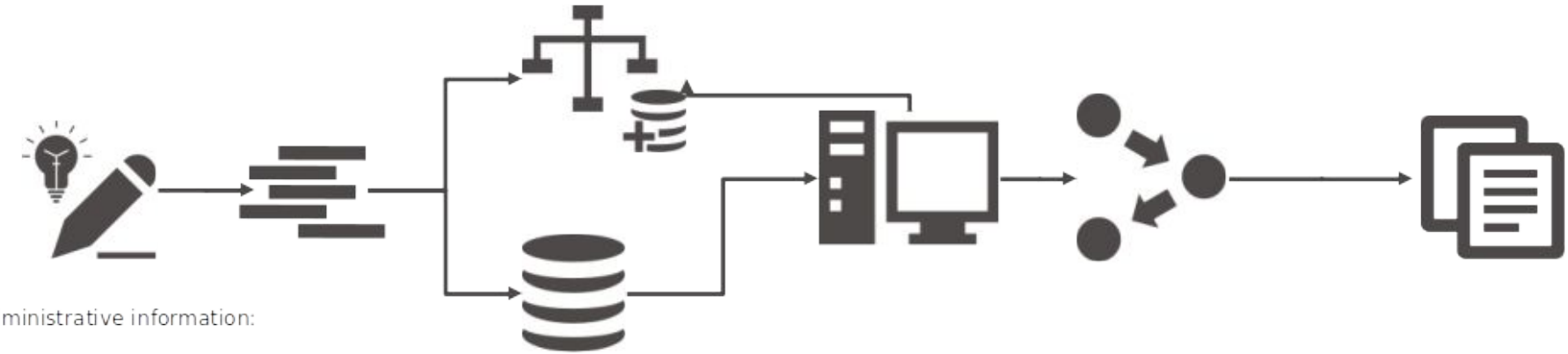
MINSEQE

ArrayExpress



MINSEQE

ArrayExpress



Administrative information:

Persons
Organizations
Publications

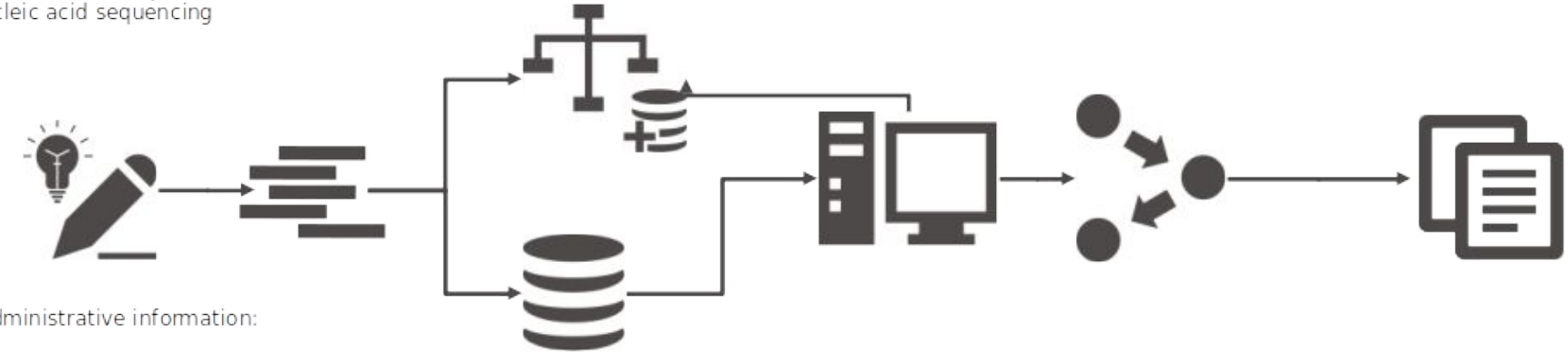
MINSEQE

ArrayExpress

Experimental conditions/design

protocols:

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



Administrative information:

