



Data Deposition Repositories



Korbinian Bösl
Data manager ELIXIR Norway
4 March 2020



NeLS

Norwegian e-Infrastructure for Life Sciences



ArrayExpress



MetaboLights



PRIDE Archive
PRoteomics IDentifications Database



EUROPEAN
GENOME-PHENOME
ARCHIVE

 **nor seq**



sensitive
data



Why should I deposit my data?

- 👁 Increased Visibility (SEO) → more citations

Why should I deposit my data?






Increased Visibility (SEO) → more citations



Value added Databases → more citations

Why should I deposit my data?

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

Why should I deposit my data?

 Increased Visibility (SEO) → more citations

 Value added Databases → more citations

 FAIRification – Funding requirements ✓✓

 Safe money on storage

Where should I deposit my data?

Nucleic Acids Research

VOLUME 48 DATABASE ISSUE JANUARY 8, 2020
<https://academic.oup.com/nar>



OXFORD
UNIVERSITY PRESS

Open Access

No barriers to access – all articles freely available online



molecular biology
1637 databases

Where should I deposit my data?

FAIRsharing


re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

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}	-



Ontology Lookup Service

Home **Ontologies** Documentation About

OLS > eNanoMapper Ontology **ENM** > **ENVO:00000447**

marine biome

http://purl.obolibrary.org/obo/ENVO_00000447

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>]

Tree view | Term history

- entity
 - material entity
 - biome
 - aquatic biome
 - marine biome**

Graph view
Reset tree
Show all siblings

Term info

- database cross reference
 - SPIRE:Marine
- has obo namespace
 - ENVO
- has related synonym
 - marine realm
- id
 - ENVO:00000447


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}	-



Ontology Lookup Service

Home **Ontologies** Documentation About

OLS > Gazetteer **GAZ** > **GAZ:00002699**

Kingdom of Norway

http://purl.obolibrary.org/obo/GAZ_00002699

A country and constitutional monarchy in Northern Europe that occupies the western portion of the Scandinavian Peninsula. It is bordered by Sweden, Finland, and Russia. The Kingdom of Norway also includes the Arctic island territories of Svalbard and Jan Mayen. Norwegian sovereignty over Svalbard is based upon the Svalbard Treaty, but that treaty does not apply to Jan Mayen. Bouvet Island in the South Atlantic Ocean and Peter I Island and Queen Maud Land in Antarctica are external dependencies, but those three entities do not form part of the kingdom. [url:<http://en.wikipedia.org/wiki/Norway>]

Synonyms: Kongeriket Norge {language: Norwegian}, Norway, Kongeriket Noreg {language: Norwegian}

Tree view | Term history

- geographic location
 - Kingdom of Norway**
 - Bouvet Islands
 - Dronning Maud Land
 - Jan Mayen
 - Metropolitan Norway
 - Lake Polden

Graph view | Reset tree | Show all siblings

Term info

database cross reference

- ISO3166-1:NO
- ISO3166-2:NO
- ISO3166-1:578
- ISO3166-1:NOR

ABBREVIATION

Norway

The GAZ ontology describes the geographical location of the sampling

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

...



PRIDE Archive

PRoteomics IDentifications Database

HUPO-PSI TraML
MIAPE

...



SRA-XML

Data format standards

Common formats

Non-proprietary formats (accessible with open source tools)

Avoiding binary data formats (data corruption)

Data format standards



ArrayExpress

FASTQ
MAGE-ML

...



FASTA
FASTQ

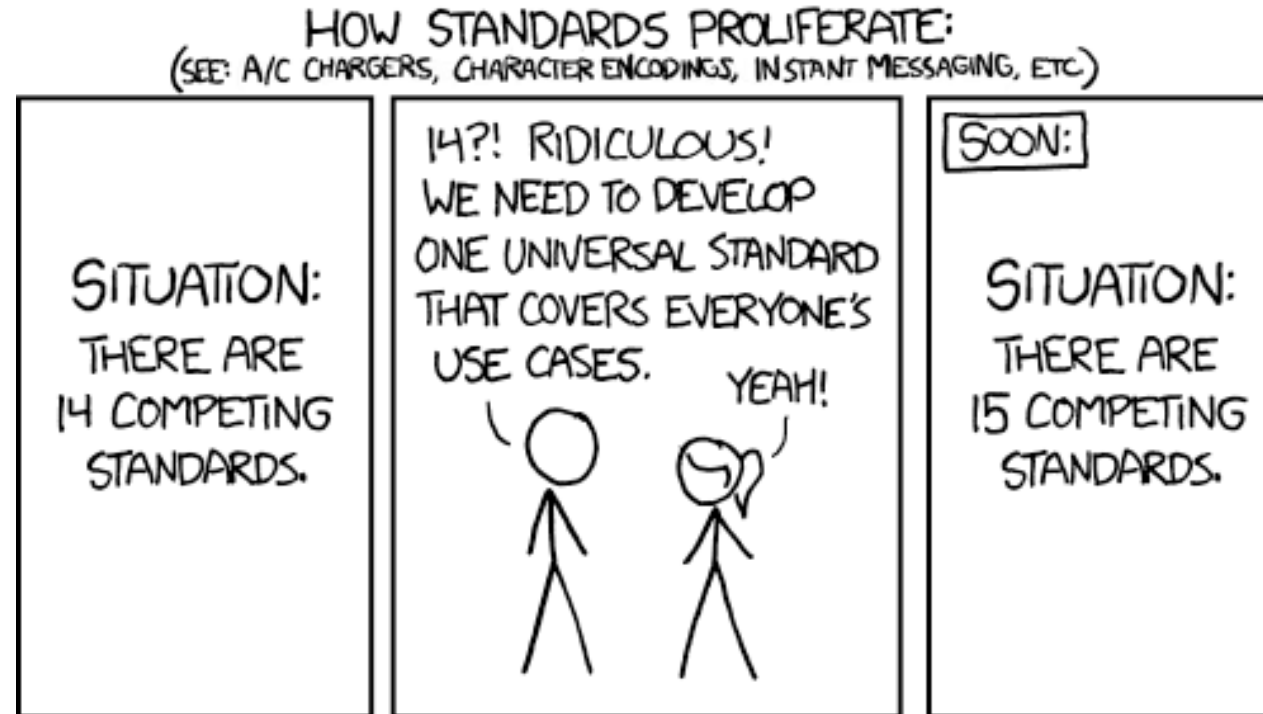
...



mzML
mzQuantML

...

Community standards vs. formal ISO standards



Where should I deposit my data?

Fits?

