A closer look at the repositories' world

Bruna Piereck

15th June 2023, Gent

Do we know why?

- Why submit data?
- Why should we be aware of the repositories?

A (somewhat) simple start

Alice is a researcher who:

- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal

A (somewhat) simple start

Alice is a researcher who:

- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal

What is written in the DMP?

A (somewhat) simple start

Alice is a researcher who:

- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal

What is written in the DMP?

Are you cleared for depositing?
Ethical and legal issues
Data availability requirements of the funder and of the journal

elixir BELGIUM

How to find a repository?

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
Search... Search

https://www.re3data.org/

elixir

Platform highlights

- ELIXIR Core Data Resources: European data resources that are of fundamental importance to research in the life sciences and are committed to the long-term preservation of data.
- ELIXIR Deposition Databases: repositories recommended for the deposition of life sciences experimental data.
- Data resource services: this list is updated as Nodes finalise or review their Service Delivery Plans (see How countries join).

https://elixir-europe.org/platforms/data

FAIRsharing.org
standards, databases, policies
search through all content
STANDARDS | DATABASES | POLICIES | COLLECTIONS

https://fairsharing.org/

EMBL-EBI home > Services > Data submission

EMBL-EBI

Data submission

Use this data submission wizard to find the right archive for your data in a few simple steps.

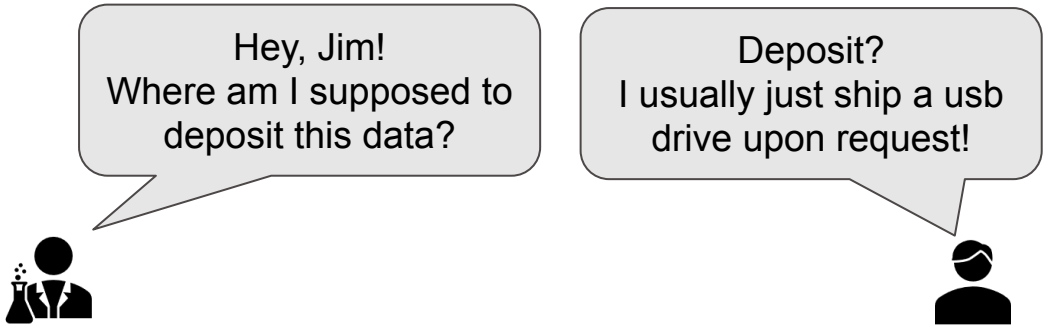
https://www.ebi.ac.uk/submission/

scientific data

Explore content | About the journal | Publish with us

nature > scientific data > policies > data repository guidance

https://www.nature.com/sdata/policies/repositories



elixir BELGIUM

Generic vs. Domain specific

Repository Name	Information on fees/costs	Size limits	Integrated with *Scientific Data*'s manuscript submission system	Re3data / FAIRSharing entry
Dryad Digital Repository	$120 USD for first 20 GB, and $50 USD for each additional 10 GB	None stated	**Yes ✓**	view FAIRsharing entry
figshare	100 GB free per *Scientific Data* manuscript.	1 TB per dataset	**Yes ✓** - To qualify for the 100 GB of free storage, data must be uploaded to figshare via our submission system. Download instructions.	view FAIRsharing entry
Harvard Dataverse	Contact repository for datasets over 1 TB	2.5 GB per file, 10 GB per dataset	No	view re3data entry
Open Science Framework	Free of charge	5 GB per file, multiple files can be uploaded	No	view FAIRsharing entry
Zenodo	Donations towards sustainability encouraged	50 GB per dataset	No	view re3data entry
Science Data Bank	Free of charge	8 GB per file, no limit to dataset size	No	view FAIRsharing entry

https://www.nature.com/sdata/policies/repositories

- There are a few generalist repositories out there
- Your institution might have a data repository that fulfils requirements from funders and journals
- They might still not match your data type

Where do you find data?

Which repository can best connect you with the broader community?

elixir BELGIUM

Let's find a repository

Alice is a researcher who:

- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

https://www.re3data.org/

FAIRsharing

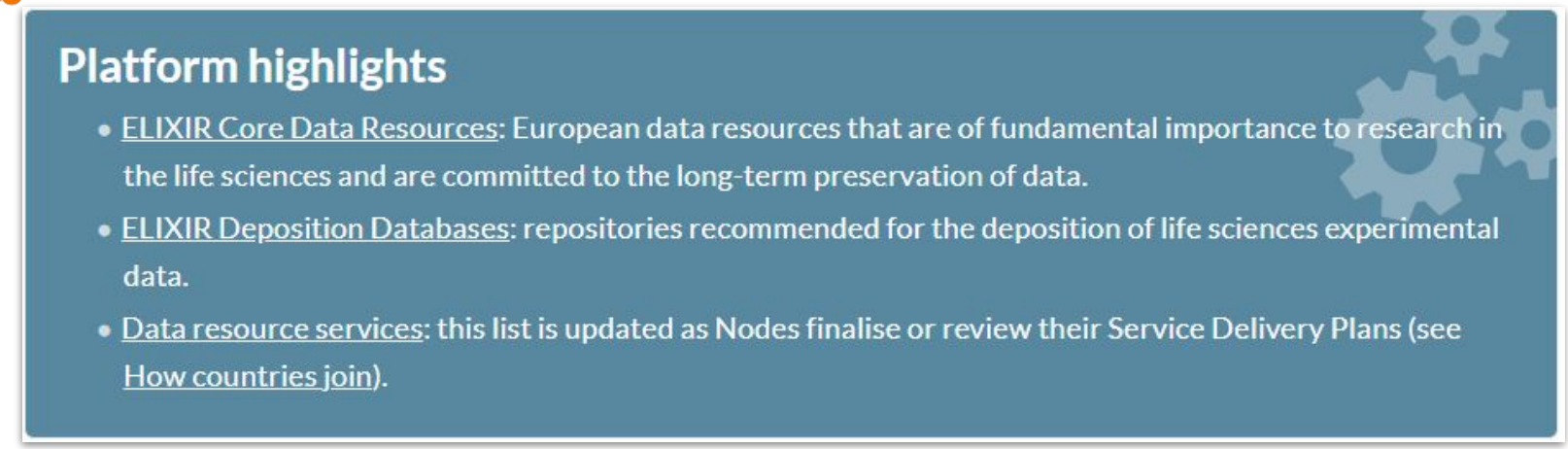
https://fairsharing.org/

Identify a couple of potential repositories using re3data and FAIRsharing

- Domain specific
- Justify the choice
- What information can you find about the repository that might help you?

