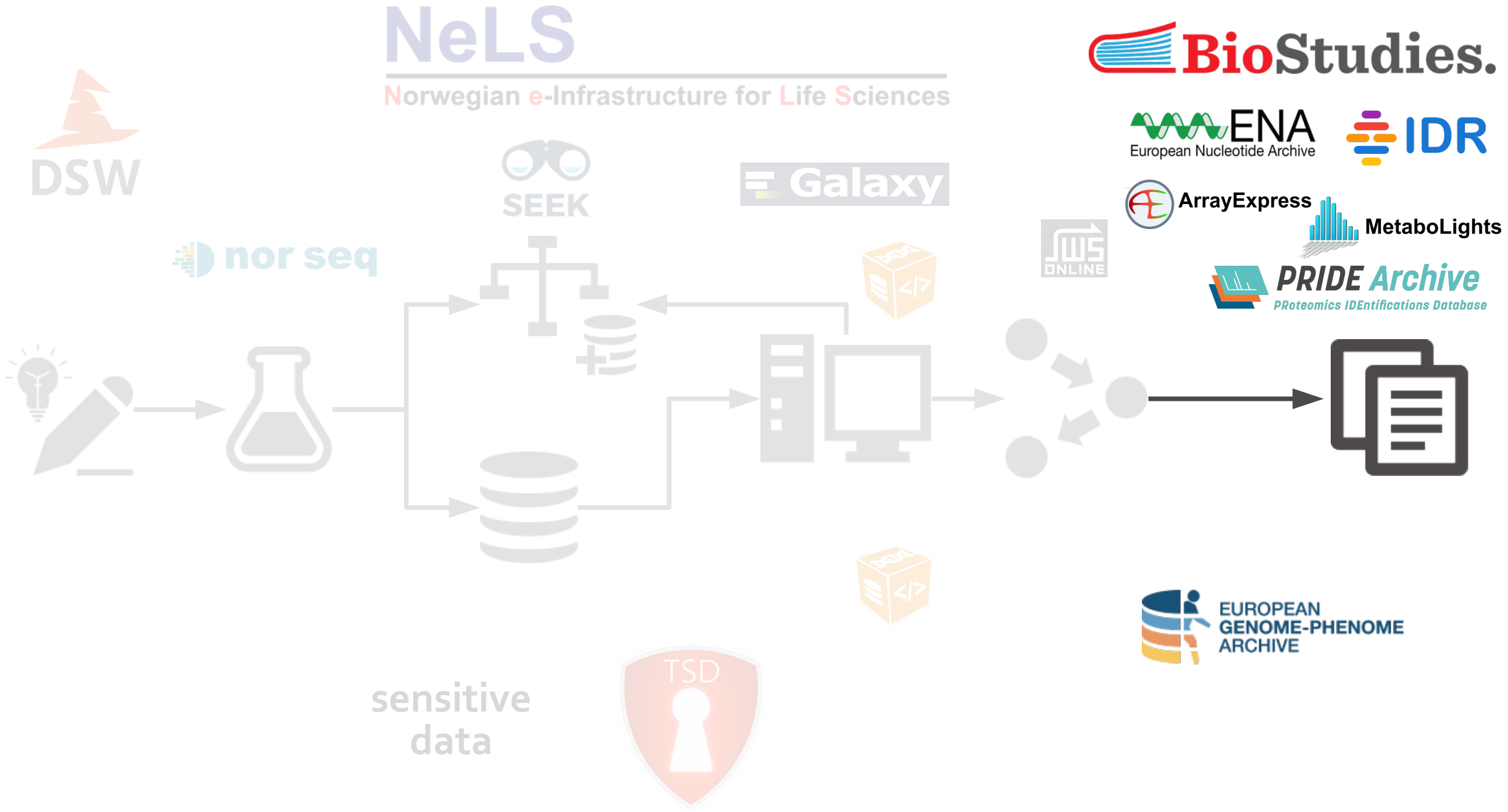
Archiving, preserving, sharing

Korbinian Bösl
Data management coordinator
ELIXIR Norway/Digital Life Norway
10th December 2020


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

Why should I deposit my data?

- Increased Visibility (SEO) → 25% more citations

doi.org/10.1371/journal.pone.0230416

Why should I deposit my data?

- Increased Visibility (SEO) → 25% more citations
- Value added Databases → more citations

doi.org/10.1371/journal.pone.0230416

Why should I deposit my data?

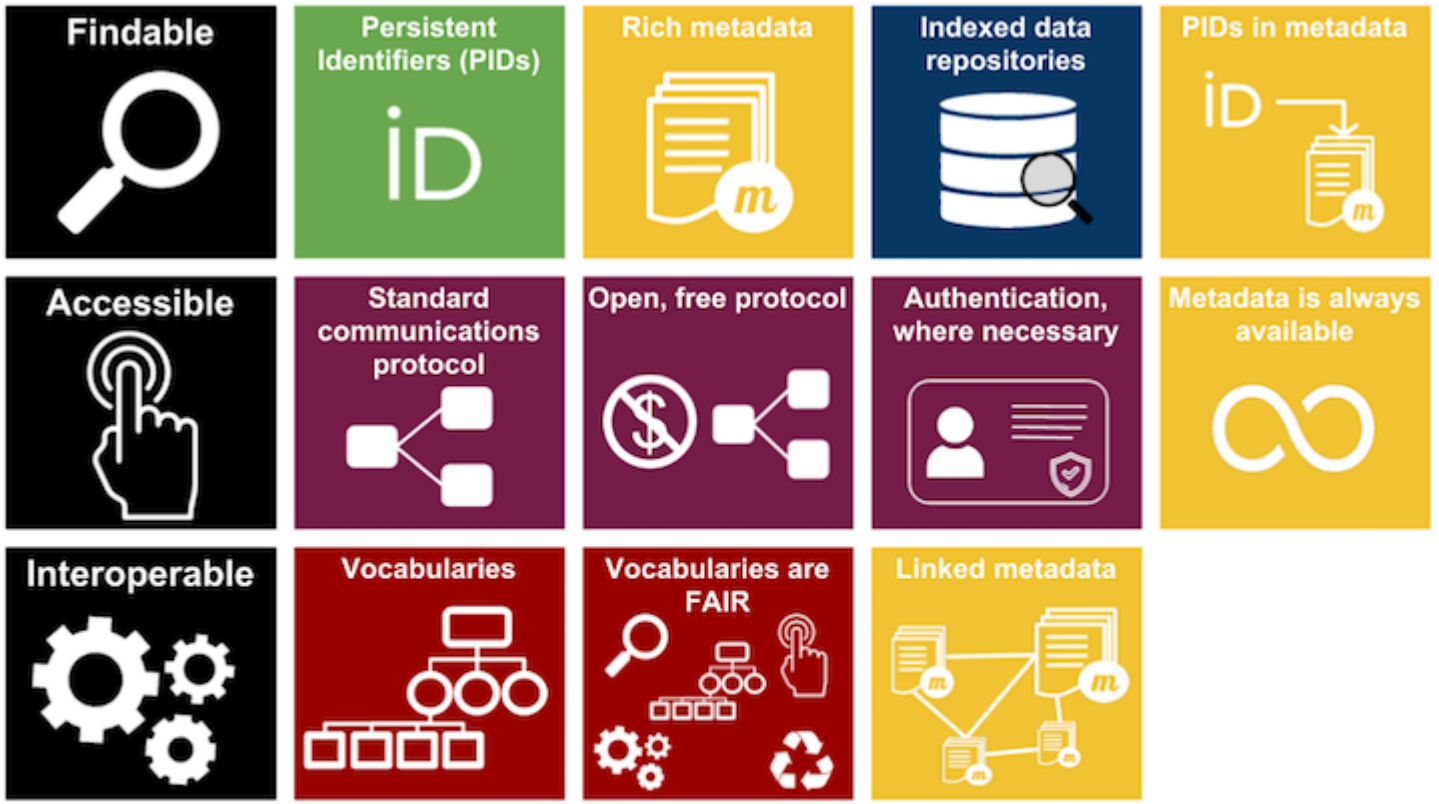
- Increased Visibility (SEO) → 25% more citations
- Value added Databases → more citations
- FAIRification – Funding requirements ✔✔

doi.org/10.1371/journal.pone.0230416

Why should I deposit my data?

- Increased Visibility (SEO) → 25% more citations
- Value added Databases → more citations
- FAIRification – Funding requirements ✔✔
- $ Safe money on storage

doi.org/10.1371/journal.pone.0230416



What is a PID and why do we need it?

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

Identifiers.org<

Archival Resource Keys (ARK)

Perisitant Uniform Resource Locator (PURL)

Universal Resource Name (URN)

PIDs exits also for

- Persons ORCID

PIDs exits also for

- Persons ORCID
- $ Funding bodies Crossref

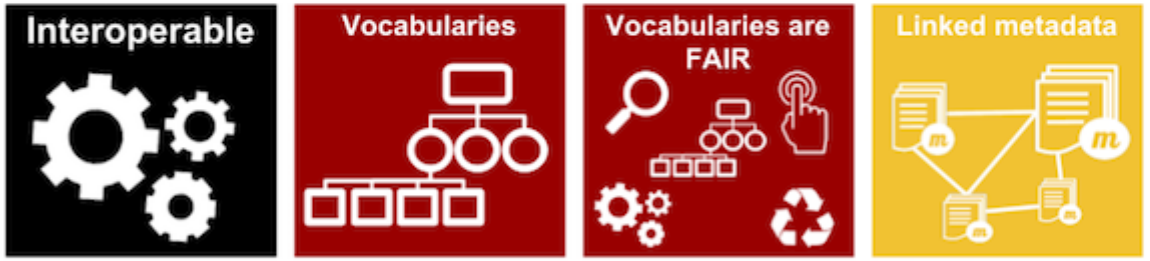
PIDs exits also for

- Persons ORCID
- $ Funding bodies Crossref
- Institutions

PIDs exits also for

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

Why do we need standard vocabularies?



How many way can you say “female”?

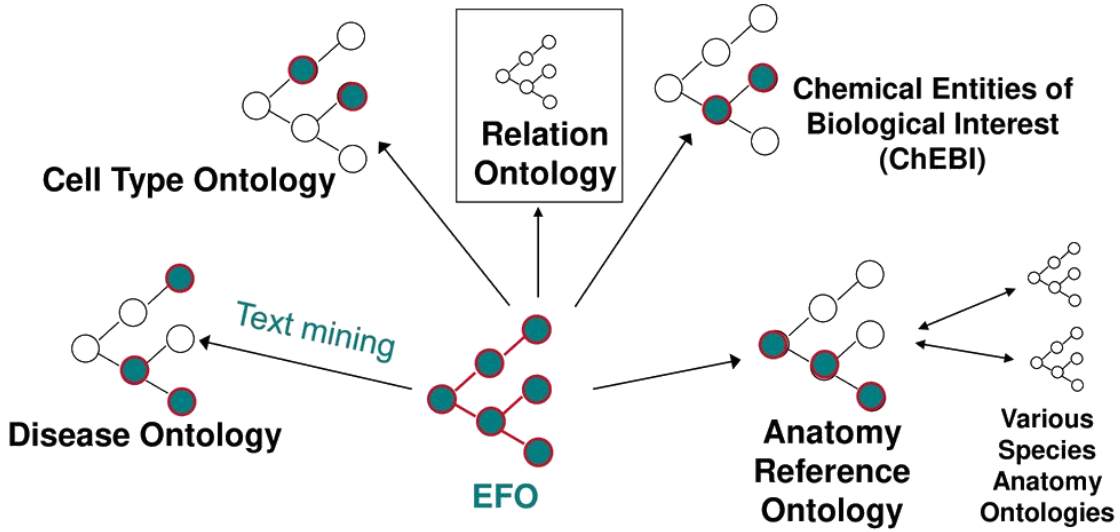
How many way can you say “female”?