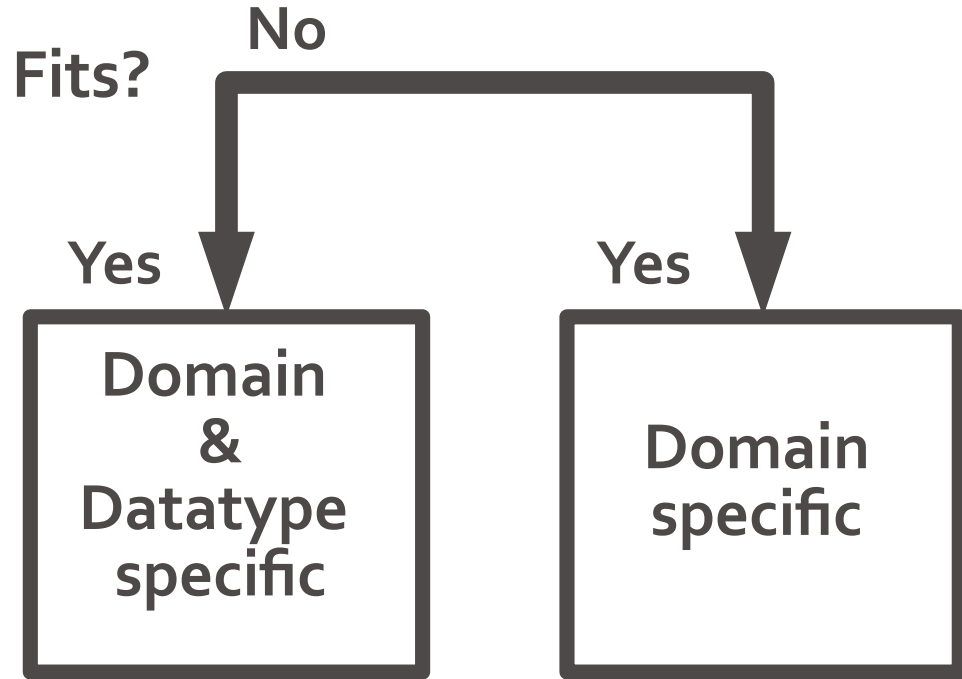


Yes

Domain
&
Datatype
specific



Where should I deposit my data?

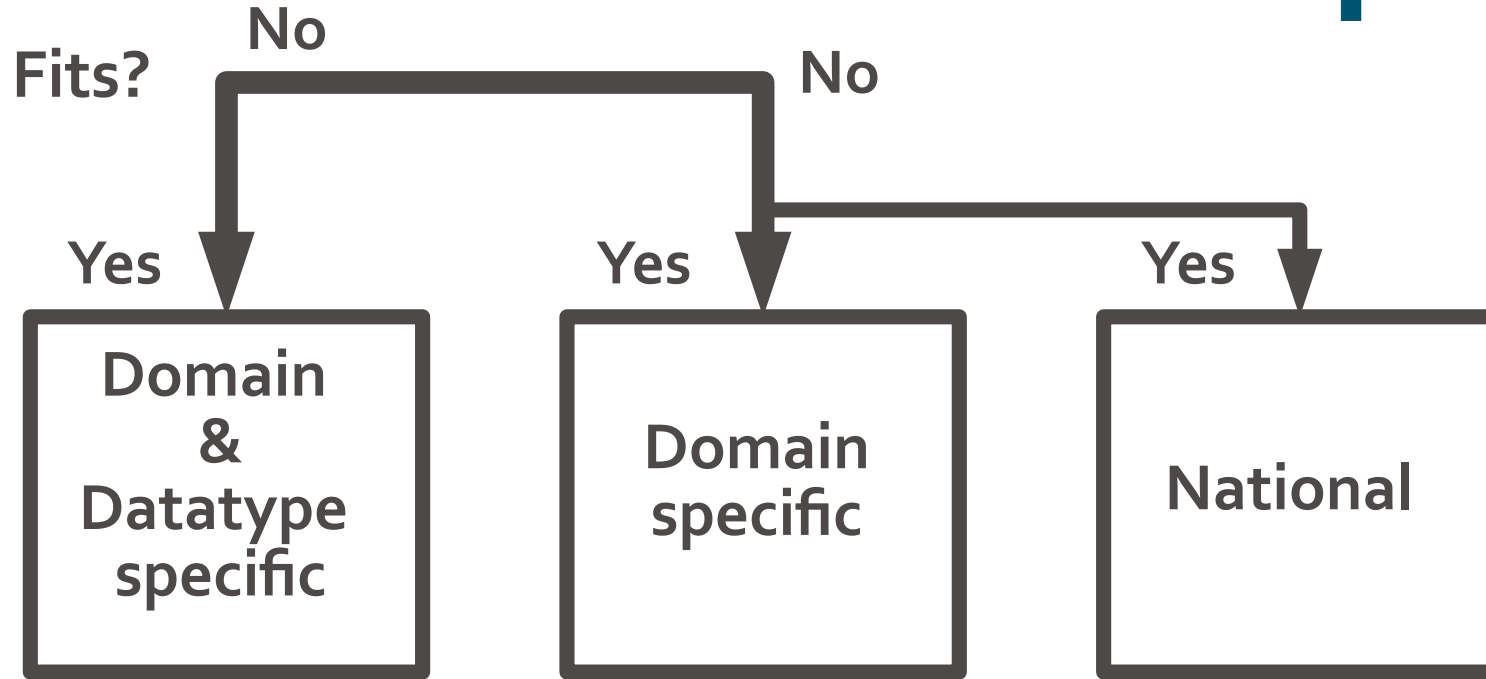


 **ENA**
European Nucleotide Archive

 **BioStudies.**

 **SRA**
NCBI

Where should I deposit my data?