elixir BELGIUM

Other ways to find a repository



https://elixir-europe.org/platforms/data

Focus on Life Sciences

The ELIXIR recommended deposition databases

ELIXIR Deposition Database list

Deposition Database	Data type	International collaboration framework [1]
ArrayExpress	Functional genomics data. Stores data from high-throughput functional genomics experiments.	
BioModels	Computational models of biological processes.	
BioSamples	BioSamples stores and supplies descriptions and metadata about biological samples used in research and development by academia and industry.	NCBI BioSamples database
BioStudies	Descriptions of biological studies, links to data from these studies in other databases, as well as data that do not fit in the structured archives.	
EGA	Personally identifiable genetic and phenotypic data resulting from biomedical research projects.	European Bioinformatics Institute and the Centre for Genomic Regulation
EMDB	The Electron Microscopy Data Bank is a public repository for electron microscopy density maps of macromolecular complexes and subcellular structures.	
ENA	Nucleotide sequence information, covering raw sequencing data, contextual data, sequence assembly information and functional and taxonomic annotation.	International Nucleotide Sequence Database Collaboration
IntAct	IntAct provides a freely available, open source database system and analysis tools for molecular interaction data.	The International Molecular Exchange Consortium
MetaboLights	Metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments.	
PDBe	Biological macromolecular structures.	wwPDB
PRIDE	Mass spectrometry-based proteomics data, including peptide and protein expression information (identifications and quantification values) and the supporting mass spectra evidence.	The ProteomeXchange Consortium

Further information: Gavin Farrell (gavin.farrell@elixir-europe.org)

[1] An International collaboration framework enables content sharing on a formal level. This is often signified by a shared Accession Number system, such that data deposited in one database becomes part of the shared data collection, and is also available through other partner portals.

https://elixir-europe.org/platforms/data/elixir-deposition-databases

Focus on experimental data

elixir BELGIUM

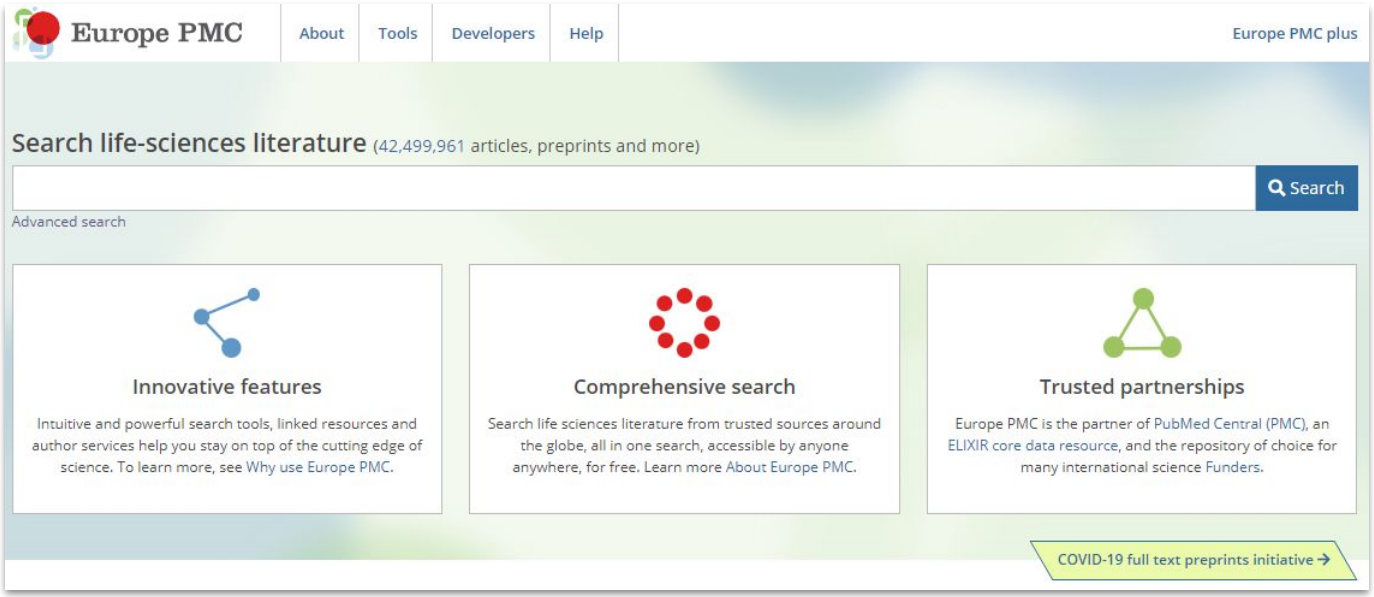
Why is it a good idea to deposit your data?

elixir BELGIUM

A short side note on visibility

Why is it a good idea to deposit your data?

Core Data Resource	Data type
Europe PMC	Europe PMC is a repository, providing access to worldwide life sciences articles, books, patents and clinical guidelines.