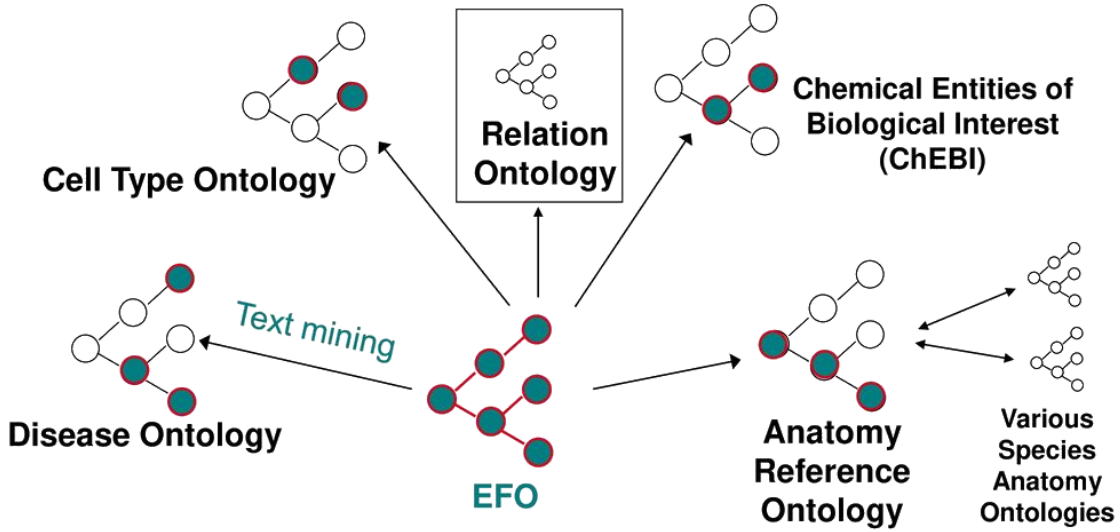
18-day pregnant females	female (lactating)	individual female	worker caste (female)
2 yr old female	female (pregnant)	lgb*cc females	sex: female
400 yr. old female	female (outbred)	mare	female, other
adult female	female parent	female (worker)	female child
asexual female	female plant	monosex female	femal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



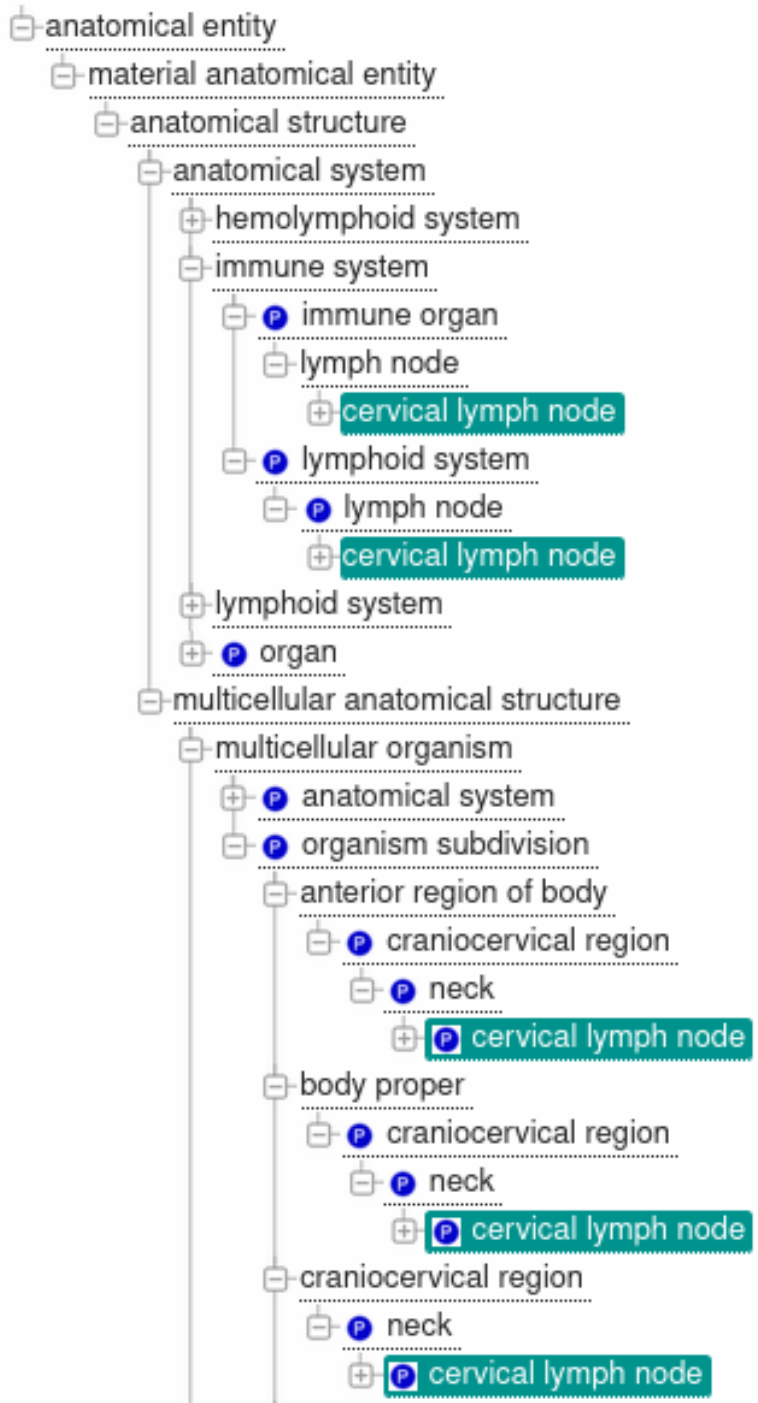
- experimental factor
  - material property
    - quality
      - sex
        - female

database cross reference

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

Uberon

Onthologies
enable hierarchical searches



Meta data standards

MINSEQE

Meta data standards

MINSEQE

Organism

Taxonomy

Experimental factors

Tissue

Sequencing data - e.g. FASTQ

Processed data

Linked publications, contact data, general information

sample - data relationships

Experimental + data processing protocols

Controlled vocabulary & Ontologies

Metadata standards – controlled vocabulary for

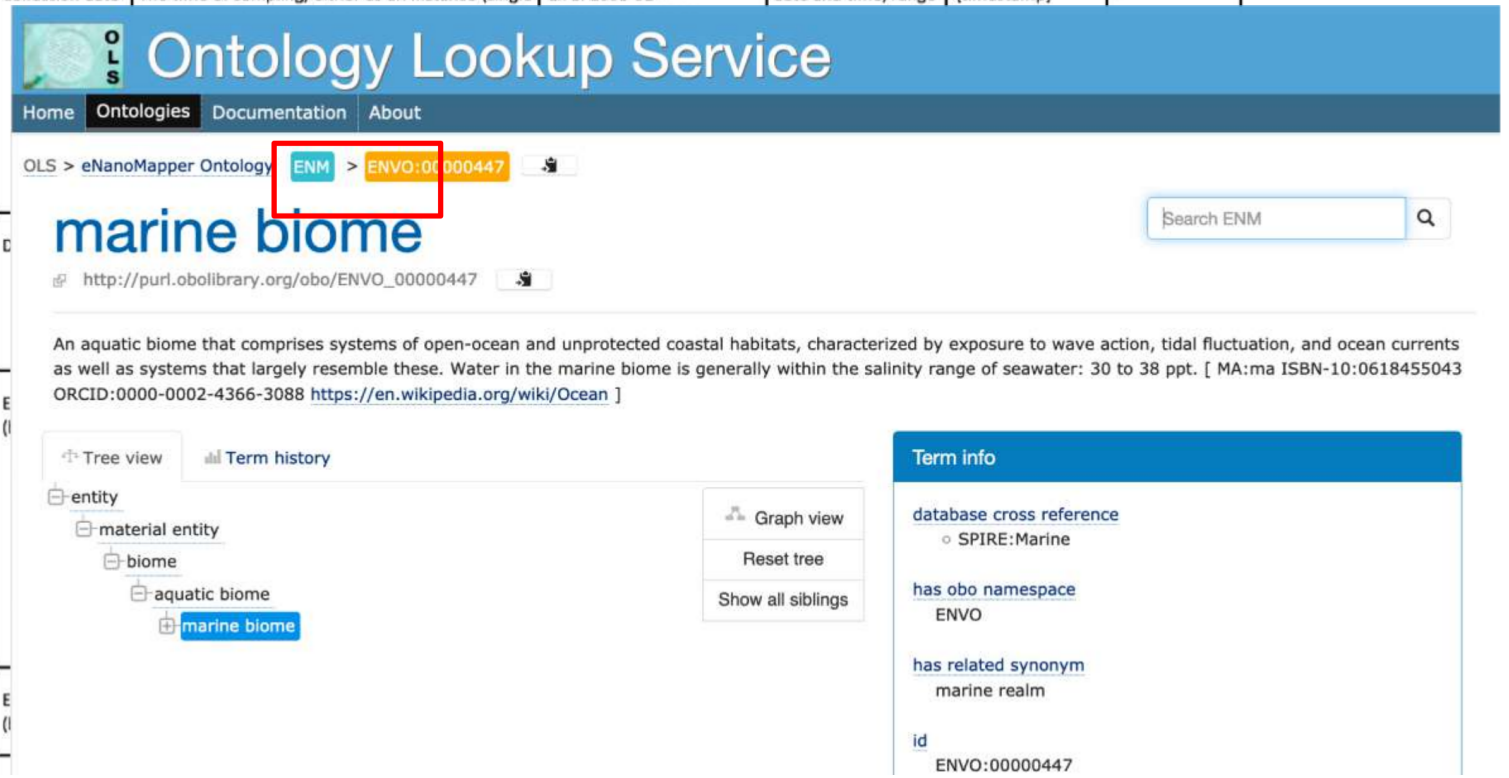
Structured comment name	Item	Description	Examples	Expected value	Value syntax	Preferred units / suffix
alt_elev	Geographic location (altitude/elevation)	Sample taken at given elevation above sea level, defined in meters(m) as a positive floating number with two decimals.	Ex 1: 3.06 Ex 2: 1.80-2.15	-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