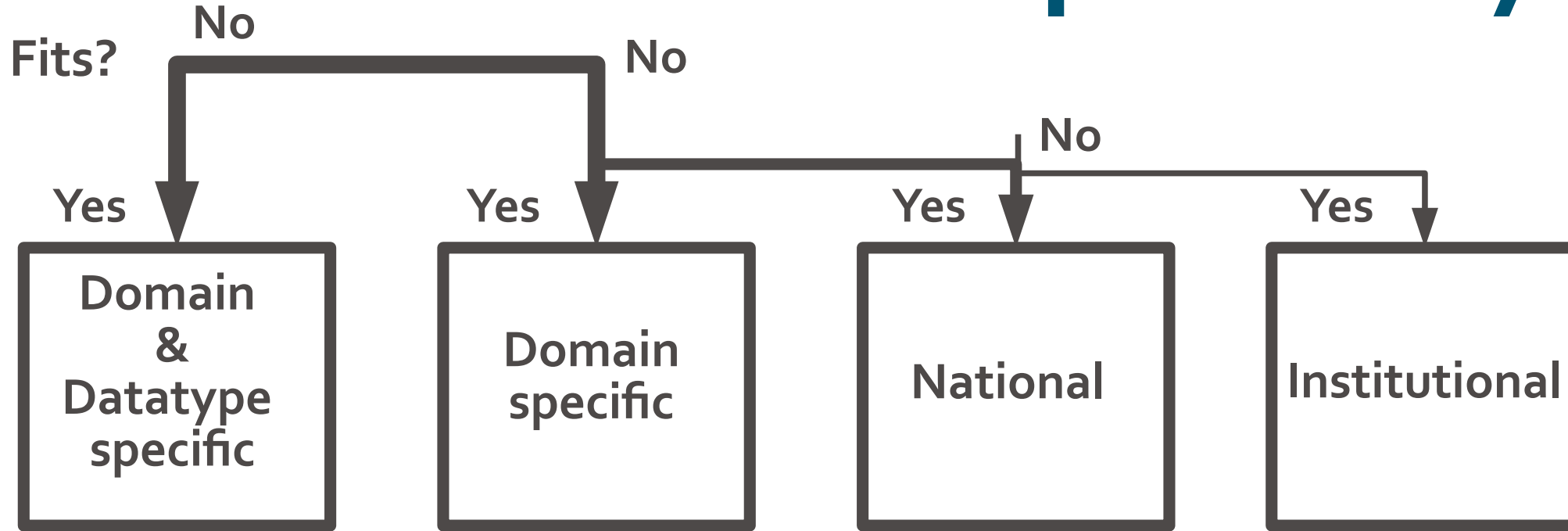
 **ENA**
European Nucleotide Archive

 **BioStudies.**

NIRD RESEARCH DATA ARCHIVE

 **SRA**
NCBI

Where should I deposit my data?



 **ENA**
European Nucleotide Archive


 **BioStudies.**


 **NIRD** RESEARCH DATA ARCHIVE

 **DataverseNO**
Dataverse Network Norway

 **NTNU Open Research Data**
NTNU

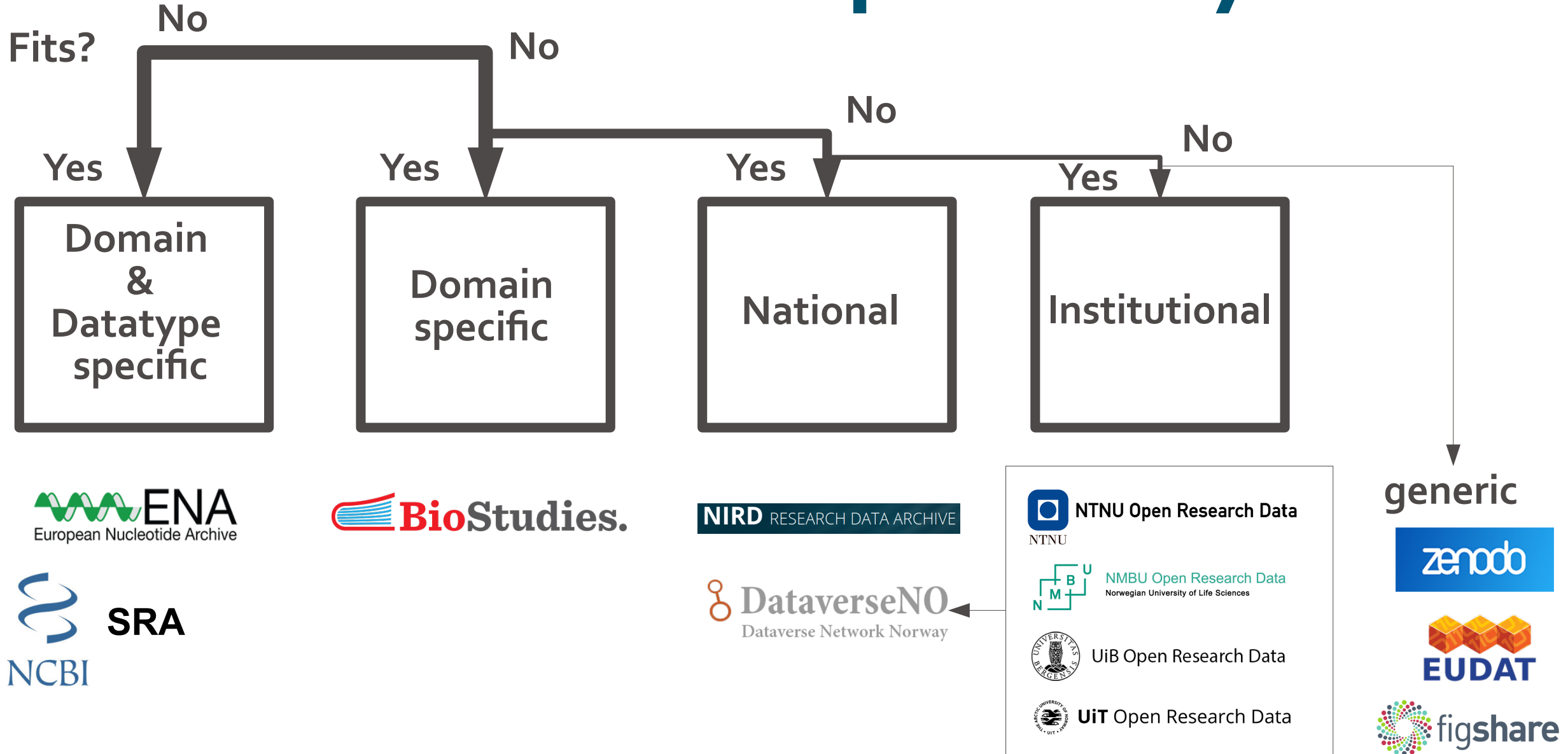
 **NMBU Open Research Data**
Norwegian University of Life Sciences

 **UiB Open Research Data**

 **UiT Open Research Data**

 **SRA**
NCBI

Where should I deposit my data?