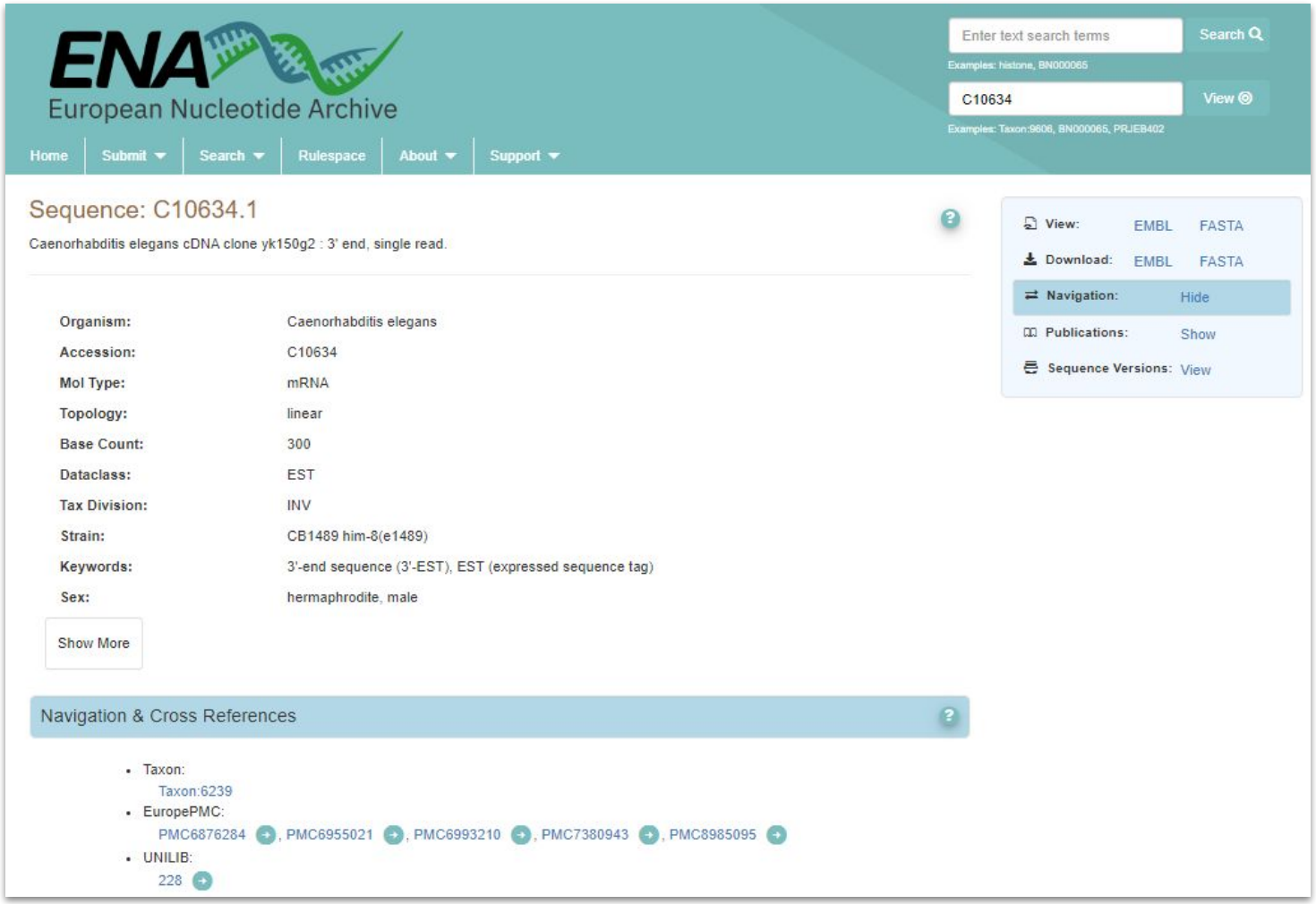
Data citations or links search

Data citations are mentions of accession numbers (unique identifiers for data) in the text of articles. Data citations are available for data in a number of different databases including: European Nucleotide Archive, UniProt, PDBe, OMIM, RefSNP, RefSeq, Pfam, InterPro, and Ensembl. Data links are citations of an article by a database record. Data links from a specific data record to a Europe PMC article are provided by a number of life science databases. The databases included are listed in sections 2.7 and 2.8 of the Search syntax reference.

https://europepmc.org/Help#searchdata

Cross referencing resources

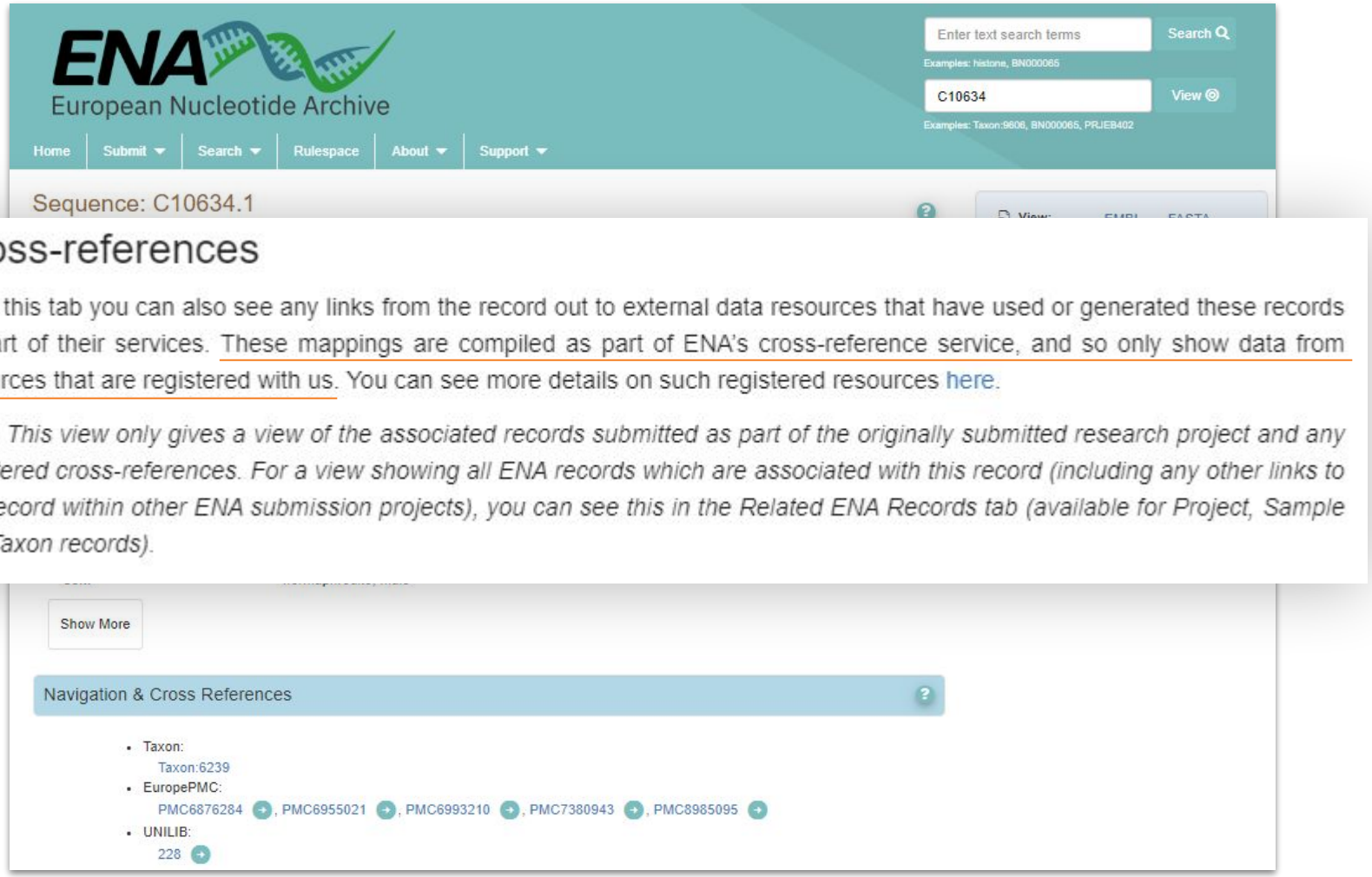
Why is it a good idea to deposit your data?



elixir BELGIUM

Cross referencing resources

Why is it a good idea to deposit your data?