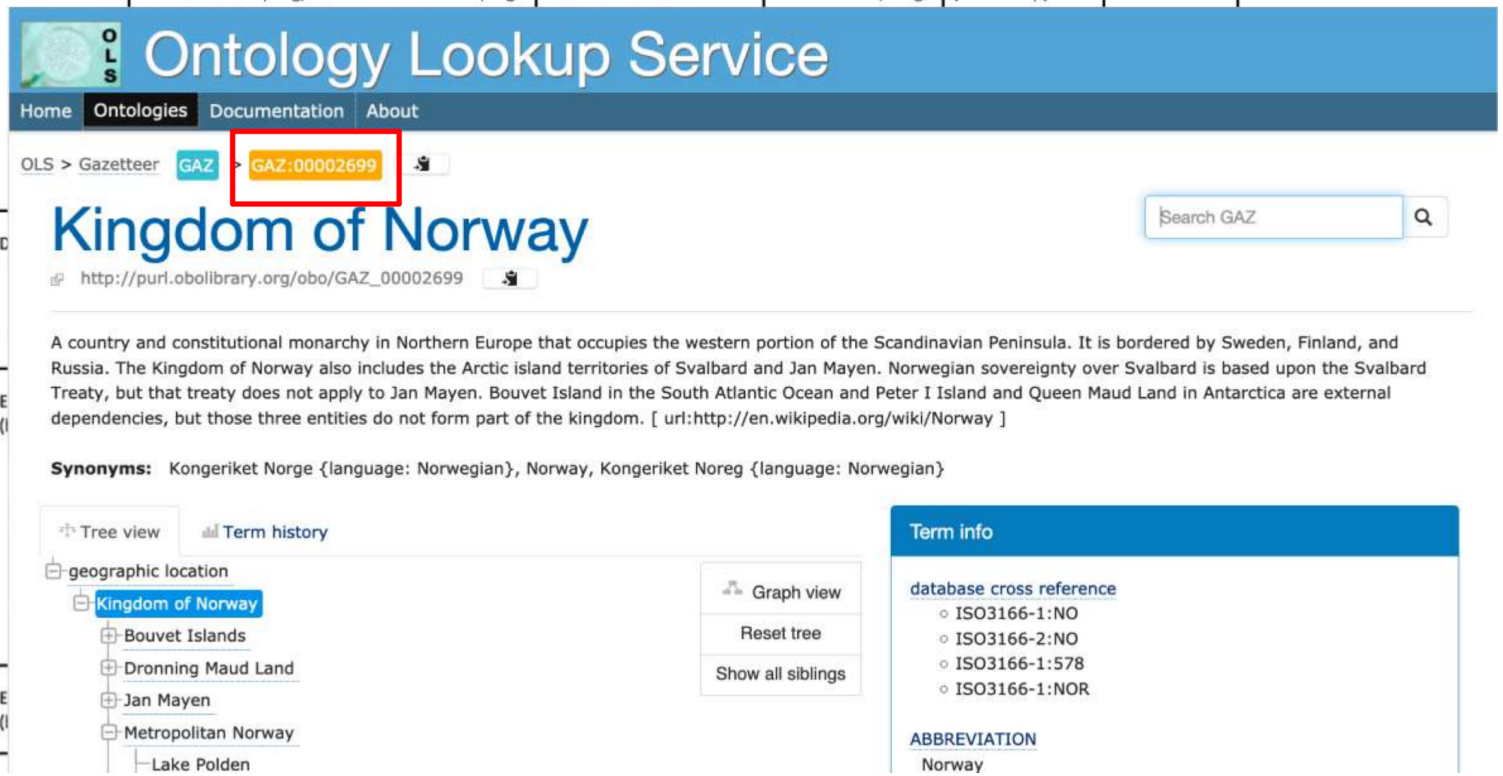


The ENVO ontology describes the environment of the sampling

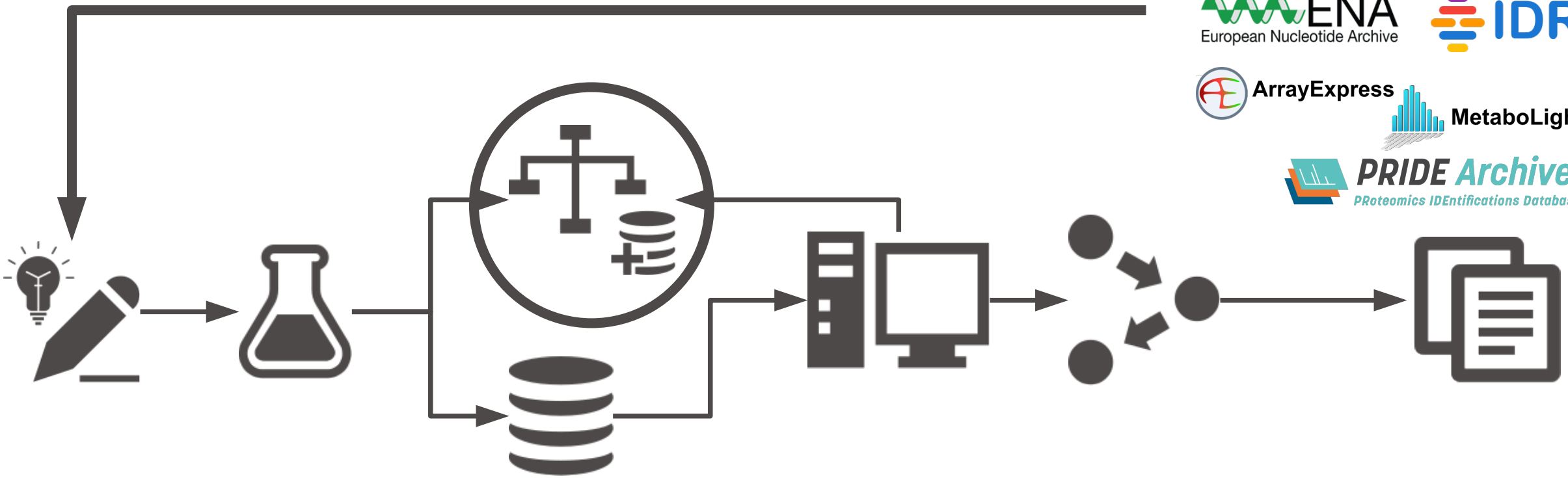
Controlled vocabulary & Ontologies

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



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
...

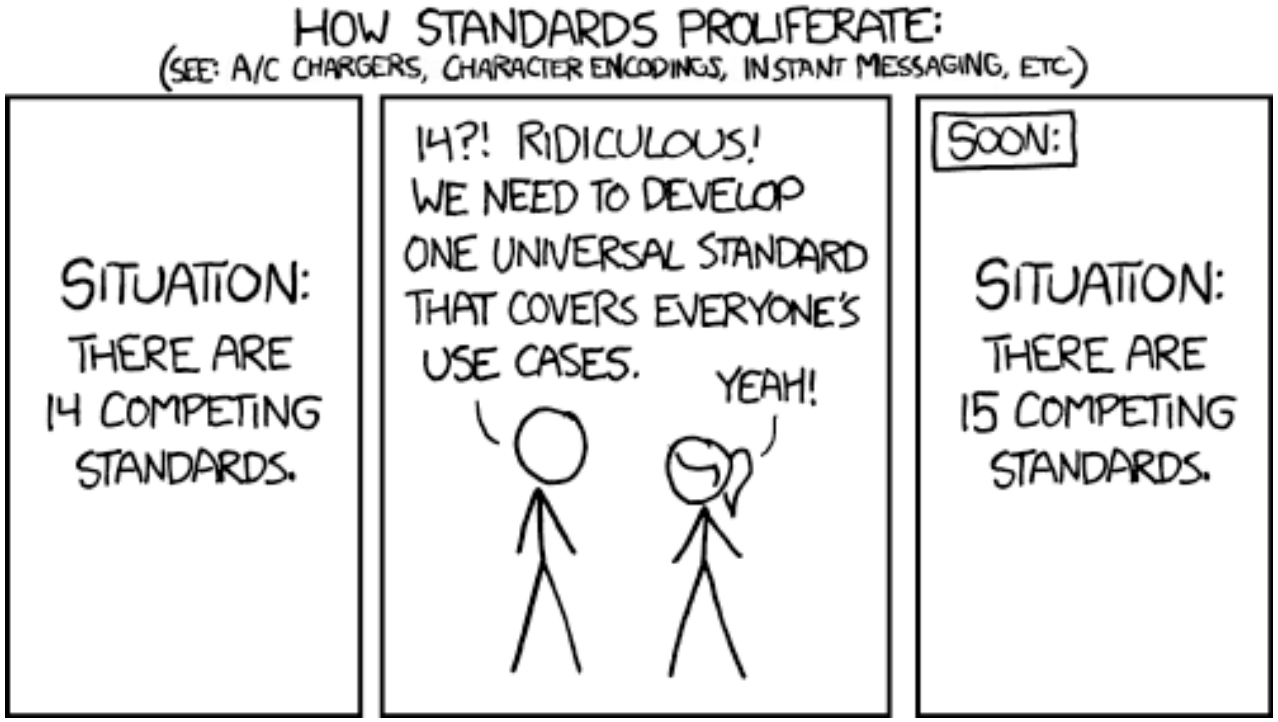
Meta data standards

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?

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
…

Where should I deposit my data?

Preferably somewhere
where most things are taken care of?

How to choose?

molecular biology
1637 databases

The 27th annual Nucleic Acids Research database issue and molecular biology database collection. Nucleic Acids Res 48, D1–D8 (2020).

eLife

ELSEVIER

F1000 Research

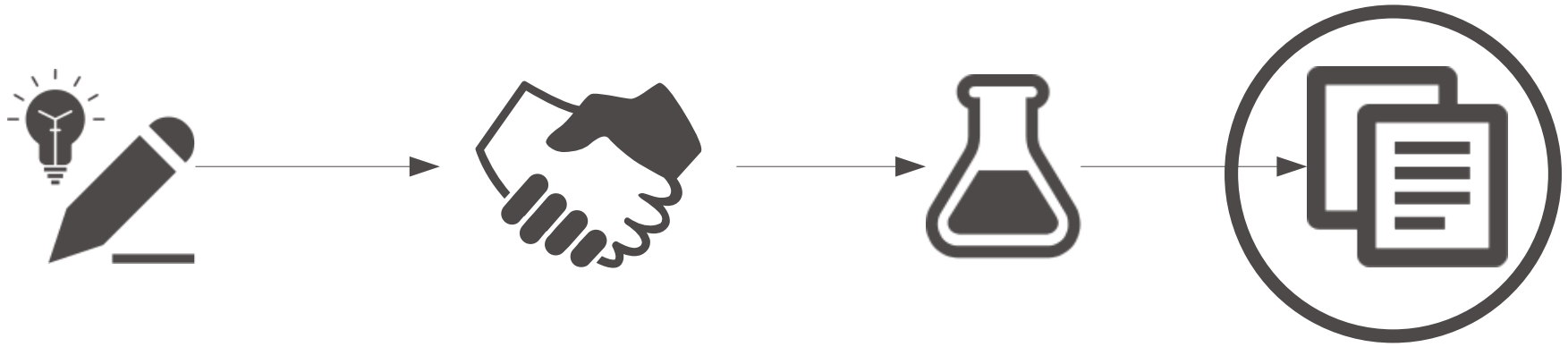
Taylor & Francis
Taylor & Francis Group

nature

EMBOpress

PLOS

SCIENTIFIC DATA



Deposition

Where should I deposit my data?

fairsharing.org
re3data.org

Where should I deposit my data?

Fits?

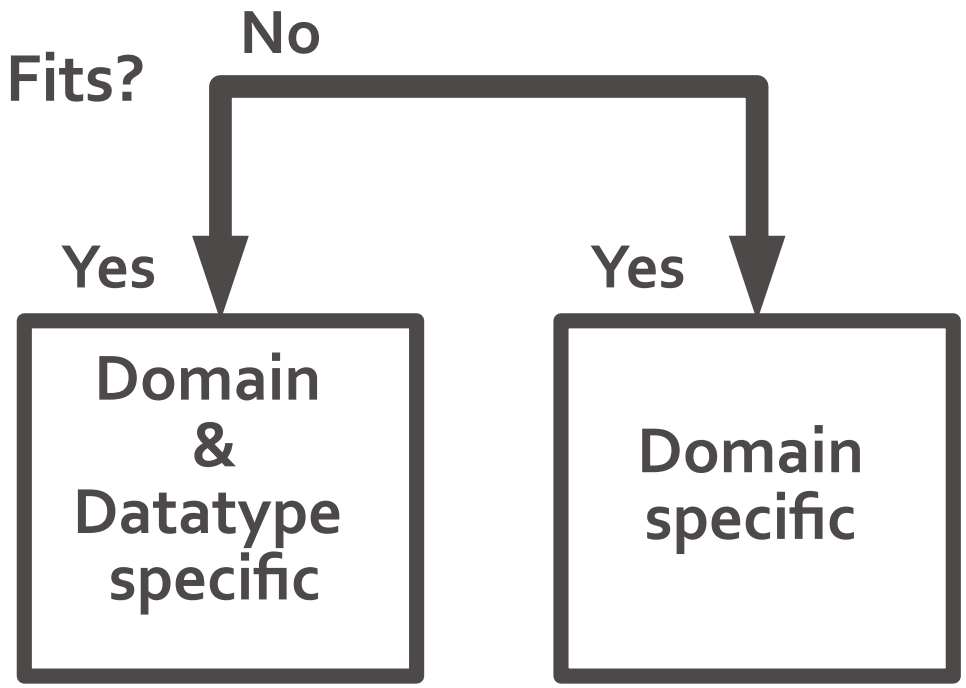
Yes

Domain
&
Datatype
specific

ENA
European Nucleotide Archive

NCBI SRA

Where should I deposit my data?