Multiple Repositories – similar datatypes



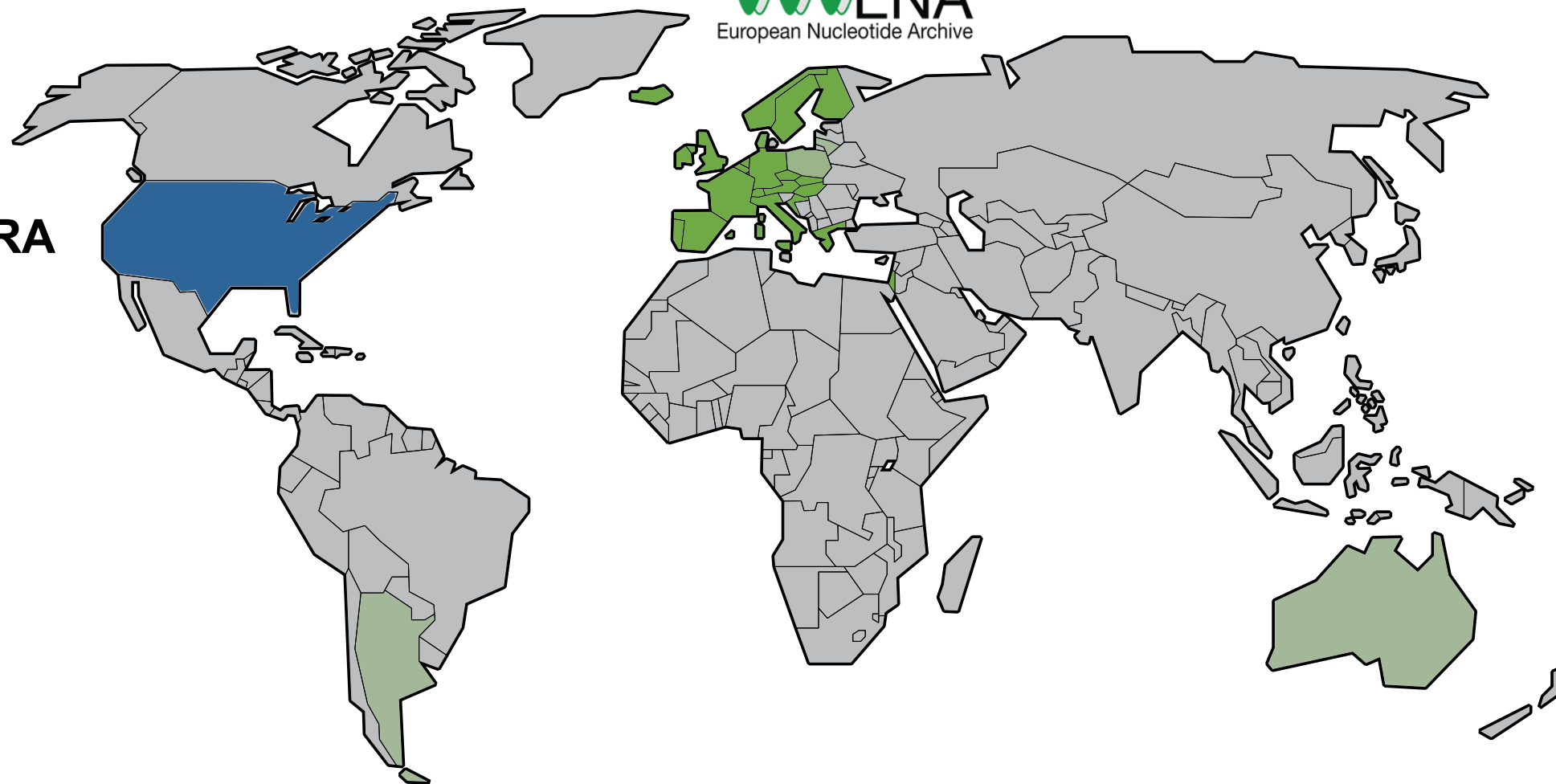


SRA



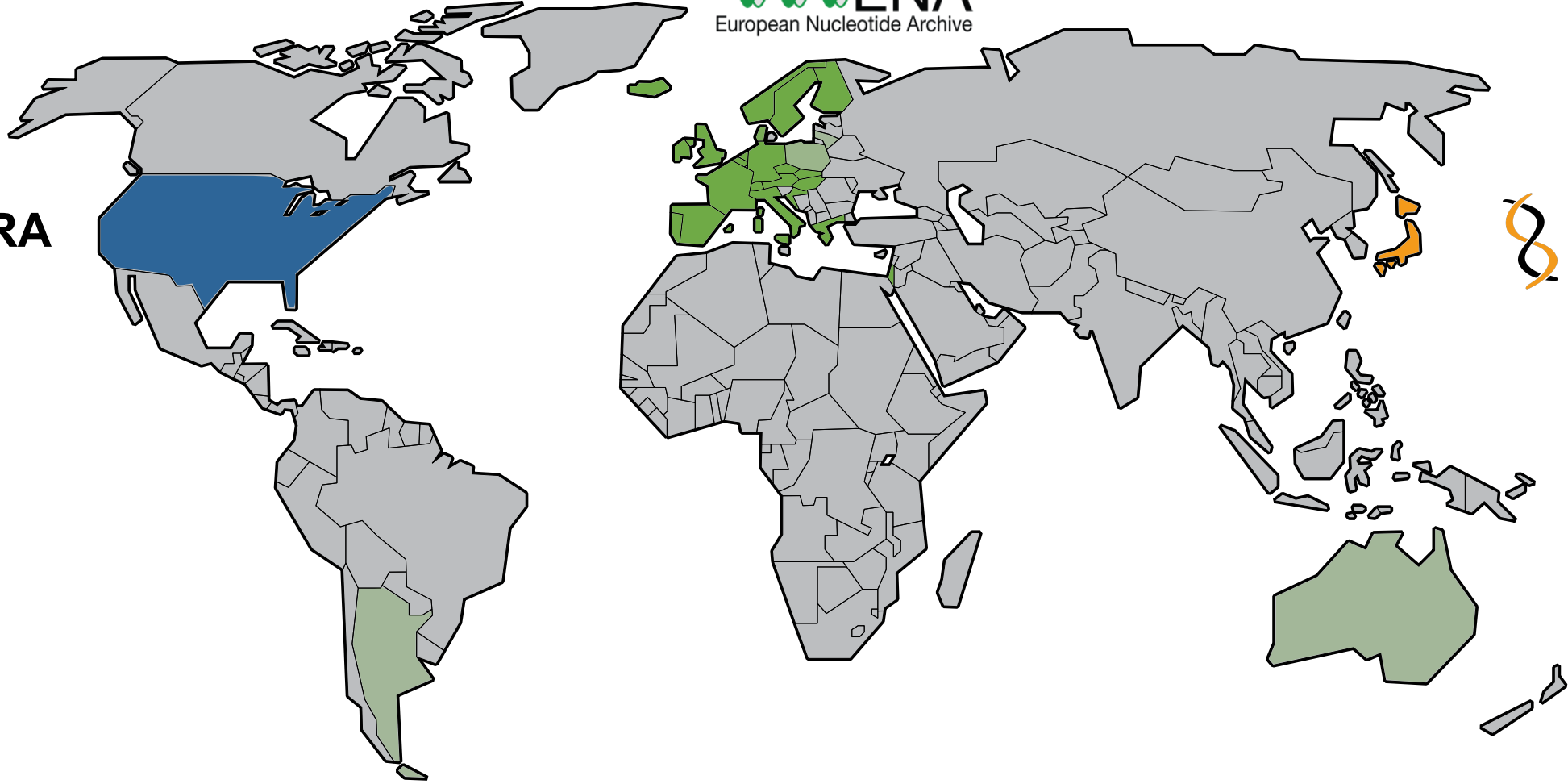


SRA





SRA

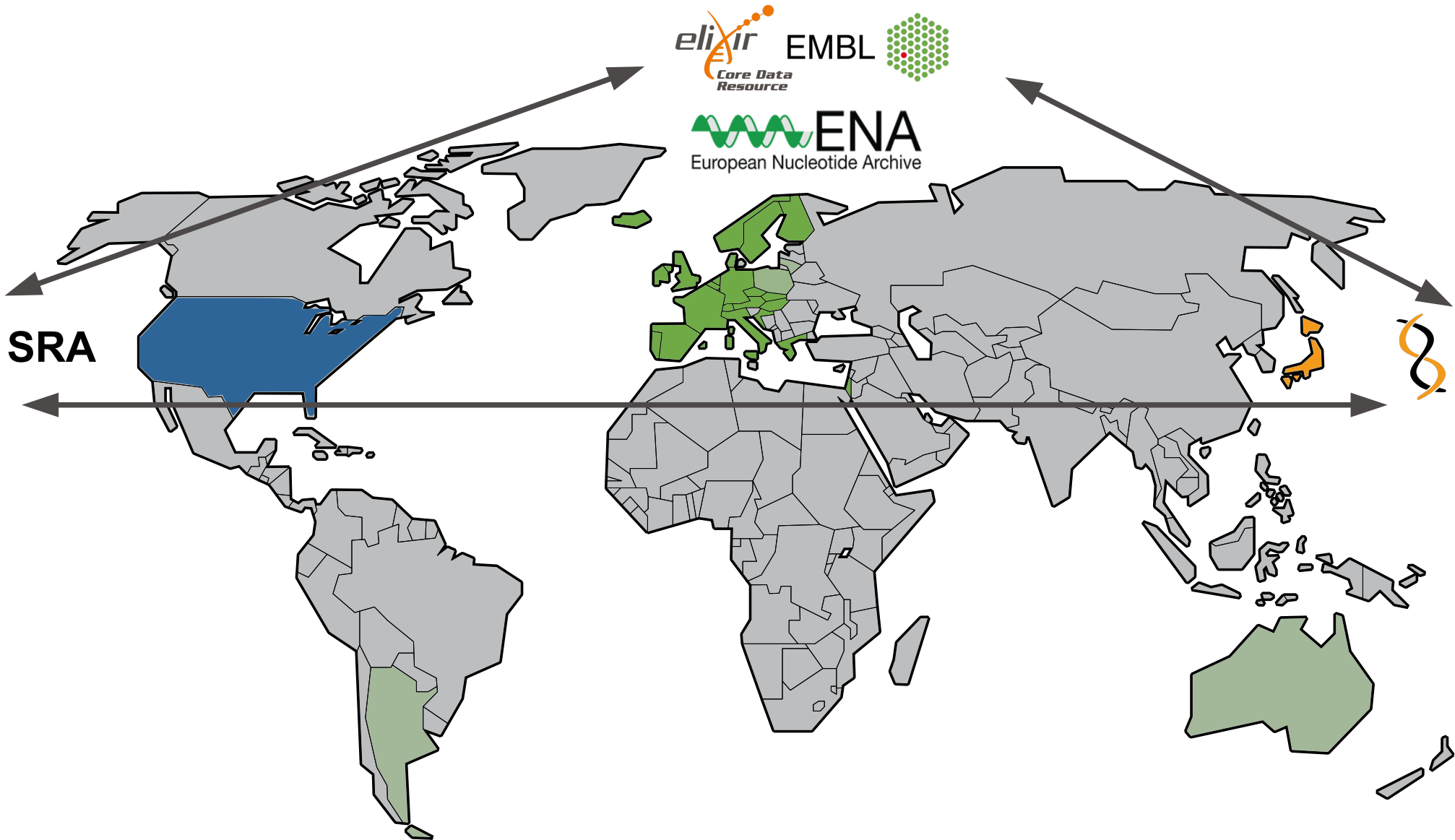


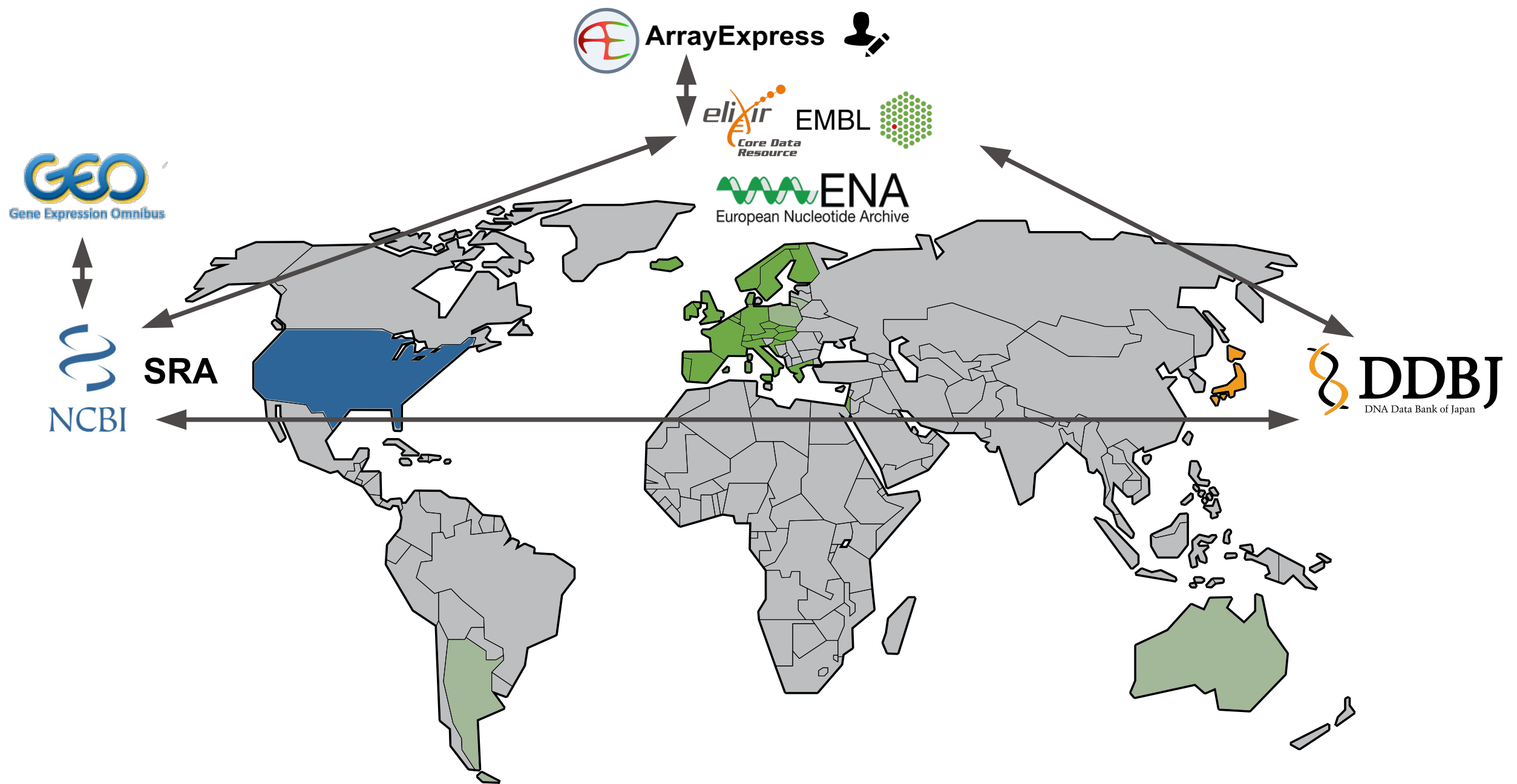


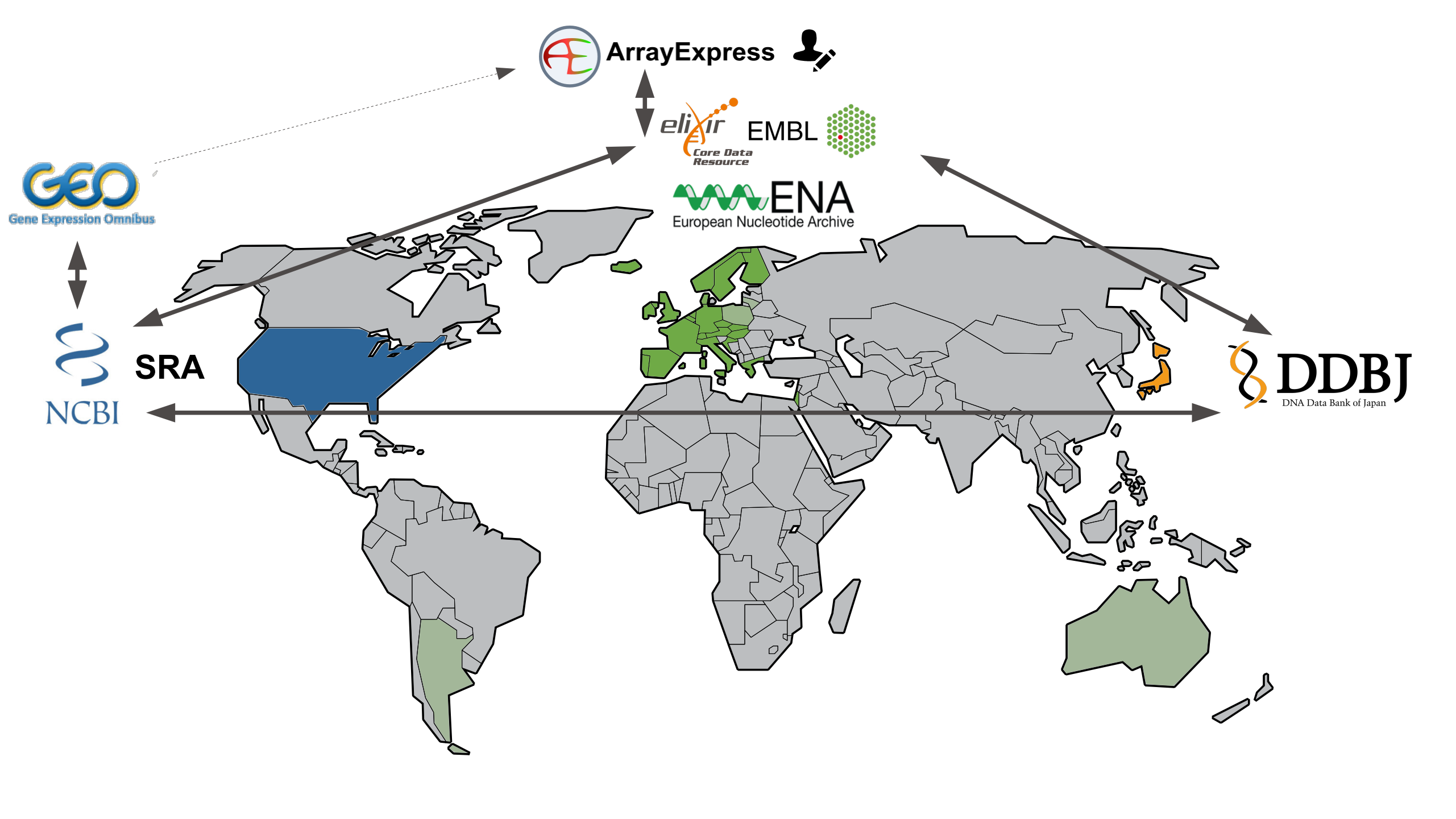
SRA

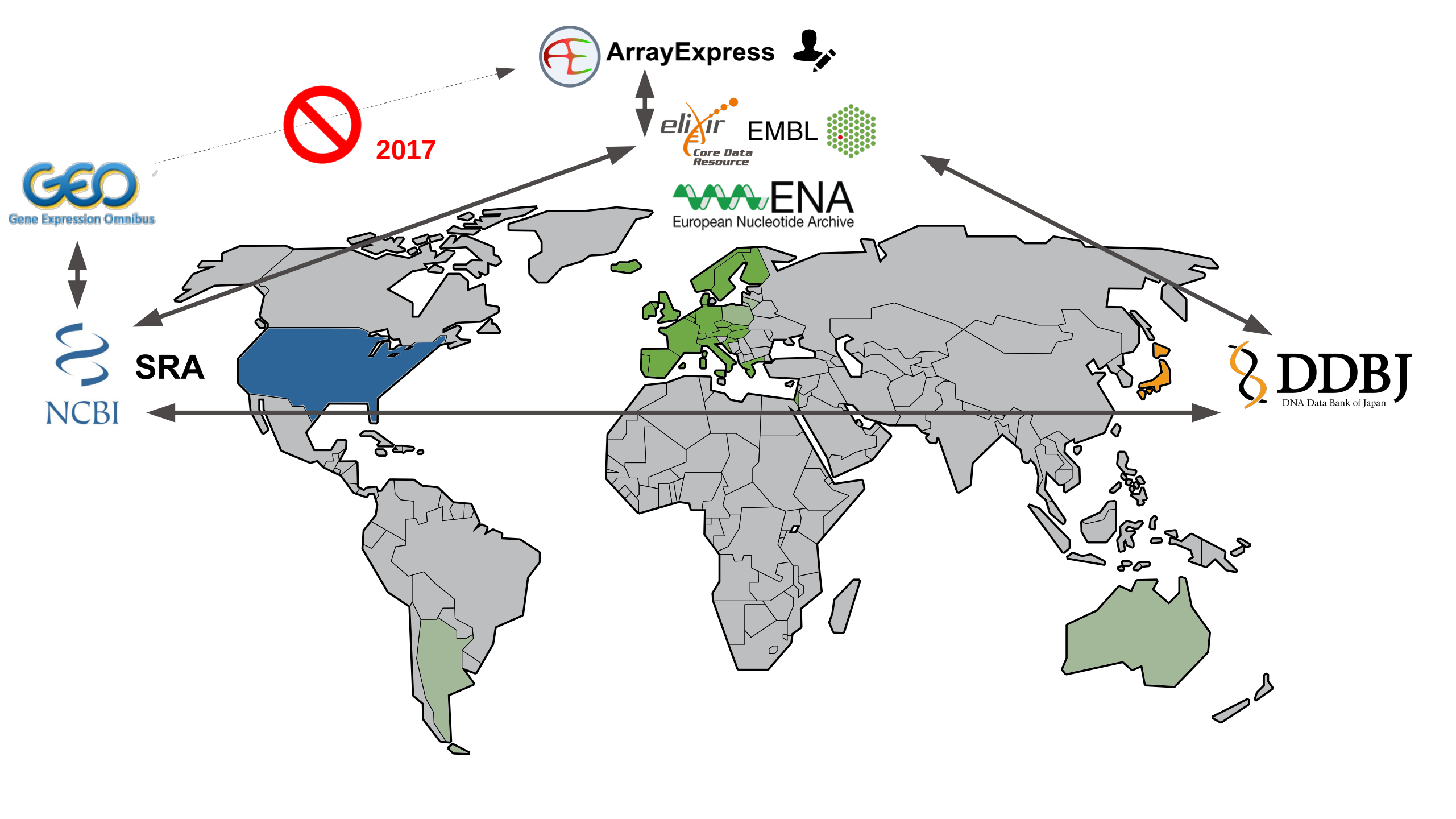


EMBL











PRIDE

*PR*oteomics *ID*entifications *D*atabase

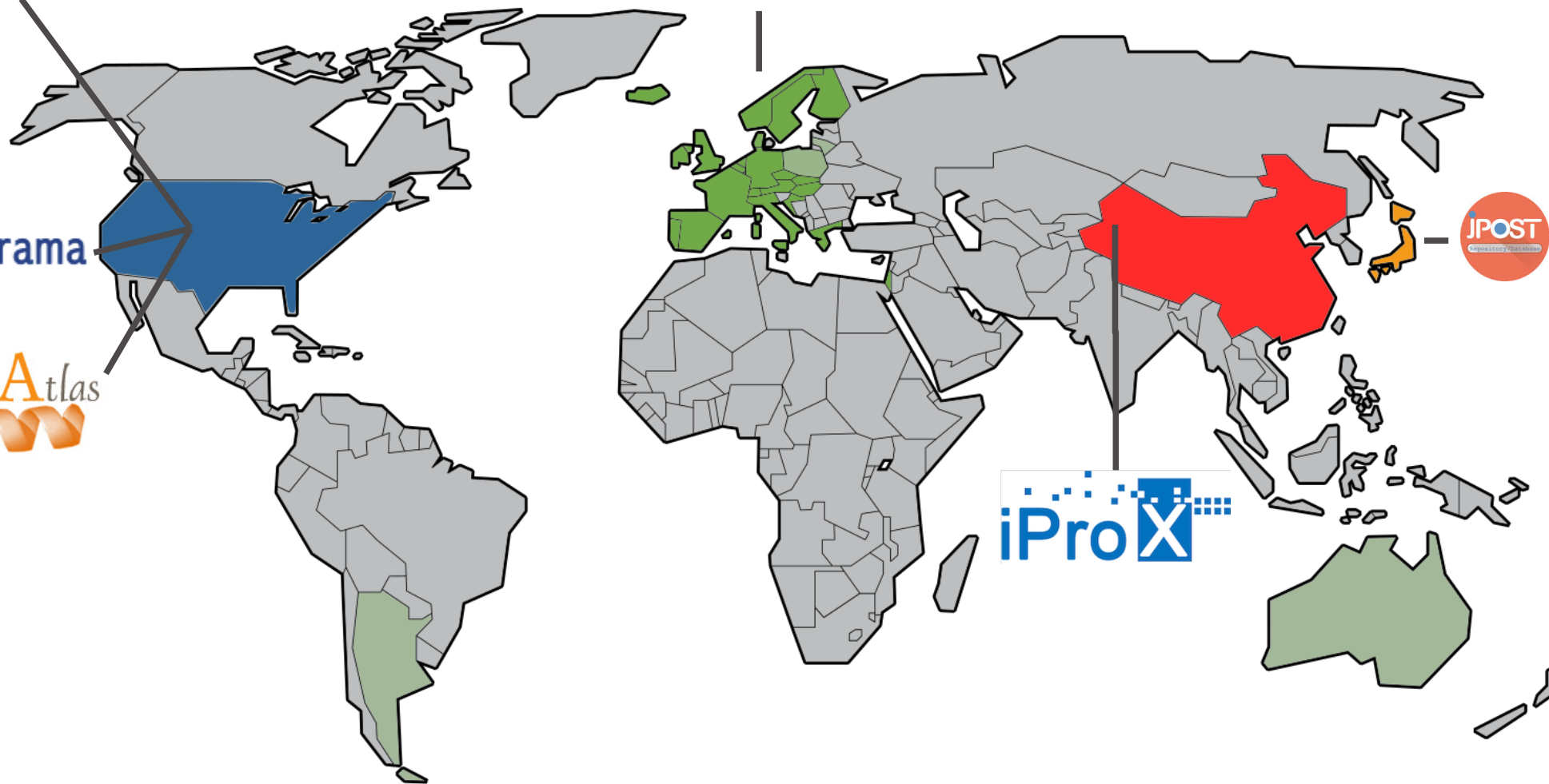


MassIVE

Mass Spectrometry