elixir BELGIUM

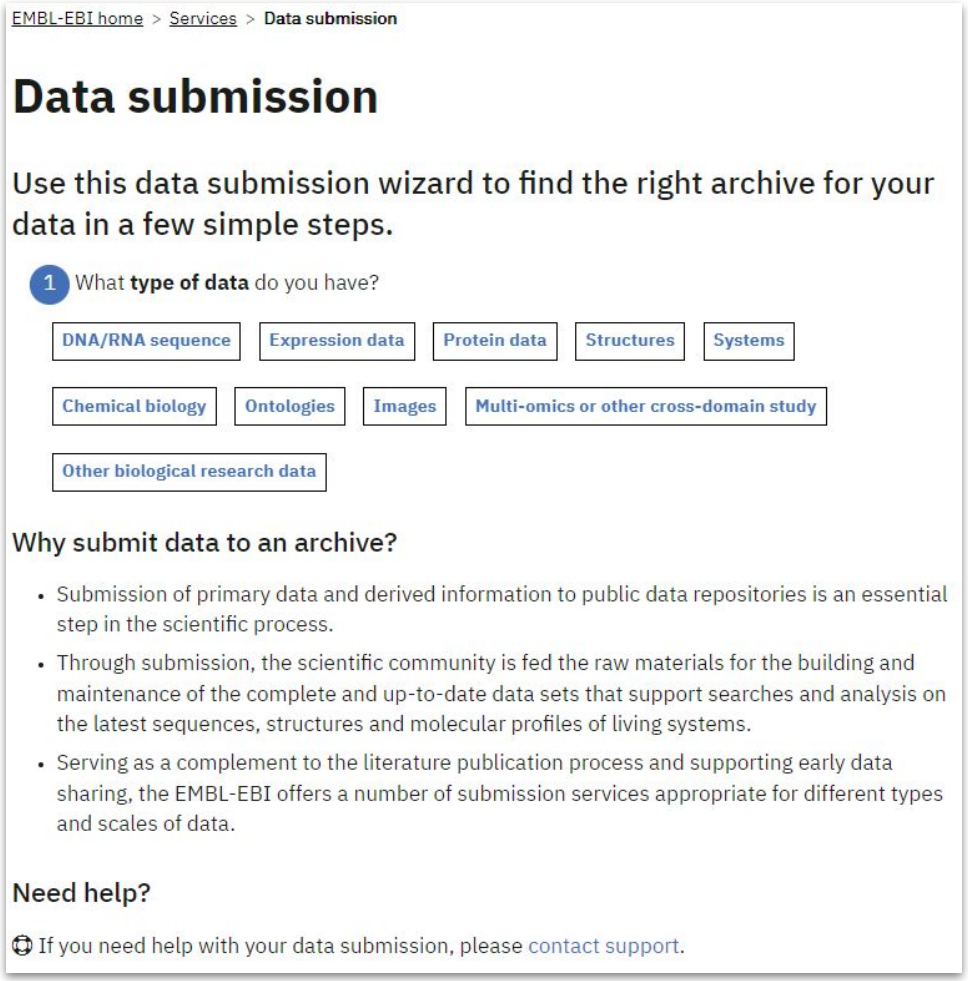
Many ways to find

How do I know which data goes where?
When should I consider it?

elixir BELGIUM

A wizard can be of help

EMBL-EBI



EMBL-EBI home > Services > Data submission

Data submission

Use this data submission wizard to find the right archive for your data in a few simple steps.

1 What **type of data** do you have?

DNA/RNA sequence | Expression data | Protein data | Structures | Systems

Chemical biology | Ontologies | Images | Multi-omics or other cross-domain study

Other biological research data

Why submit data to an archive?

- Submission of primary data and derived information to public data repositories is an essential step in the scientific process.
- Through submission, the scientific community is fed the raw materials for the building and maintenance of the complete and up-to-date data sets that support searches and analysis on the latest sequences, structures and molecular profiles of living systems.
- Serving as a complement to the literature publication process and supporting early data sharing, the EMBL-EBI offers a number of submission services appropriate for different types and scales of data.

Need help?

If you need help with your data submission, please contact support.

https://www.ebi.ac.uk/submission/

All EMBL-EBI data repositories

Array Express: functional genomics data

BioImage Archive: bioimaging data

BioModels: computational models

BioSamples: reference sample data

BioStudies: biological research data

ChEBI: chemical entities

DGVa: structural genetic variation data

EFO: experimental variables

EGA: human data that requires controlled access

EMPIAR: raw image data

ENA: nucleotide sequence data

EVA: genetic variation data

GO: Gene ontology annotations

GWAS Catalog: Genome-wide association study data

IntAct: molecular interactions

IntEnz: enzyme nomenclature

MetaboLights: metabolomics data

Metagenomics: raw sequence data & associated meta-data

OneDep: electron microscopy, X-ray crystallography & NMR data

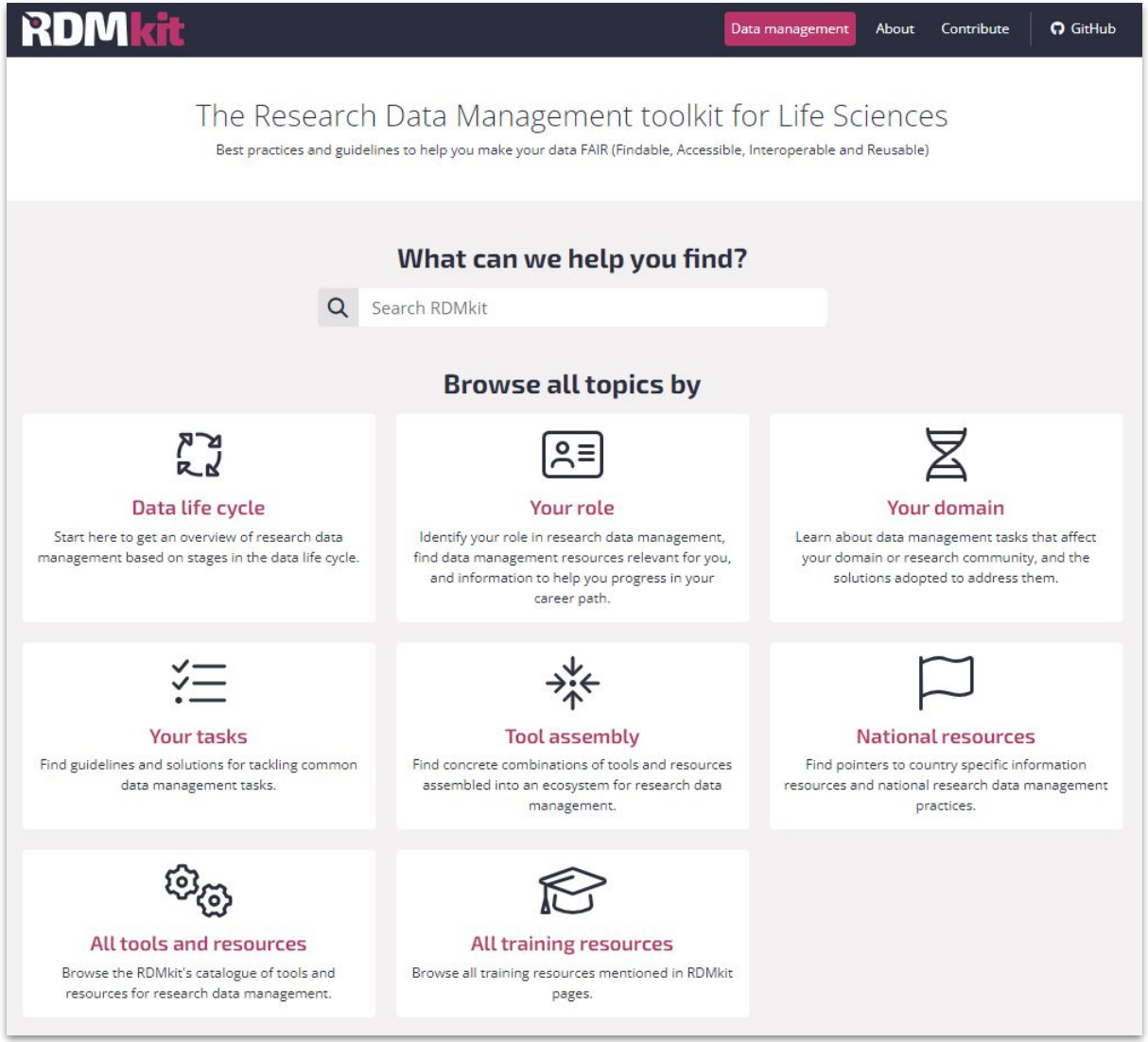
PRIDE: protein & peptide identification data