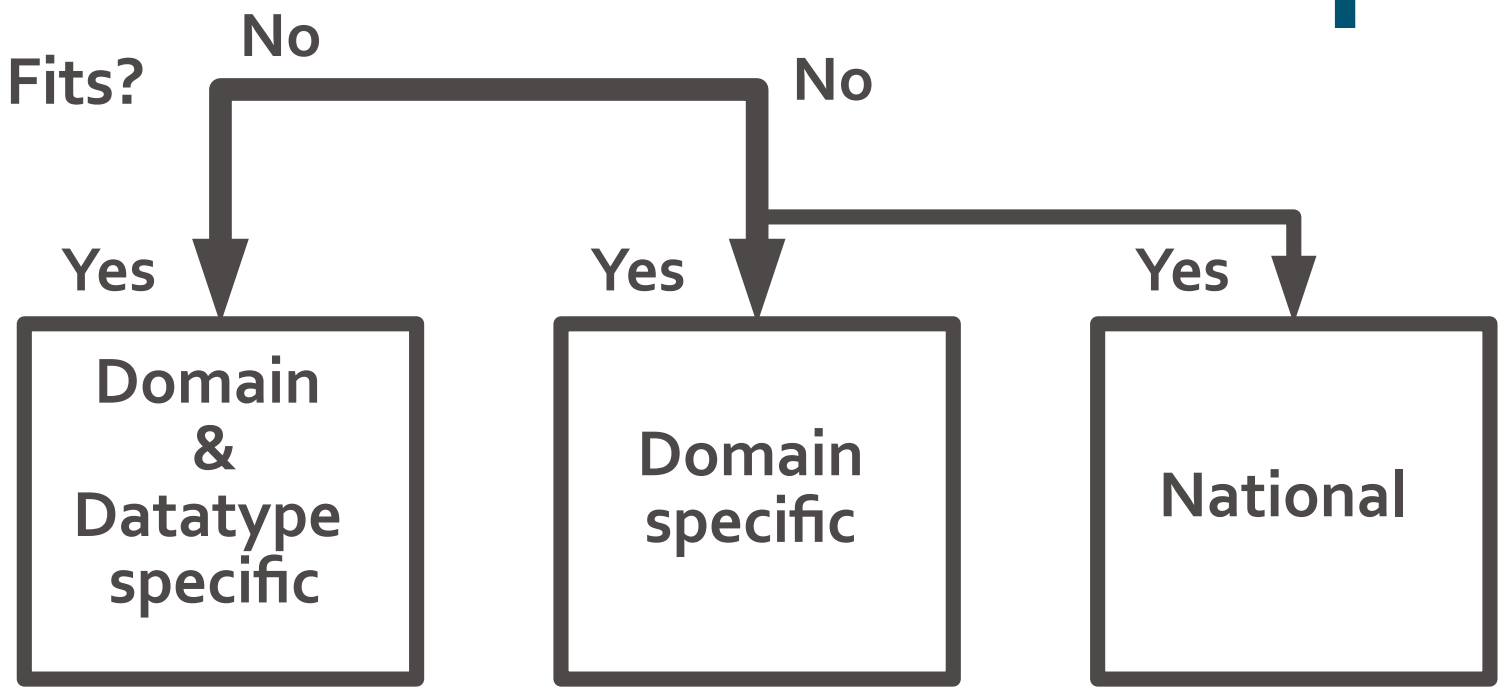


ENA
European Nucleotide Archive

BioStudies.

NCBI SRA

Where should I deposit my data?



ENA
European Nucleotide Archive

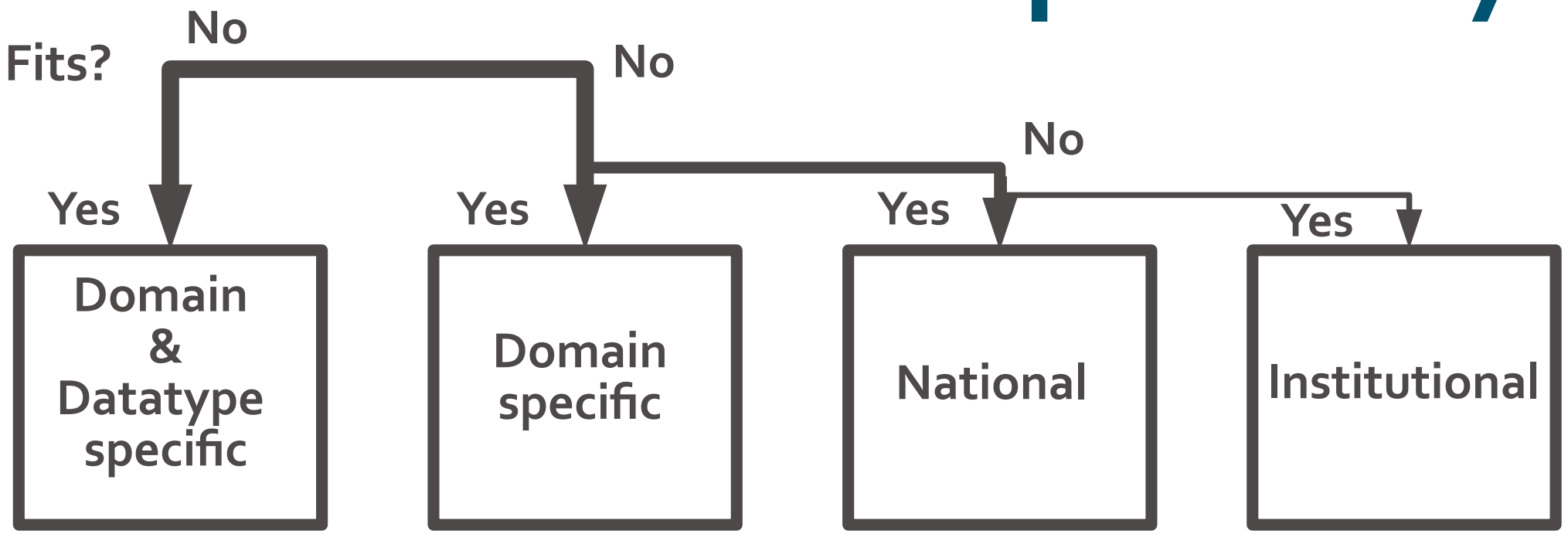
BioStudies.

NIRD RESEARCH DATA ARCHIVE

SRA
NCBI

DataverseNO
Dataverse Network Norway

Where should I deposit my data?



ENA
European Nucleotide Archive

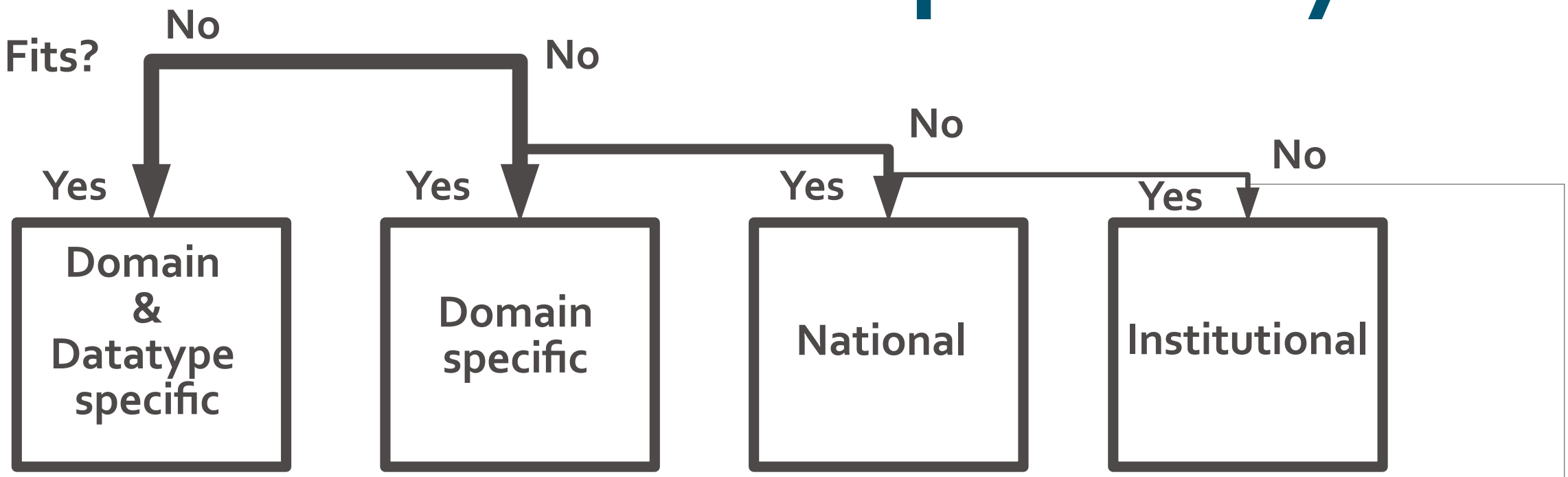
BioStudies.

NIRD RESEARCH DATA ARCHIVE

SRA
NCBI

DataverseNO
Dataverse Network Norway

Where should I deposit my data?



ENA European Nucleotide Archive

BioStudies.

NIRD RESEARCH DATA ARCHIVE

zenodo

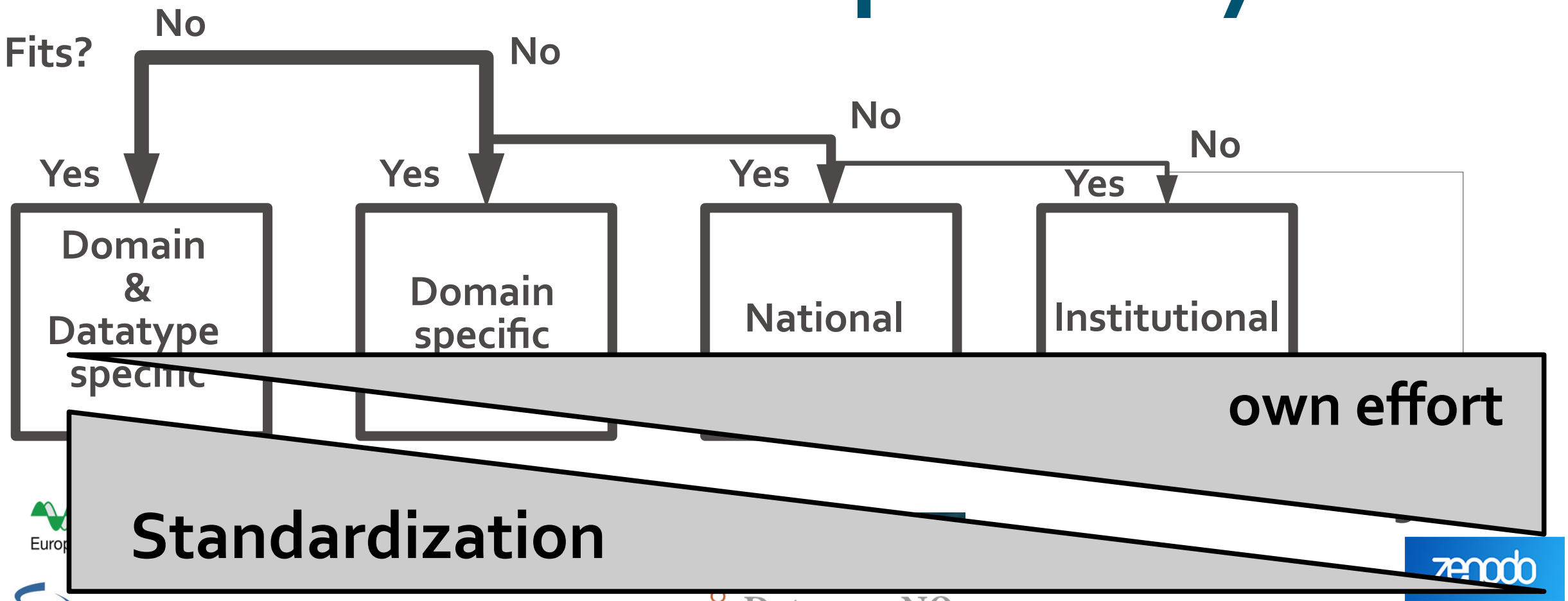
NCBI SRA

DataverseNO Dataverse Network Norway

EUDAT

figshare

Where should I deposit my data?