Persons
Organizations
Publications

MINSEQE

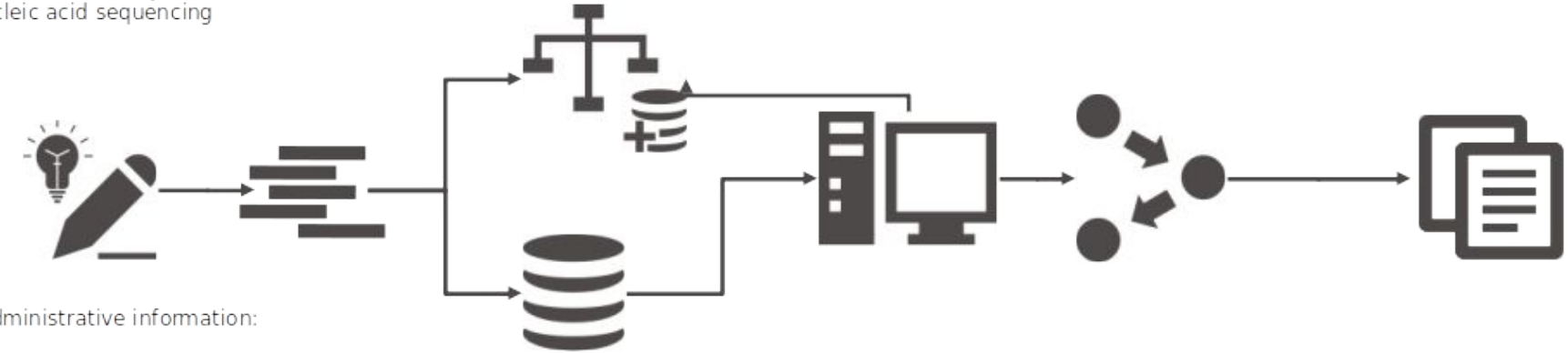
Experimental conditions/design

protocols:

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

protocols:

high throughput sequence alignment
normalization data transformation



Administrative information:

Persons
Organizations
Publications

fastq
bam
csv/tsv

ArrayExpress

MINSEQE

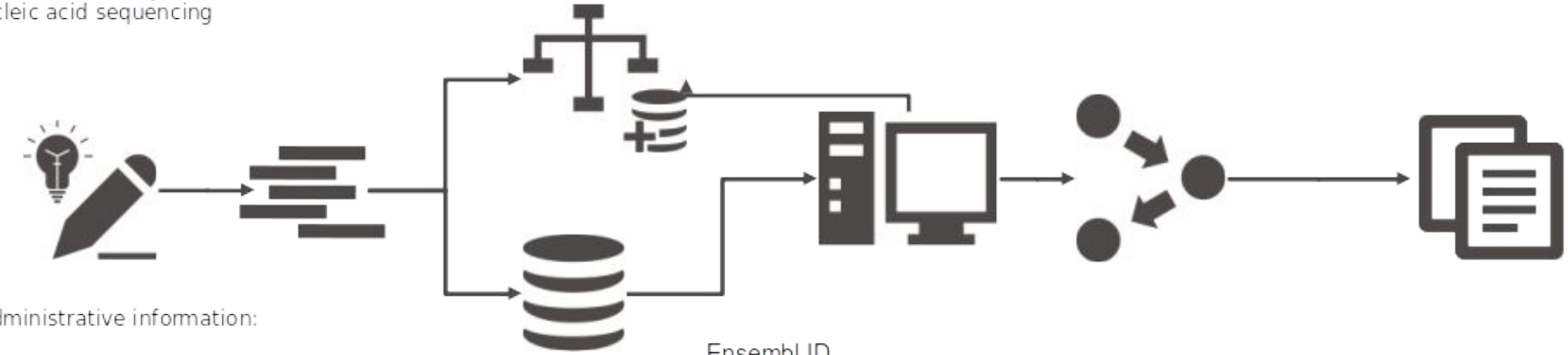
Experimental conditions/design

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

MINSEQE

Experimental conditions/design

protocols:

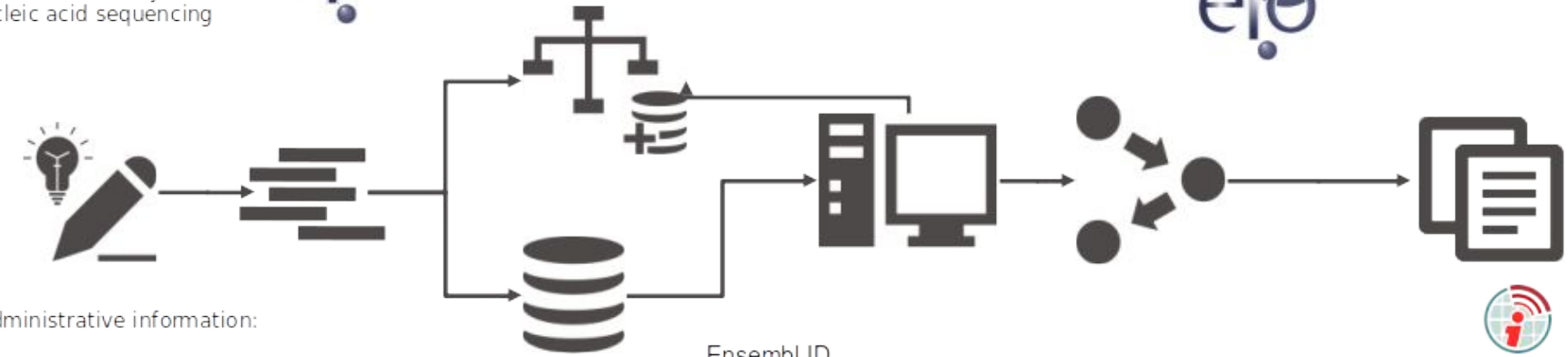
treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library constructi
nucleic acid sequencing

Taxonomy

protocols:

high throughput sequence alignment
normalization data transformation

ArrayExpress

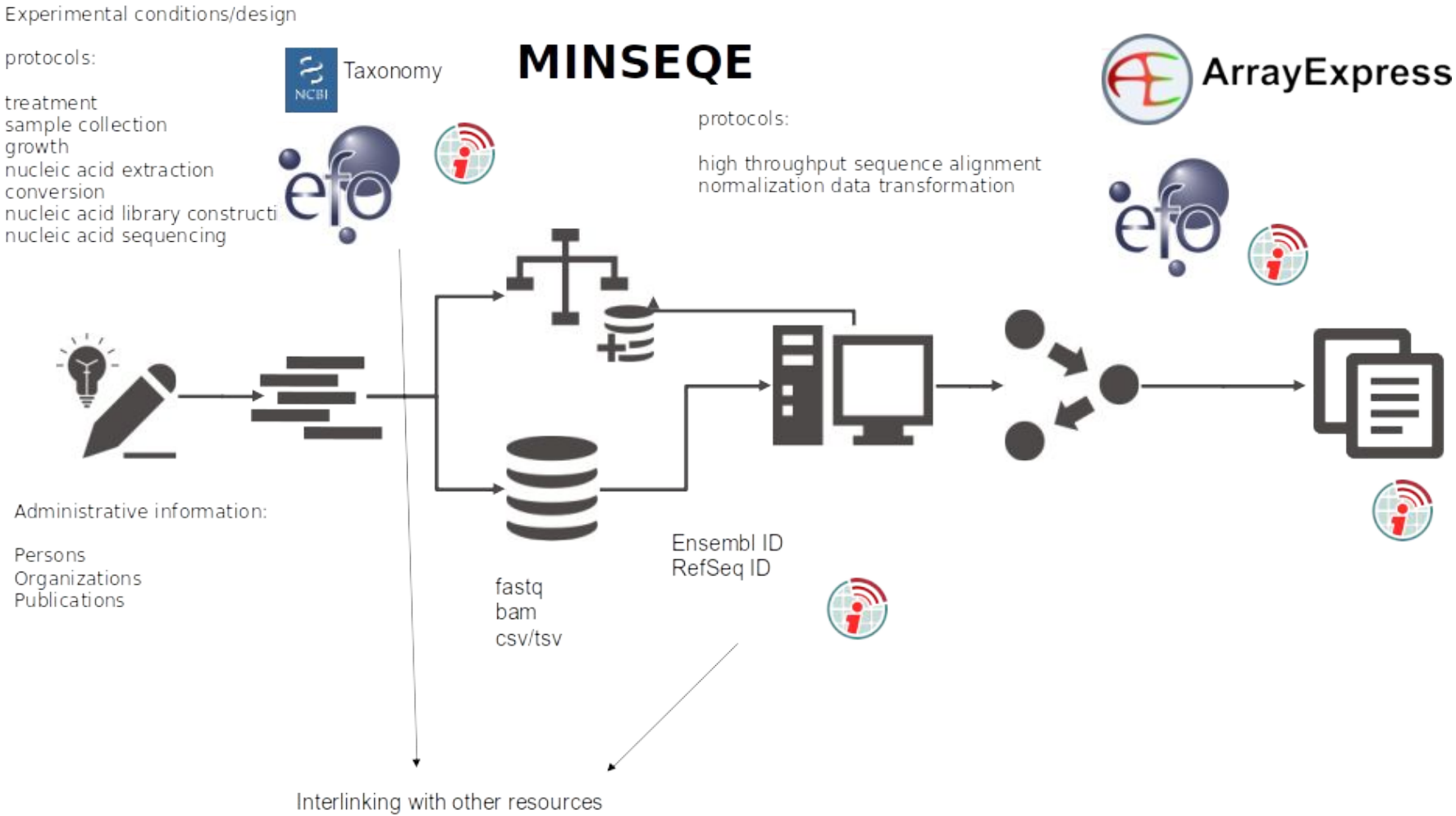


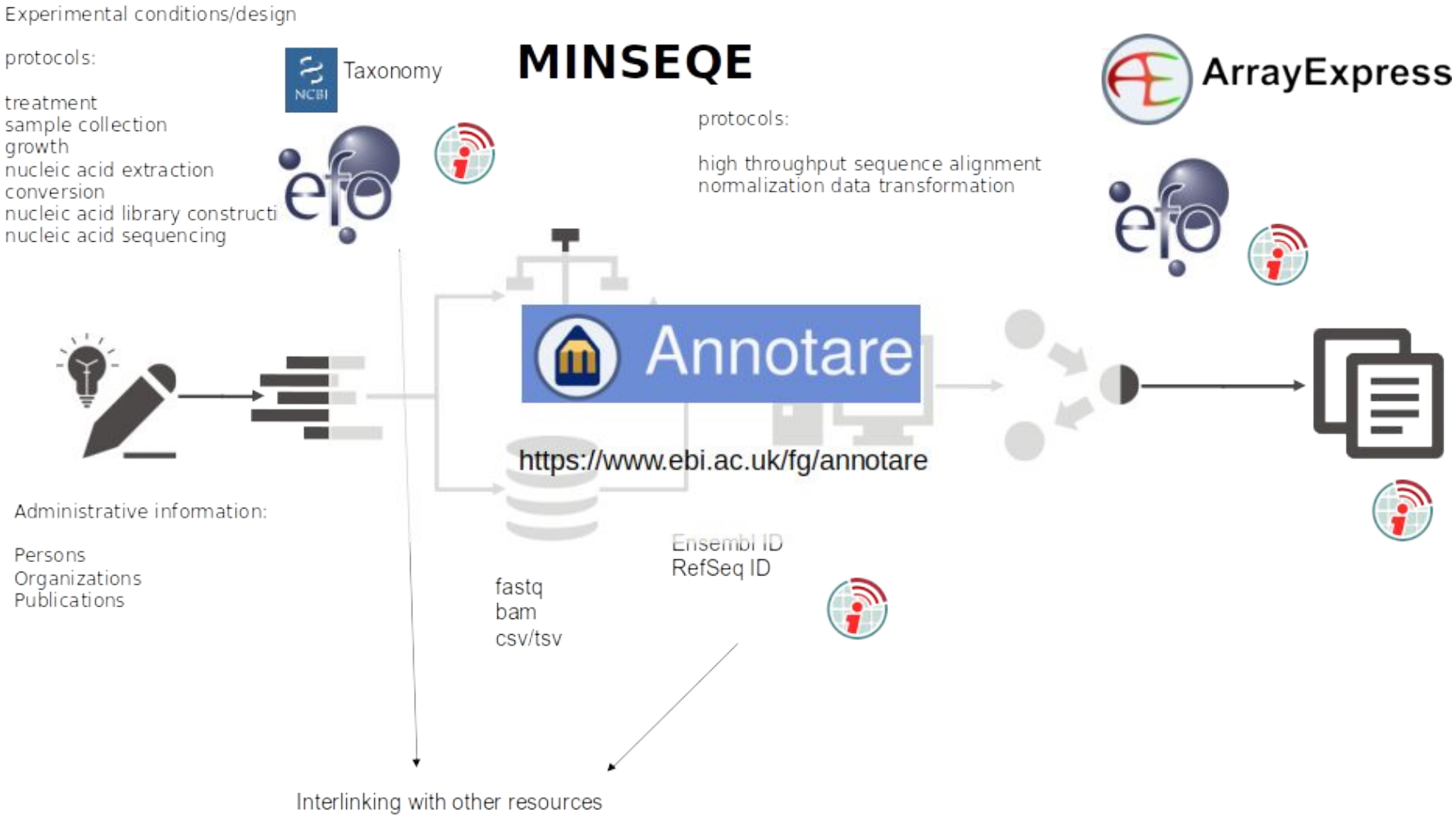
Administrative information:

Persons
Organizations
Publications

fastq
bam
csv/tsv

Ensembl ID
RefSeq ID


Experimental conditions/design
protocols:
treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library constructi
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

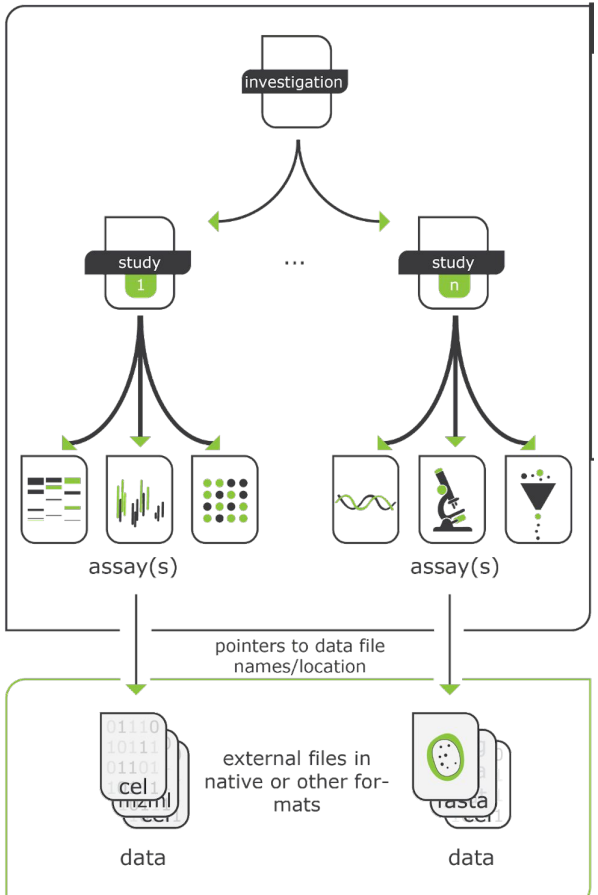

Experimental conditions/design
protocols:
treatment
sample collection
growth
nucleic acid extraction
conversion
nucleic acid library constructi
nucleic acid sequencing
Taxonomy
MINSEQE
ArrayExpress
protocols:
high throughput sequence alignment
normalization data transformation
Annotare
https://www.ebi.ac.uk/fg/annotare
Administrative information:
Persons
Organizations
Publications
fastq
bam
csv/tsv
Ensembl ID
RefSeq ID
Interlinking with other resources

Which metadata standard?