UniProt SPIN: protein sequences & annotations

UniProt: updates or corrections

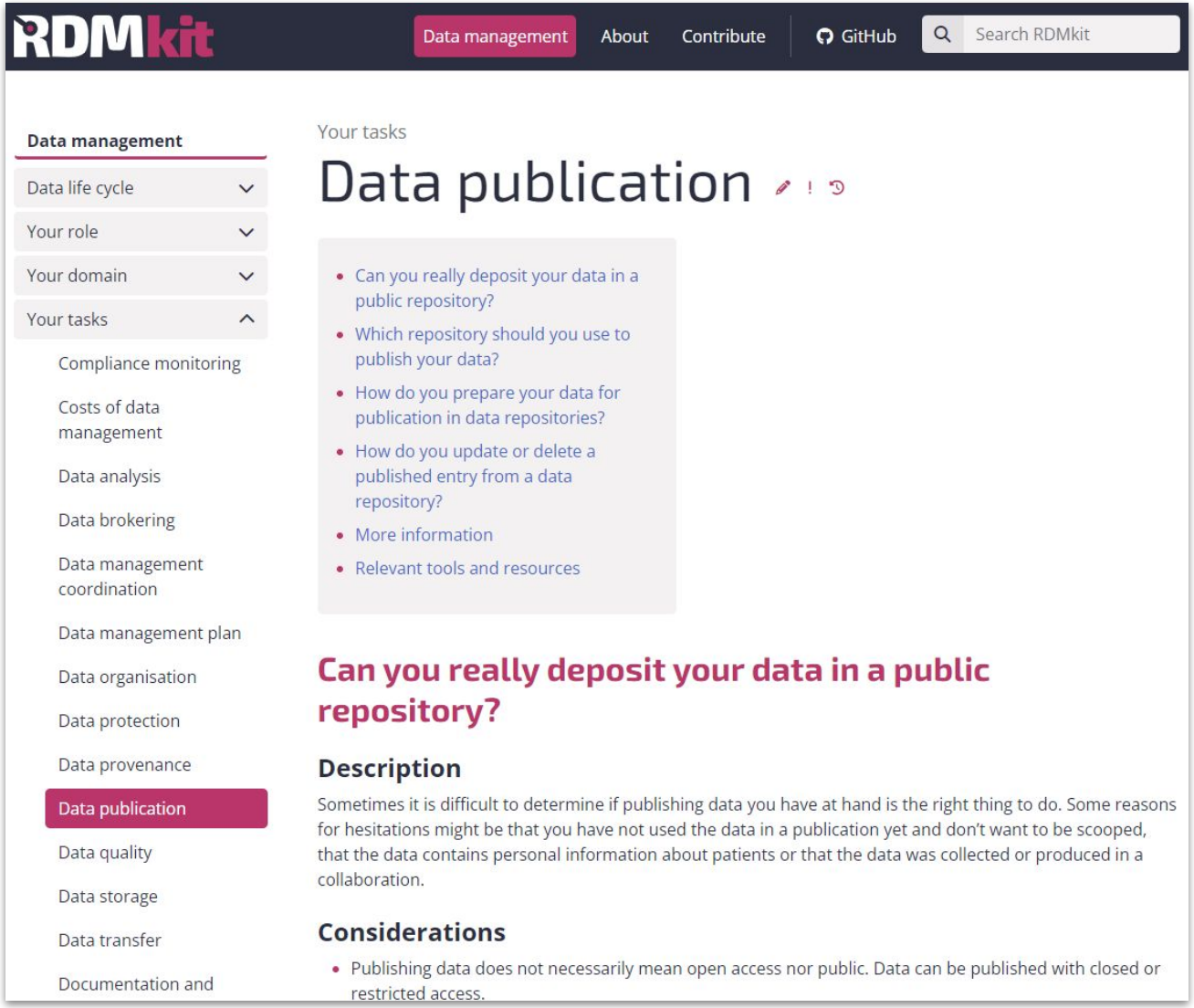
Other ways to find a repository

Community driven content



https://rdmkit.elixir-europe.org

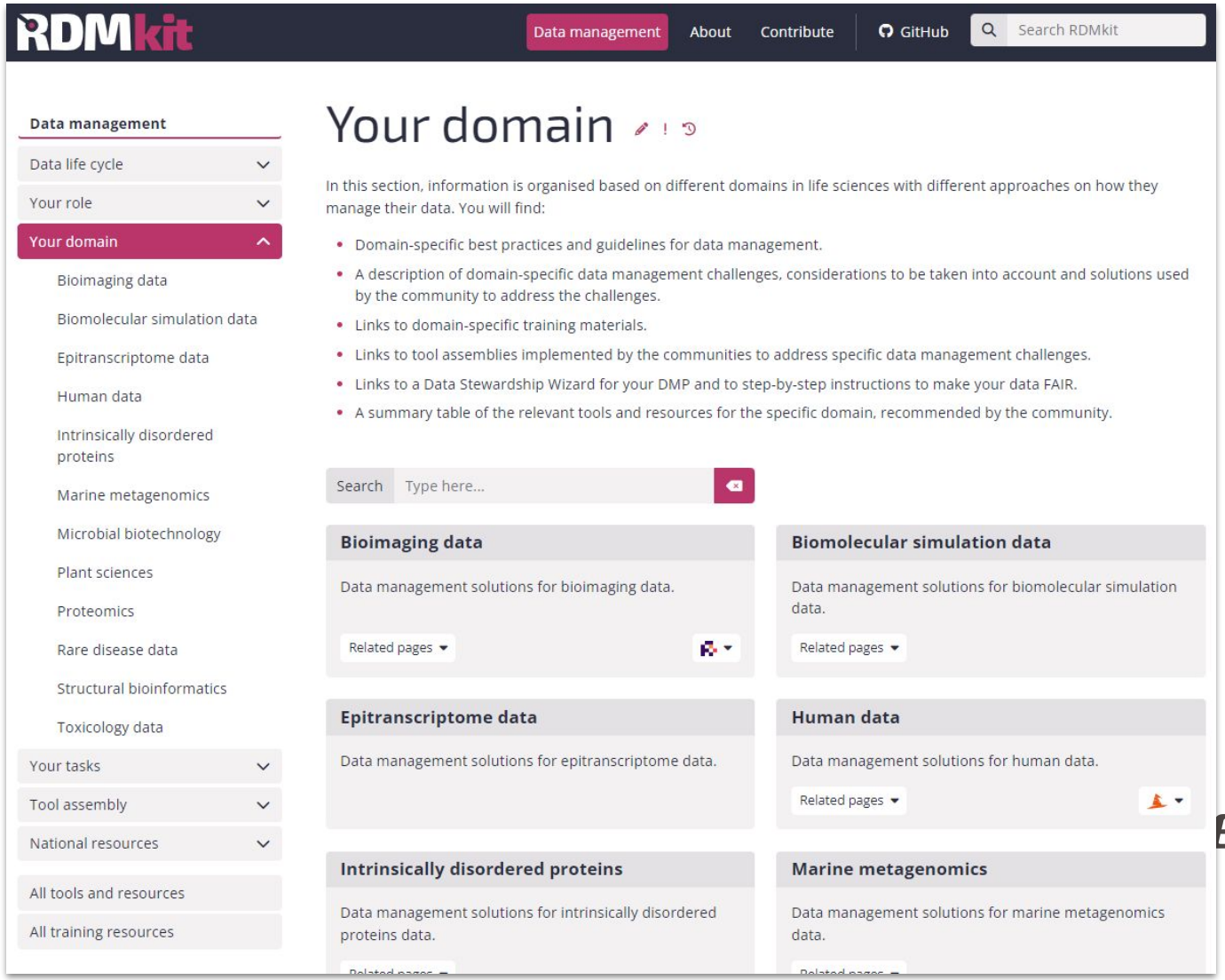
Other ways to find a repository



https://rdmkit.elixir-europe.org/data_publication

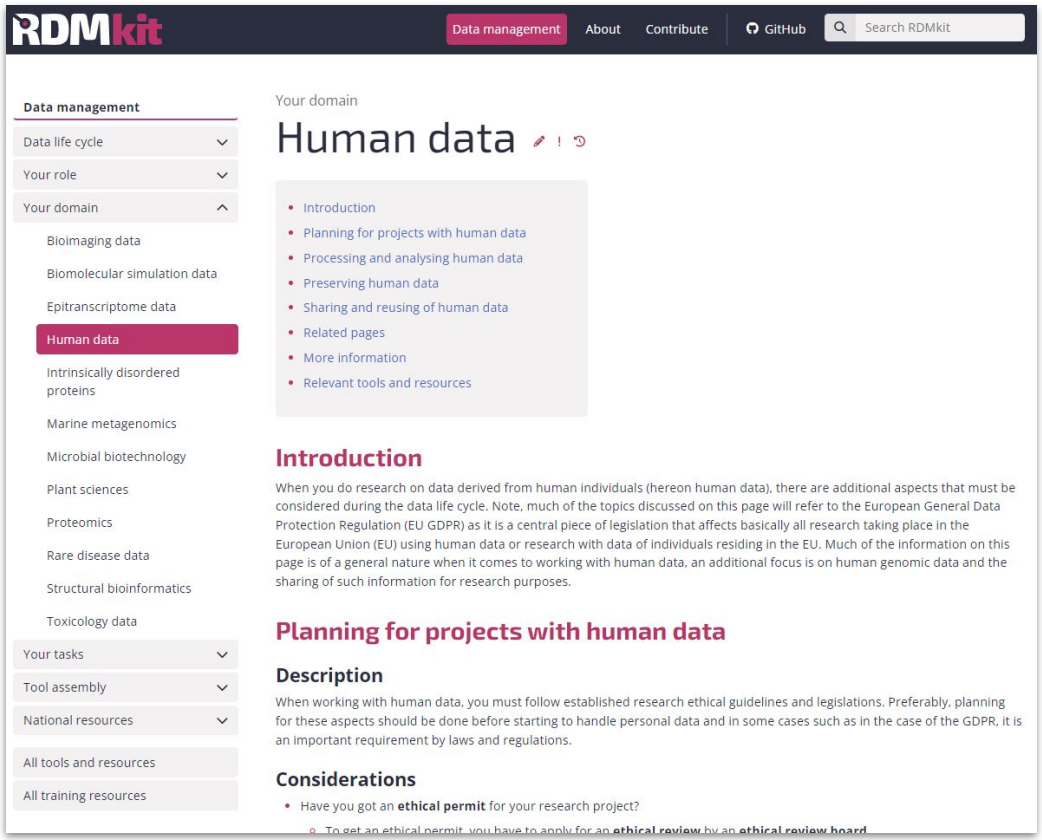
Other ways to find a repository

Are there recommendations for the domain?



https://rdmkit.elixir-europe.org

Other ways to find a repository



https://rdmkit.elixir-europe.org/human_data

Solutions

- The European Genome-phenome Archive (EGA) is an international service for secure archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical studies and healthcare centres. All services are free of charge. The EGA platform offers secure and European law-compliant data sharing. Data treatment is FAIR-compliant, thus data is discoverable in the EGA website and shareable with other researchers through authorisation and authentication protocols. The right to allow access to any dataset belongs to the Data controllers (and not to the EGA), who are responsible to sign a Data Access Agreement (DAA) with researchers requesting access to their data. Templates of the legal documents are provided. The EGA hosts data from all around the world and distributes it where and when the data controllers permit.
- dbGAP and JGA are other international data repositories, based in the USA and Japan respectively, that adopt a controlled-access model based on their national regulations. Due to European GDPR specific requirements, it may not be possible to deposit EU subjects' data to these repositories.
- The GA4GH Beacon project is a GA4GH initiative that enables genomic and clinical data sharing across federated networks. A Beacon is defined as a web-accessible service that can be queried for information about a specific allele with no reference to a specific sample or patient, thereby reducing privacy risks.
- The GA4GH Data Use Ontology DUO is an international standard, which provides codes to represent data use restrictions for controlled access datasets.
- Crypt4gh is a Python tool to encrypt, decrypt or re-encrypt files, according to the GA4GH encryption file format.