Europ

NCBI

SRA

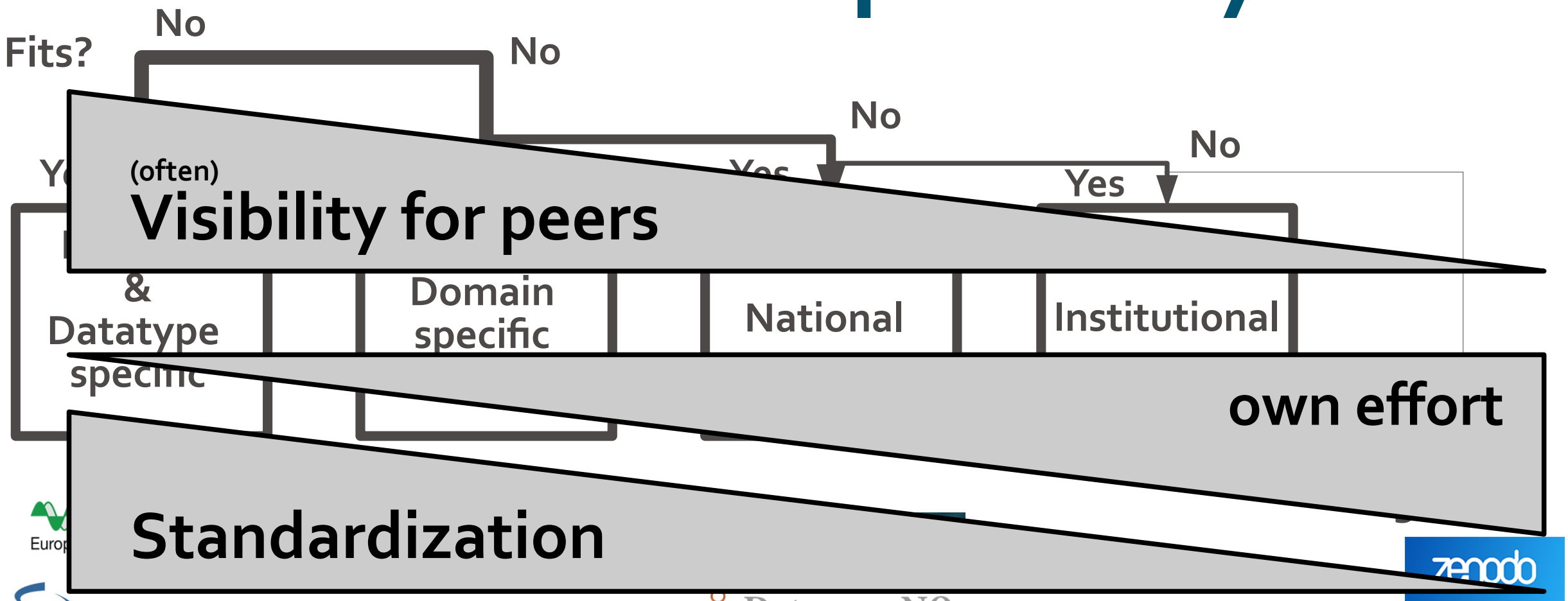
DataverseNO
Dataverse Network Norway

zenodo

EUDAT

figshare

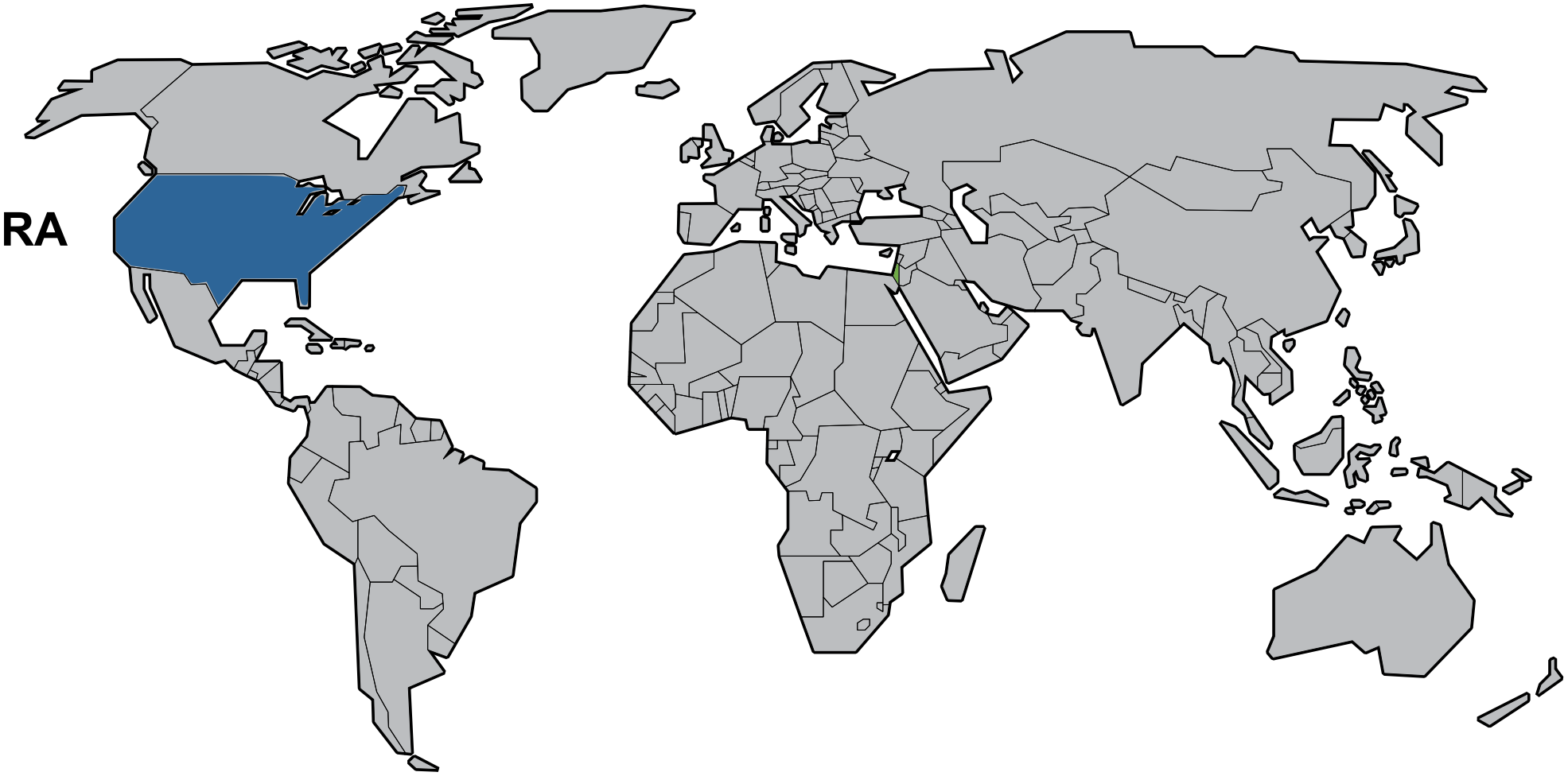
Where should I deposit my data?

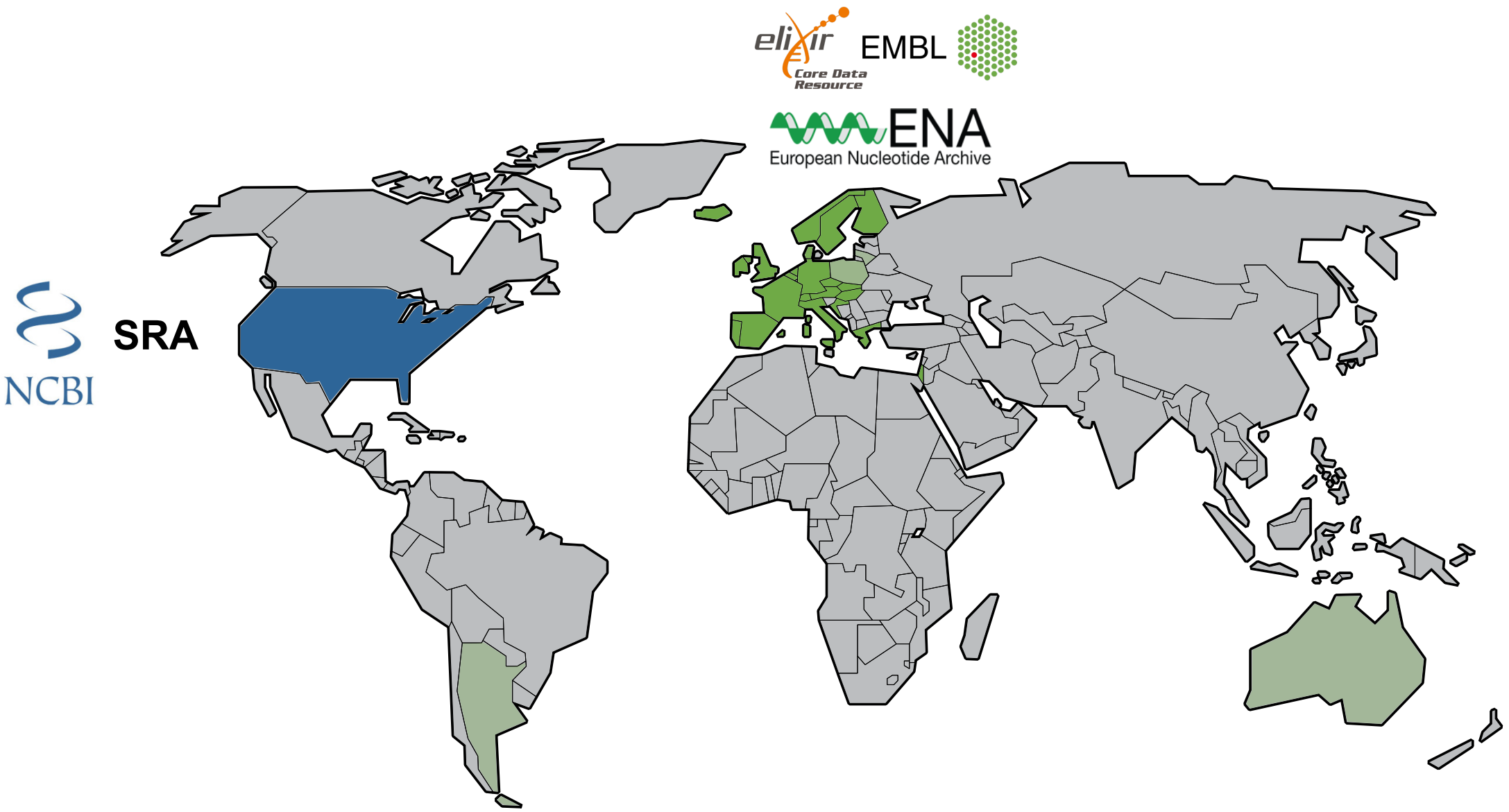


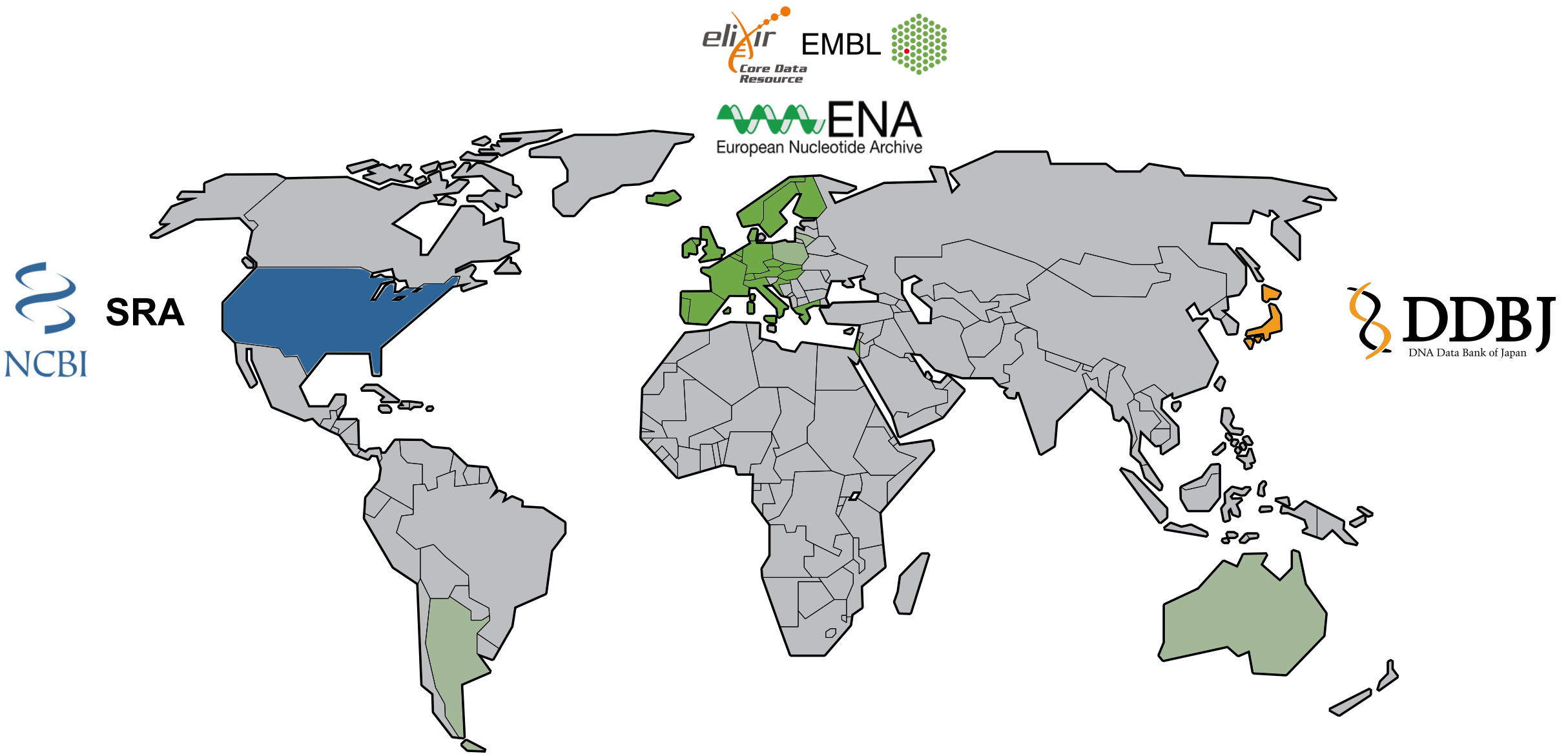
Multiple Repositories – similar datatypes