Demo

elixir
Recommended
Interoperability
Resources

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

Metadata tracking platforms

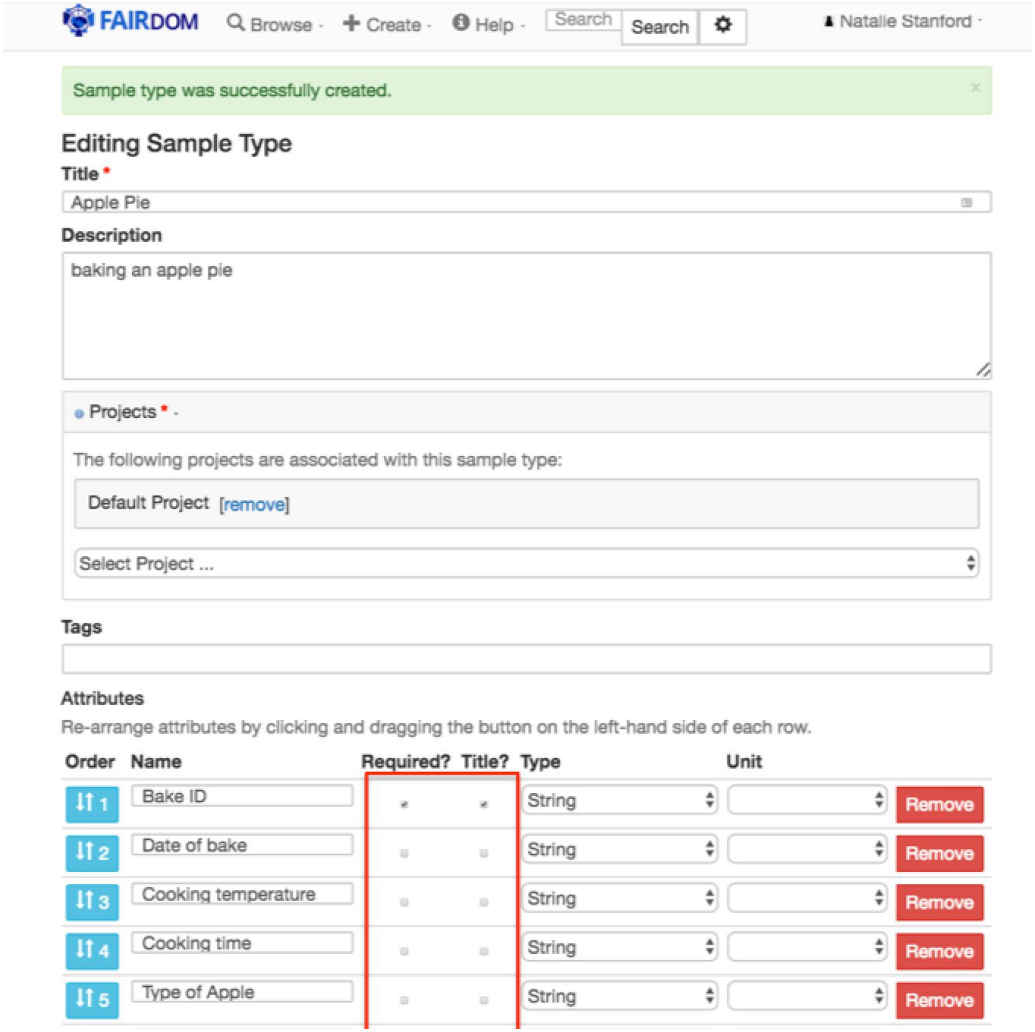
Domain specific:

- 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

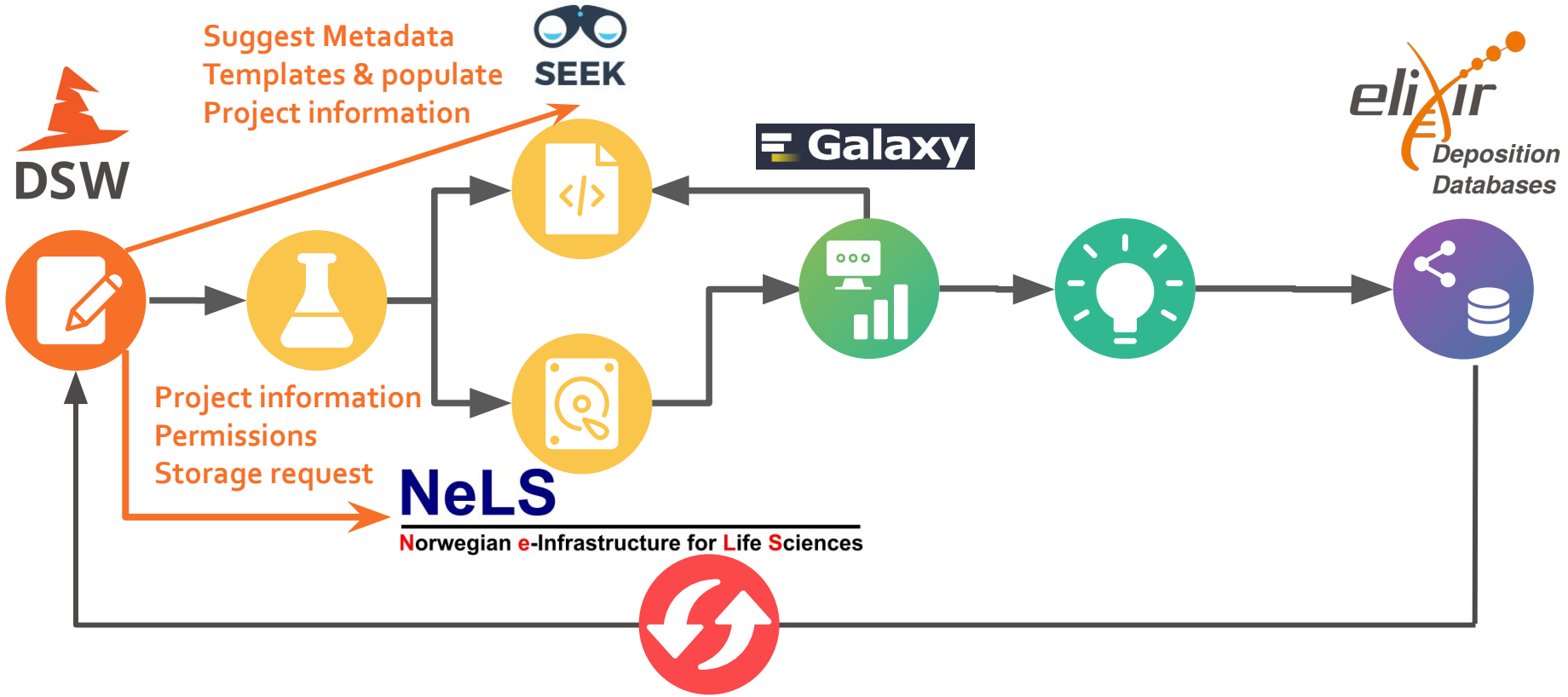
- Target new users (via ELIXIR)

nor seq

NAPI

Plans for further integration

K Bosl “NeLS (Norwegian e-Infrastructure for Life Sciences) tool assembly” originally presented at ELIXIR All Hands meeting 2022



Thank you!

elixir-norway.org

@elixirnorway

contact@bioinfo.no