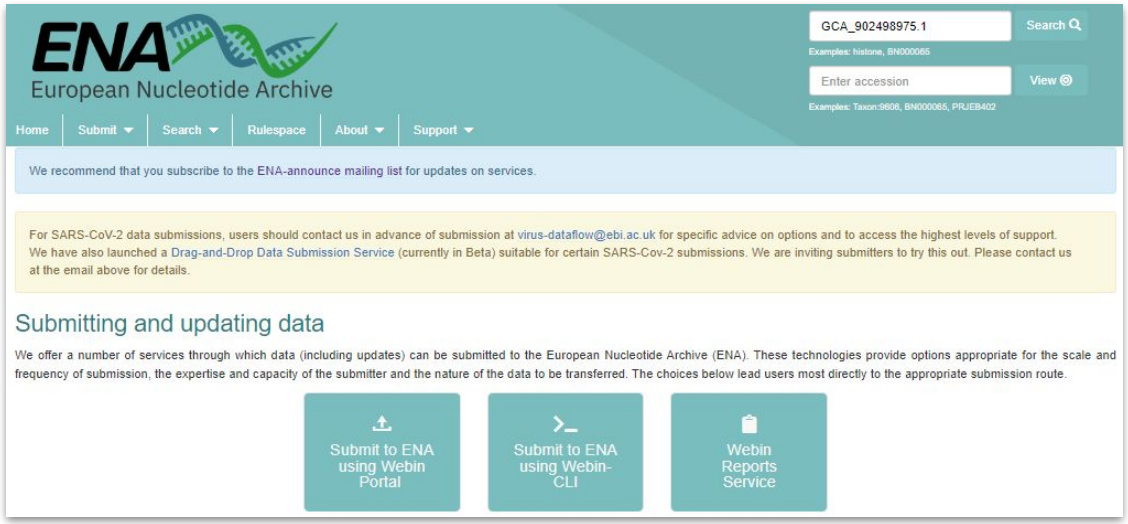
Relevant tools and resources

Tool or resource	Description	Related pages	Registry
BBMRI-ERIC's ELSI Knowledge Base	The ELSI Knowledge Base is an open-access resource platform that aims at providing practical know-how for responsible research.	Data protection, Data sensitivity, Data Steward: policy, Data Steward: research	
Beacon	The Beacon protocol defines an open standard for genomics data discovery.	Researcher, Data Steward: research, Data Steward: infrastructure	Tool info, Standards/Databases, Training
BIONDA	BIONDA is a free and open-access biomarker database, which employs various text mining	Data storage, Researcher, Proteomics	Tool info

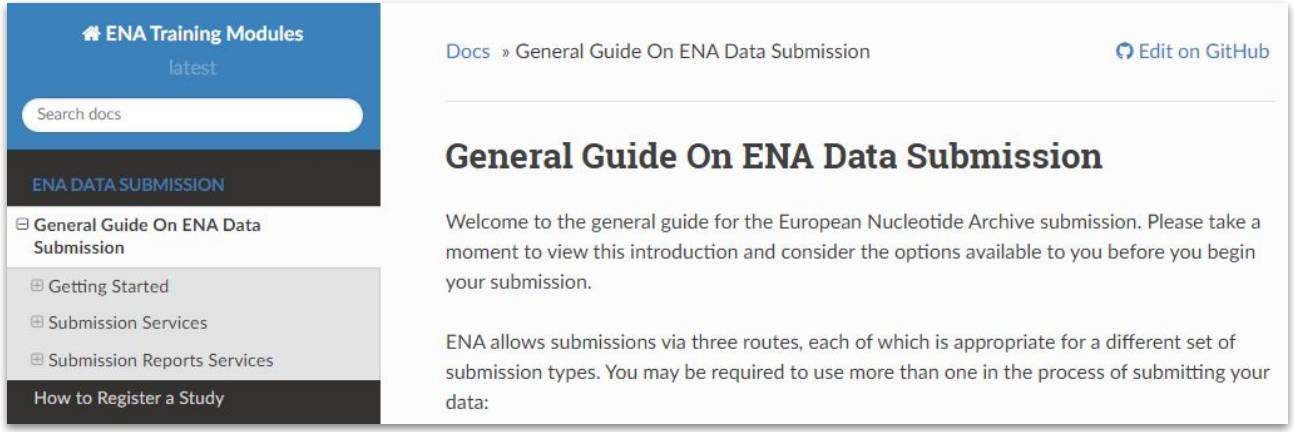
Data submission

Different repositories = different routes

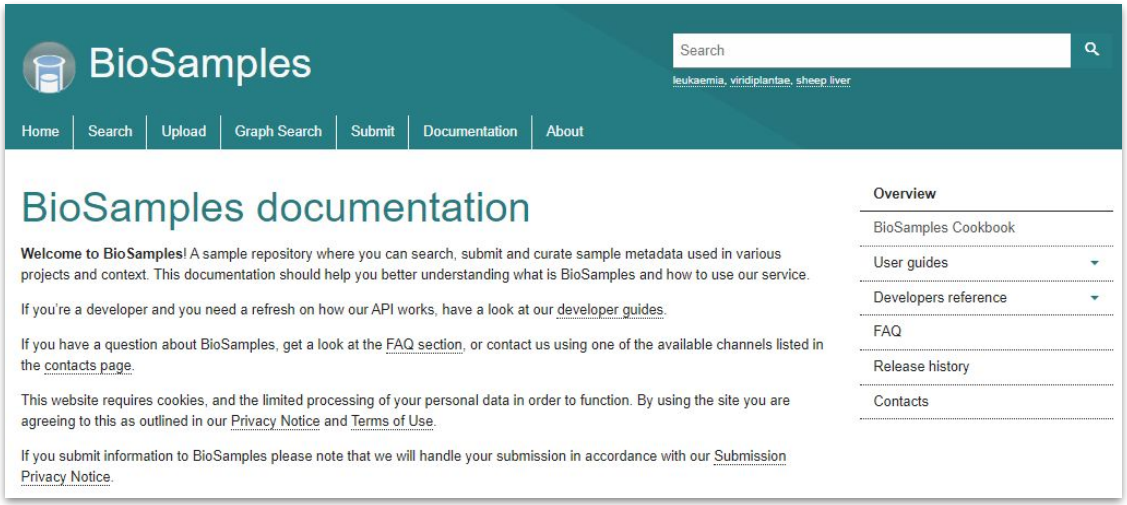
- What file formats are expected?
- What metadata is expected?
- Can ontologies and controlled vocabularies be used?
- Are there costs involved?
- Is there a license to be applied? Is it implicit?