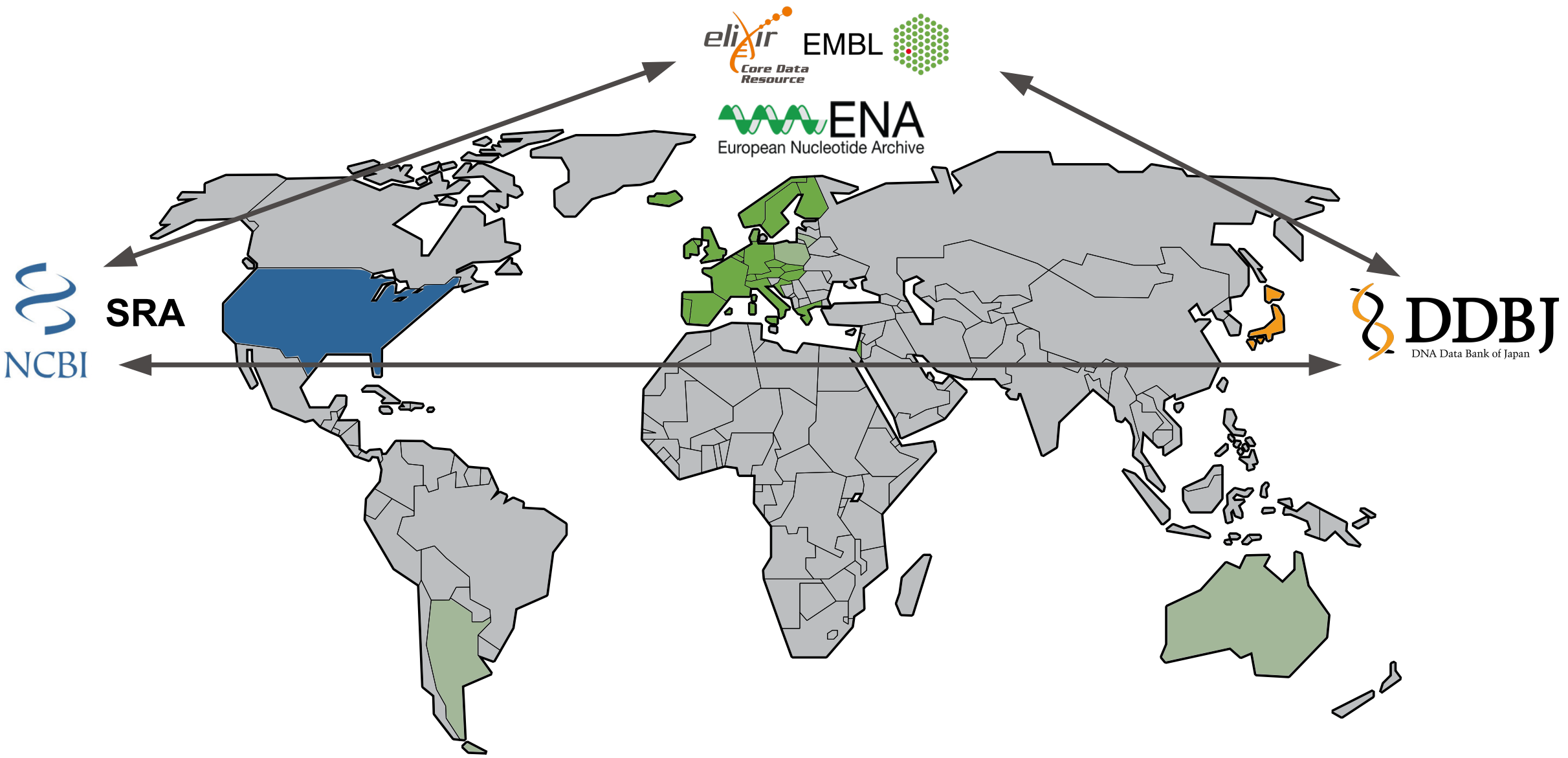
NCBI

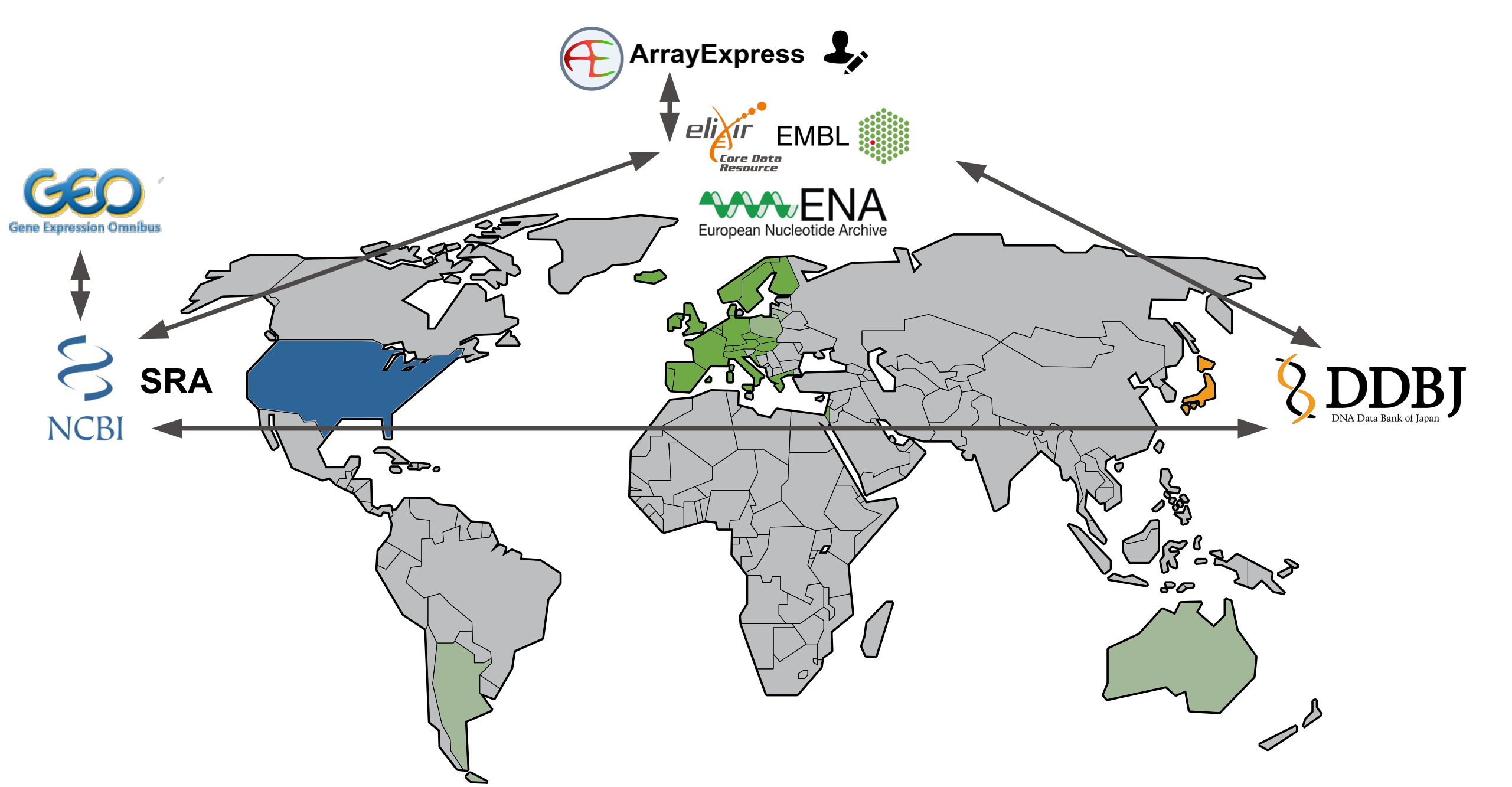
SRA

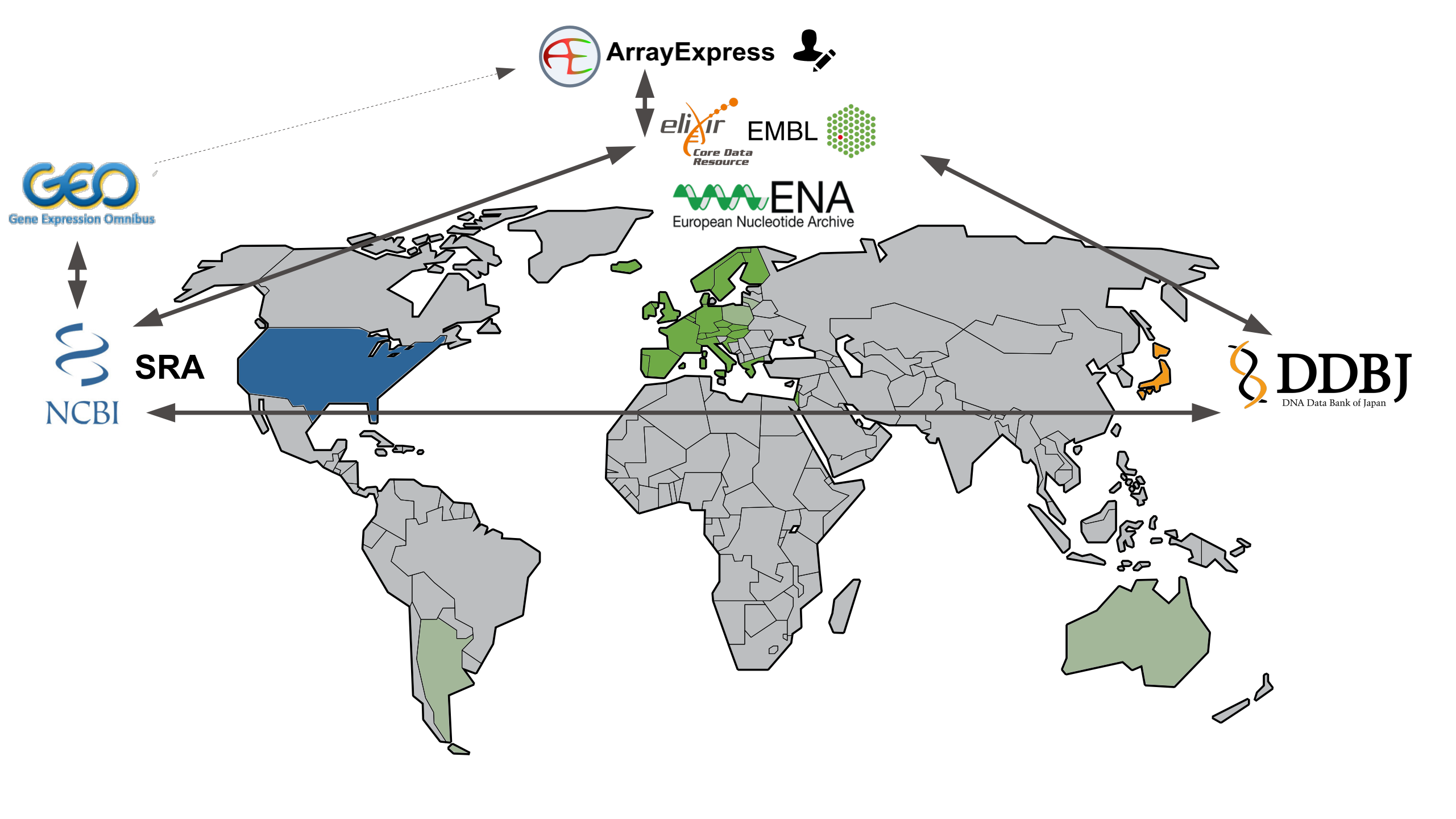


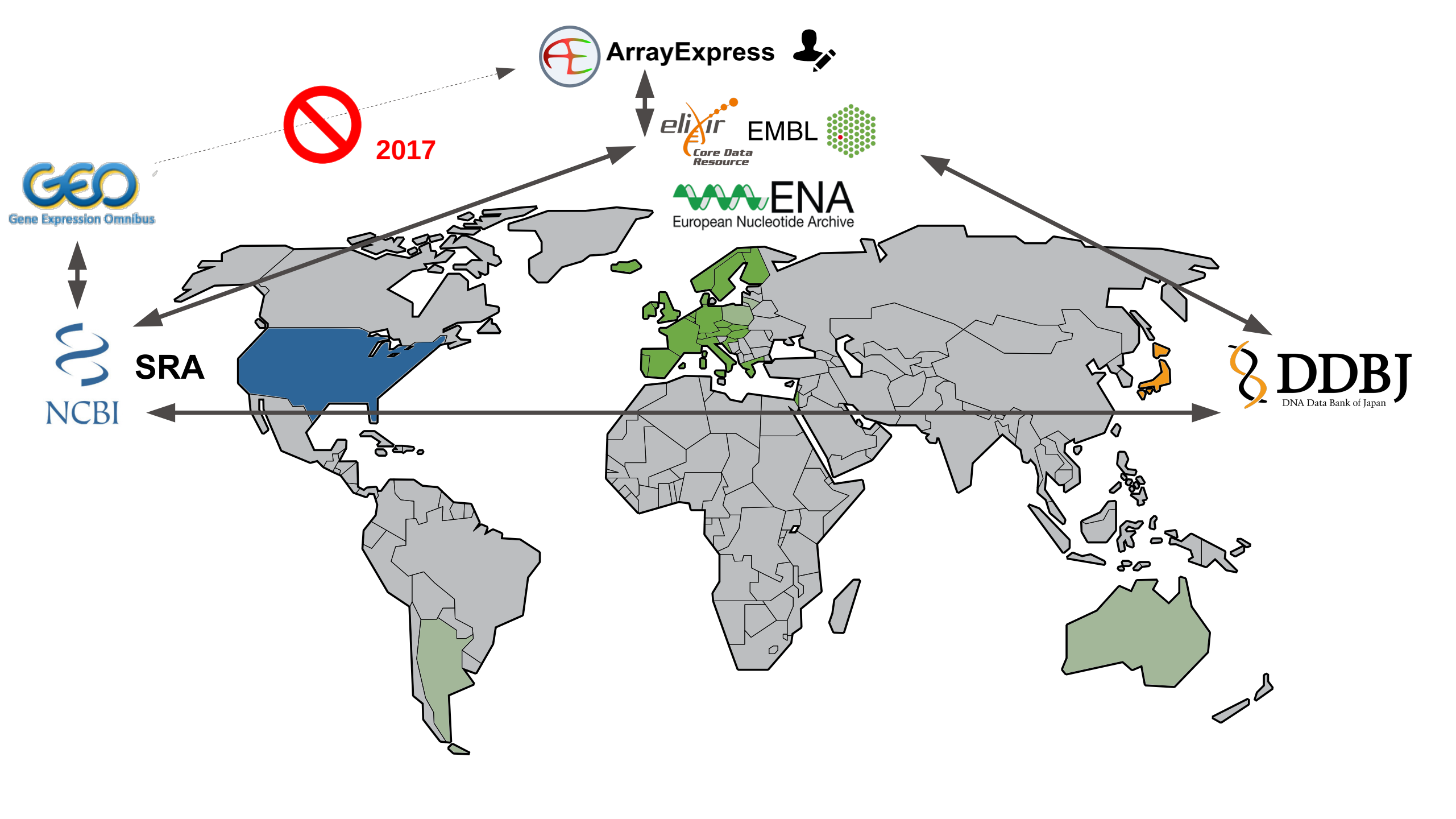

elixir
Core Data Resource
EMBL
ENA
European Nucleotide Archive
NCBI
SRA

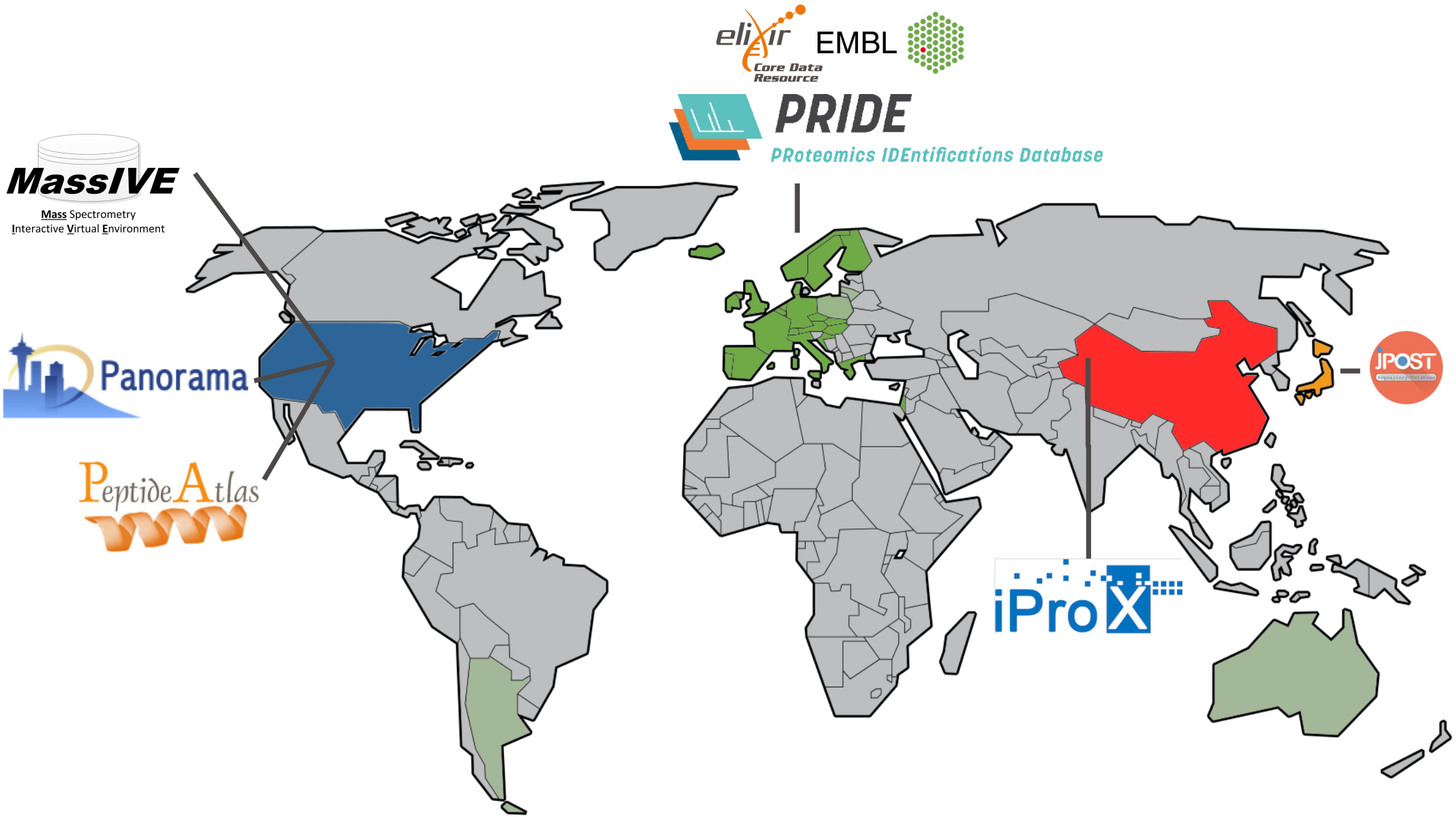

elixir
Core Data Resource
EMBL
ENA
European Nucleotide Archive
NCBI
SRA
DDBJ
DNA Data Bank of Japan

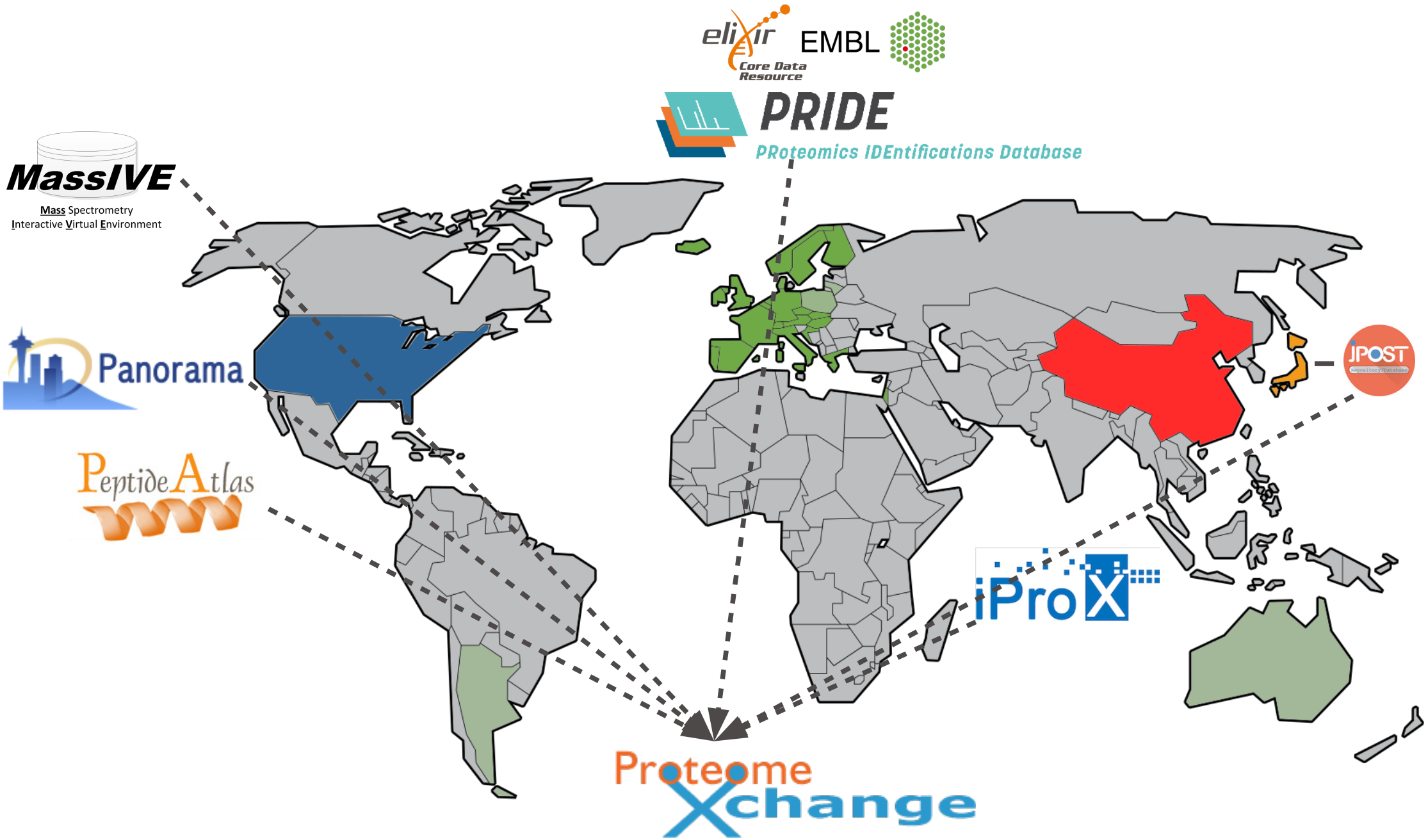

elixir
Core Data Resource
EMBL
ENA
European Nucleotide Archive
NCBI
SRA
DDBJ
DNA Data Bank of Japan


ArrayExpress
elixir
Core Data Resource
EMBL
ENA
European Nucleotide Archive
GEO
Gene Expression Omnibus
SRA
NCBI
DDBJ
DNA Data Bank of Japan


ArrayExpress
elixir
Core Data Resource
EMBL
ENA
European Nucleotide Archive
GEO
Gene Expression Omnibus
SRA
NCBI
DDBJ
DNA Data Bank of Japan


ArrayExpress
2017
elixir
Core Data Resource
EMBL
GEO
Gene Expression Omnibus
ENA
European Nucleotide Archive
SRA
NCBI
DDBJ
DNA Data Bank of Japan


elixir
Core Data Resource
EMBL
PRIDE
PRoteomics IDEntifications Database
MassIVE
Mass Spectrometry
Interactive Virtual Environment
Panorama
PeptideAtlas
iProX
JPOST


elixir
Core Data Resource
EMBL
PRIDE
PRoteomics IDEntifications Database
MassIVE
Mass Spectrometry
Interactive Virtual Environment
Panorama
PeptideAtlas
JPOST
iProX
ProteomeXchange

Where should I deposit my data?