Interactive Virtual Environment



Panorama





PRIDE

PRoteomics IDentifications Database



MassIVE

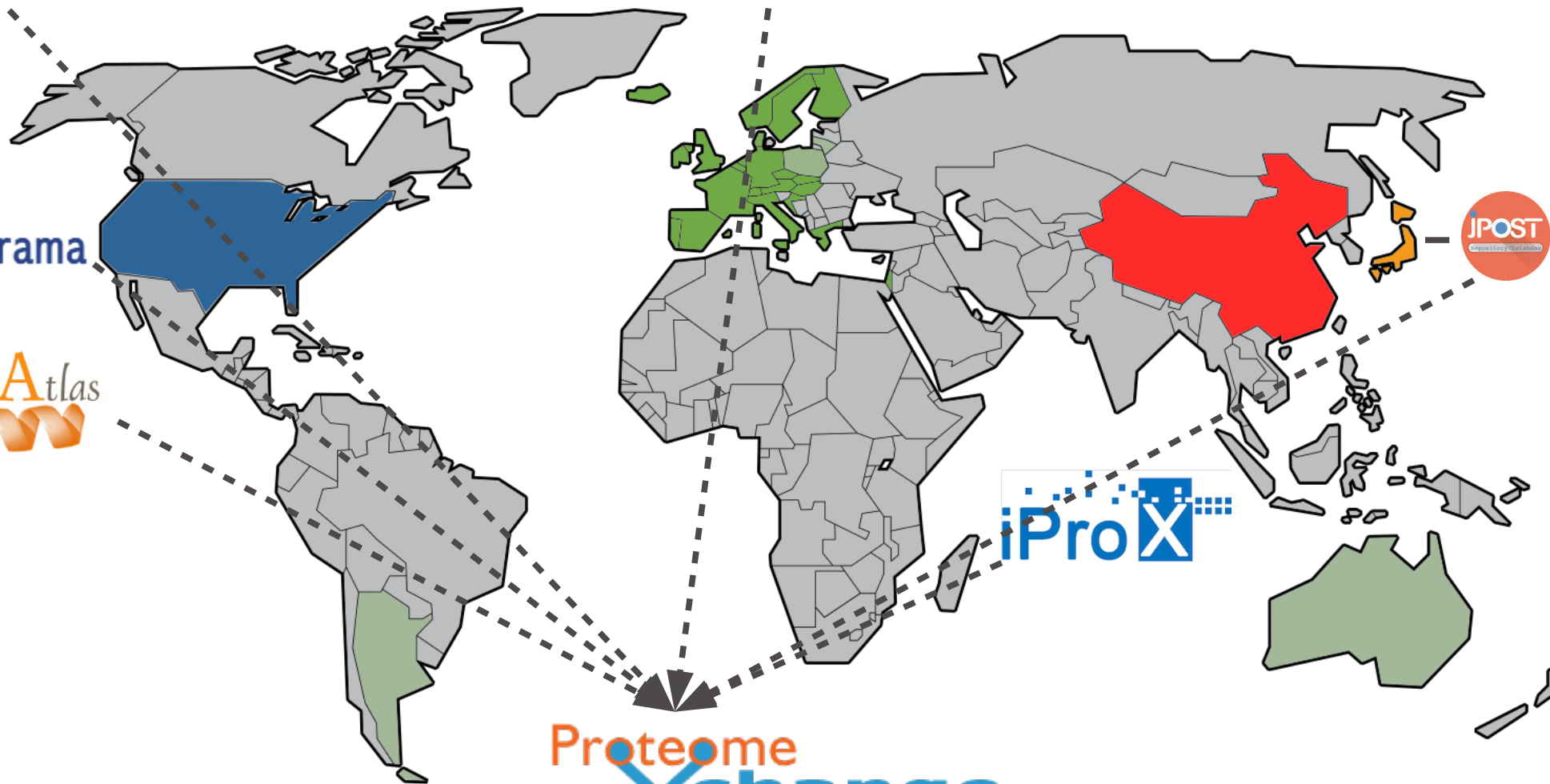
Mass Spectrometry
Interactive Virtual Environment



Panorama



**Proteome
Xchange**



Where should I deposit my data?

- Free of charge



Where should I deposit my data?

- Free of charge
- Quick upload (Aspera)



Where should I deposit my data?

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



Where should I deposit my data?