- Free of charge

elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

Where should I deposit my data?

- Free of charge
- Quick upload (Aspera)

elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

Where should I deposit my data?

- Free of charge
- Quick upload (Aspera)
- Not operated by single university/government

elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

Where should I deposit my data?

- Free of charge
- Quick upload (Aspera)
- Not operated by single university/government
- Guided submission process

elixir Core Data Resource

Annotare

elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

Where should I deposit my data?

- Free of charge
- Quick upload (Aspera)
- Not operated by single university/government
- Guided submission process
- Generous embargo regulation (2yrs →)

elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

Where should I deposit my data?

- Free of charge
- Quick upload (Aspera)
- Not operated by single university/government
- Guided submission process 

- Generous embargo regulation (2yrs →)
- API access (upload from NeLS planned)

elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

Where should I deposit my data?

ENA
European Nucleotide Archive

ArrayExpress

European Variation Archive

elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

Where should I deposit my data?

ENA
European Nucleotide Archive

ArrayExpress

European Variation Archive

PDBe
Protein Data Bank in Europe

elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

Where should I deposit my data?

ENA
European Nucleotide Archive

ArrayExpress

European Variation Archive

PRIDE
PRoteomics IDEntifications Database

PDBe
Protein Data Bank in Europe

elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

Where should I deposit my data?

ENA
European Nucleotide Archive

ArrayExpress

European Variation Archive

PRIDE
PRoteomics IDEntifications Database

PDBe
Protein Data Bank in Europe

MetaboLights

elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

Where should I deposit my data?

ENA
European Nucleotide Archive

ArrayExpress

BioImage Archive

EMPIAR

IDR

European Variation Archive

PRIDE
PRoteomics IDEntifications Database

BioModels

PDBe
Protein Data Bank in Europe

MetaboLights

elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

Where should I deposit my data?

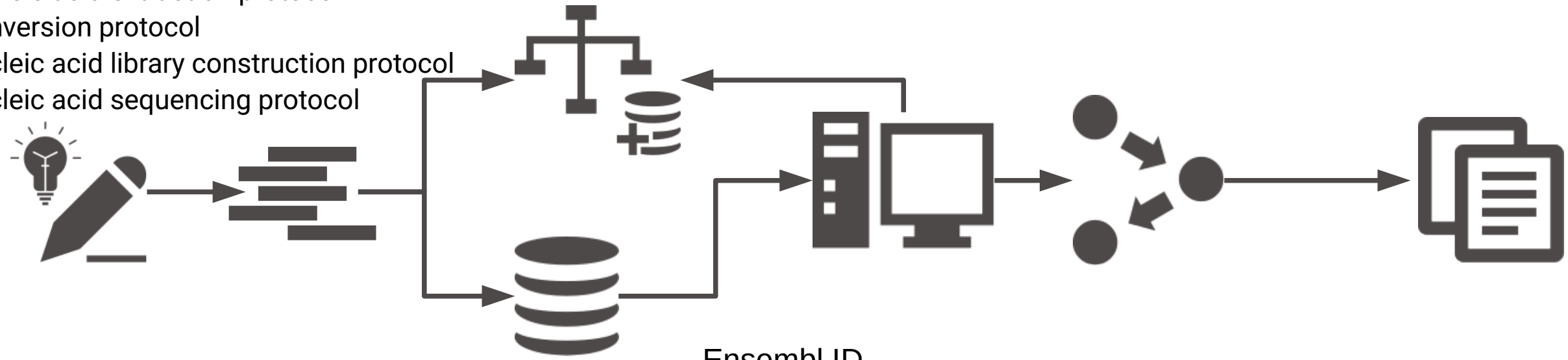
elixir-europe.org/platforms/data/core-data-resources
elixir-europe.org/platforms/data/elixir-deposition-databases

MINSEQE

Experimental conditions/design

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

SOPs
high throughput sequence alignment protocol
normalization data transformation protocol



fastq
bam
csv/tsv

Ensembl ID
RefSeq ID

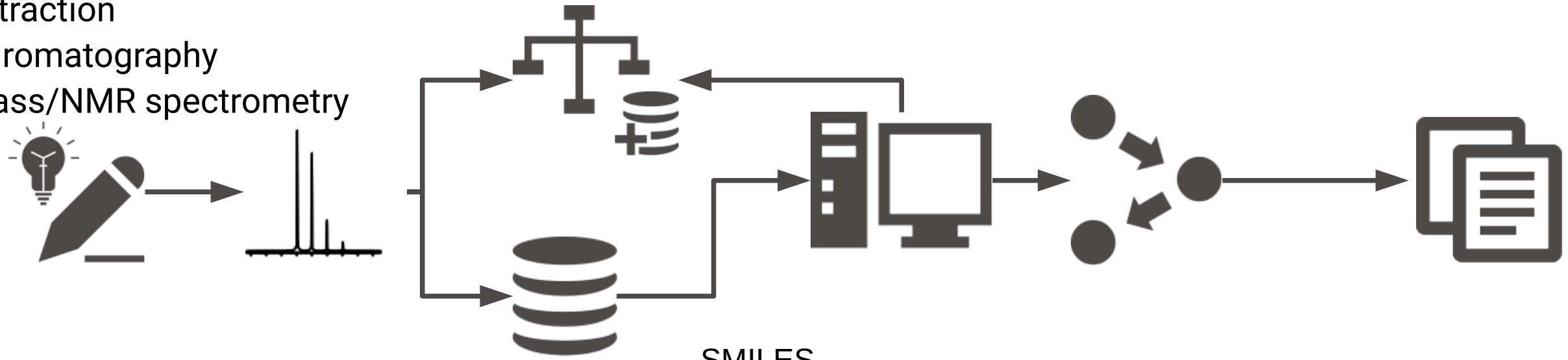
CIMR - ISA-tab

Experimental conditions/design

SOPs
Sample collection
Extraction
Chromatography
Mass/NMR spectrometry

SOPs
Data transformation
Metabolite identification

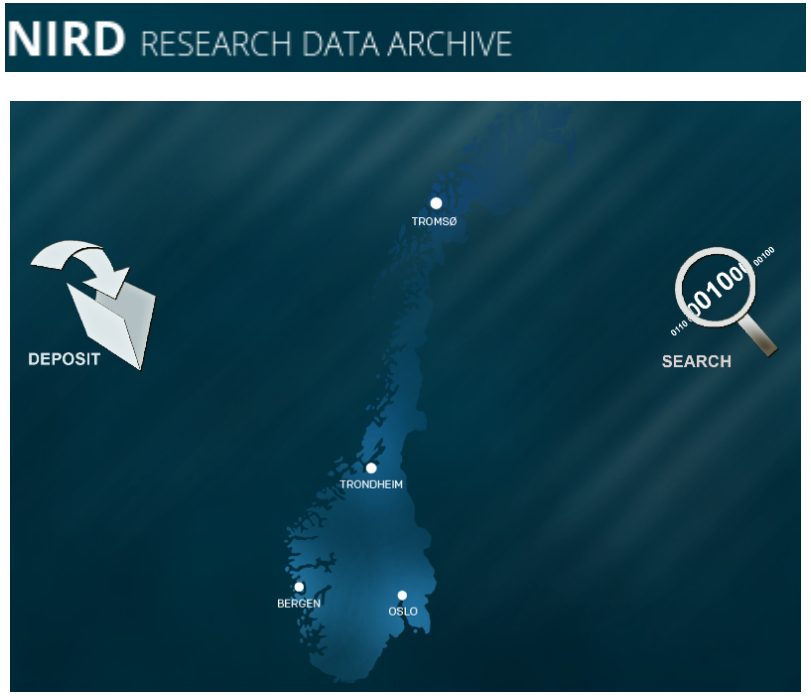
MetaboLights



mzML
nmrML

SMILES
InCh

NIRD archive



National, free depositing repository

Domain agnostic

Dublin-core metadata standard

DOI accessible

Supports machine readable metadata harvesting

https://archive.norstore.no/

Generic archives

Less standard metadata → Re-usability ⇓

Domain agnostic → Findability ⇓

Guided submission process

Commercial actors: Size limitations, …

https://zenodo.org/
https://b2share.eudat.eu/
https://figshare.com/

What about sensitive data?

EUROPEAN
GENOME-PHENOME
ARCHIVE

Central metadata accessibility

Secure storage

Implemented data access committees

Norwegian EGA
(for data that can not leave the country)

NSD archive

NSD NORSK SENTER FOR FORSKNINGSDATA

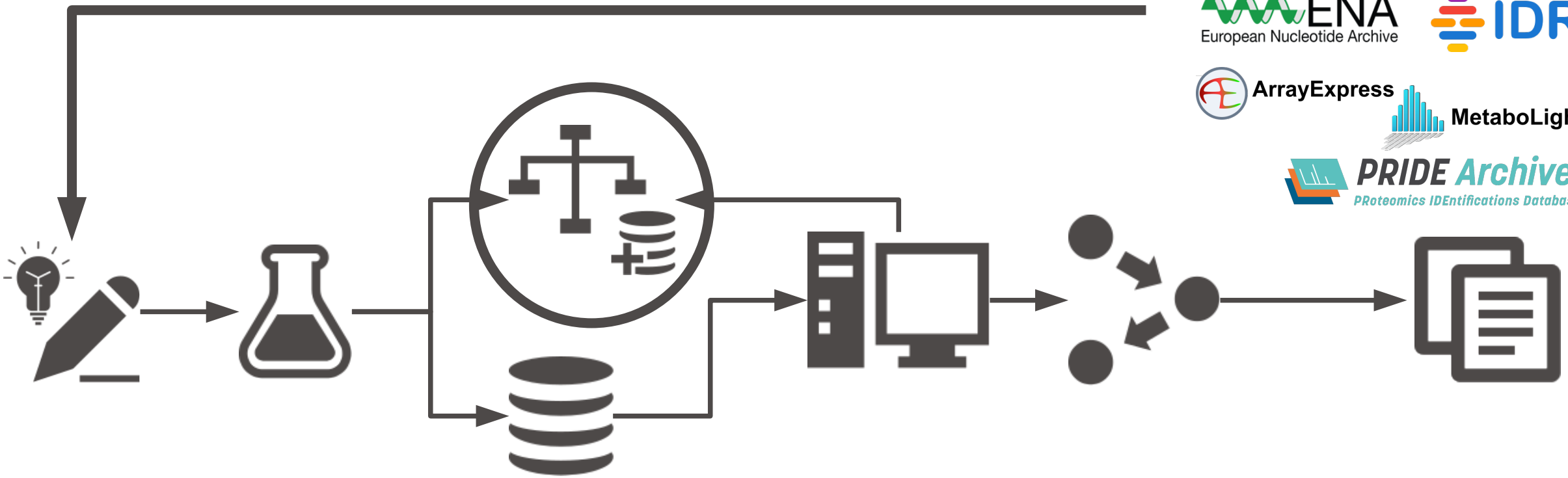
Domain agnostic

DOI accessible

Implemented data access committees

Main data type: Questionaires, …

https://arkivering.nsd.no/



BioStudies.

ENA European Nucleotide Archive

IDR

ArrayExpress

MetaboLights

PRIDE Archive
PRoteomics IDEntifications Database

EUROPEAN GENOME-PHENOME ARCHIVE