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




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 →)



Where should I deposit my data?

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



Where should I deposit my data?



Where should I deposit my data?



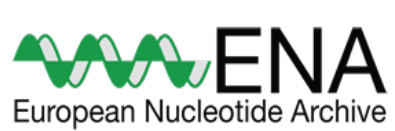
Where should I deposit my data?



Where should I deposit my data?



Where should I deposit my data?



Where should I deposit my data?

 **BioStudies.**

 **ENA**
European Nucleotide Archive

 **ArrayExpress**

 **BioImage Archive**

 **EMPIAR**  **IDR**

 **European Variation Archive**

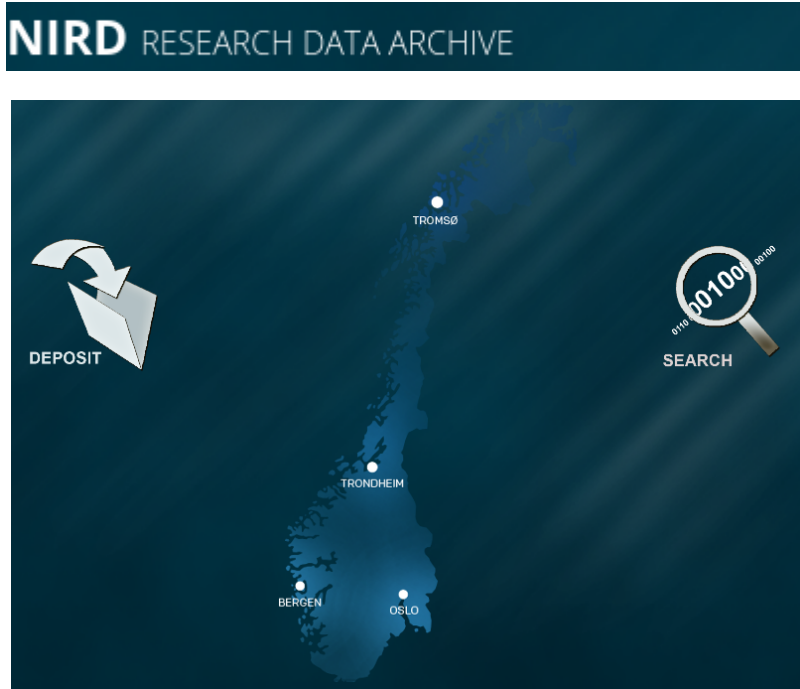
 **PRIDE**
PRoteomics IDentifications Database

 **BioModels**

 **PDBe**
Protein Data Bank in Europe

 **MetaboLights**

NIRD archive



National, free depositing repository

Domain agnostic

Dublin-core metadata standard

DOI accessible

Supports machine readable metadata harvesting

Generic archives



Less standard metadata → Re-usability ↓↓



Domain agnostic → Finability ↓↓

Guided submission process



Commercial actors: Size limitations, ...

What about sensitive data?

Central metadata accessibility

Secure storage

Implemented data access committees

Local EGA in development



EUROPEAN
GENOME-PHENOME
ARCHIVE



NSD archive

NSD NORSK SENTER FOR FORSKNINGSDATA

Domain agnostic

DOI accessible

Implemented data access committees