https://www.ebi.ac.uk/ena/browser/submit



https://ena-docs.readthedocs.io/en/latest/submit/general-guide.html



https://www.ebi.ac.uk/biosamples/docs

elixir BELGIUM

Let's find a repository

Alice is a researcher who:

- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal

RDMkit

https://rdmkit.elixir-europe.org

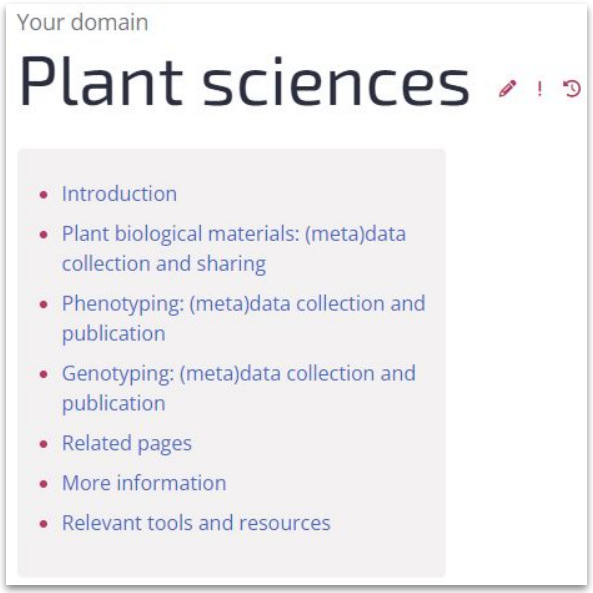
Identify a repository based on information from RDMkit

- Domain specific
- Justify the choice
- Bonus: can you find further resources that can help you?

elixir BELGIUM

Other ways to find a repository

RDMkit



Genotyping: (meta)data collection and publication

Description

Here are described the mandatory, recommended and optional metadata fields for data interoperability and re-use, as well as for data deposition in EVA (European Variation Archive), the EMBL-EBI's open-access genetic variation archive connected to BioSamples, described above.

Considerations

- Did you collect the metadata for the identification of your plant samples according to the recommendations provided in the above section?
- Is the reference genome assembly available in an INSDC archive and has a Genome Collections Accession number, either GCA or GCF?
- Is the analytic approach used for creating the VCF file available in a publication and has a Digital Object Identifier (DOI)?

Solutions

Checklists, ontologies and file formats

- Sharing plant genotyping data files involves the use of the Variant Call Format (VCF) standard.
- Findability and reusability of VCF files depends on the supplied metadata and in particular with MIAPPE compliant biological material description: the plant genomic and genetic variation data submission recipe helps you on that topic.

Data sharing and publication

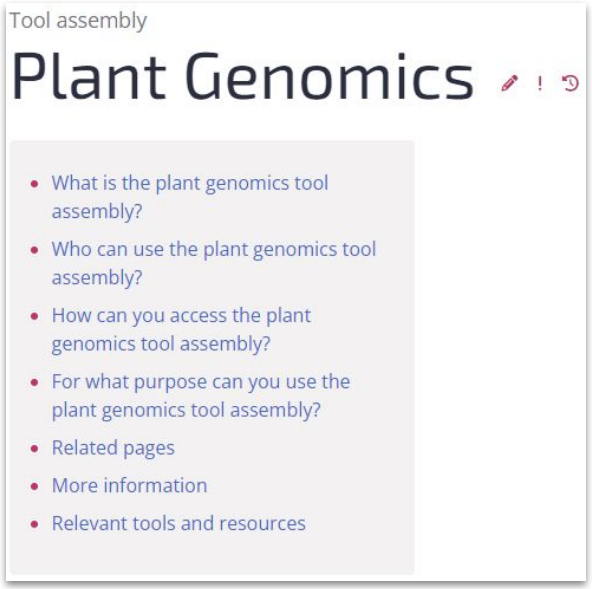
- Once the VCF file is ready with all necessary metadata, it can be submitted to the European Variation Archive (EVA). You will find all necessary information on the submission steps on the EVA submission page.

https://rdmkit.elixir-europe.org/plant_sciences

elixir BELGIUM

Other ways to find a repository

RDMkit


Tool assembly

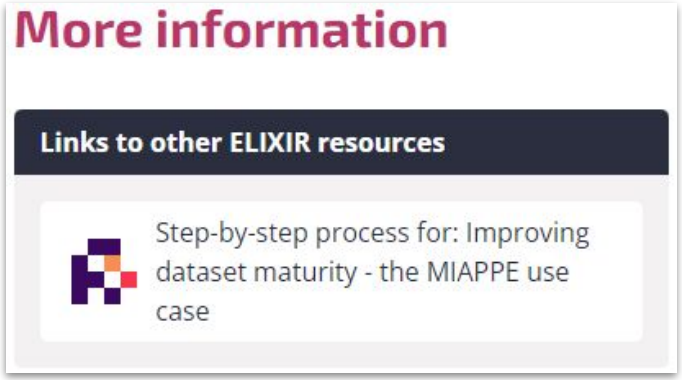
Plant Genomics

- What is the plant genomics tool assembly?
- Who can use the plant genomics tool assembly?
- How can you access the plant genomics tool assembly?
- For what purpose can you use the plant genomics tool assembly?
- Related pages
- More information
- Relevant tools and resources


Data sharing and publishing

All sequencing data collected in plant genotyping experiments should be submitted to ENA together with metadata compliant to the GSC MIxS plant associated checklist. Final results of such studies in the form of VCF files should be submitted to EVA. Additionally, supplemental data complementing these two data types is encouraged to be submitted to e!DAL-PGP or Recherche Data Gouv.

https://rdmkit.elixir-europe.org/plant_genomics_assembly


More information

Links to other ELIXIR resources

Step-by-step process for: Improving dataset maturity - the MIAPPE use case

elixir BELGIUM

Multiple repositories with a similar data type

The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between DDBJ, EMBL-EBI and NCBI.

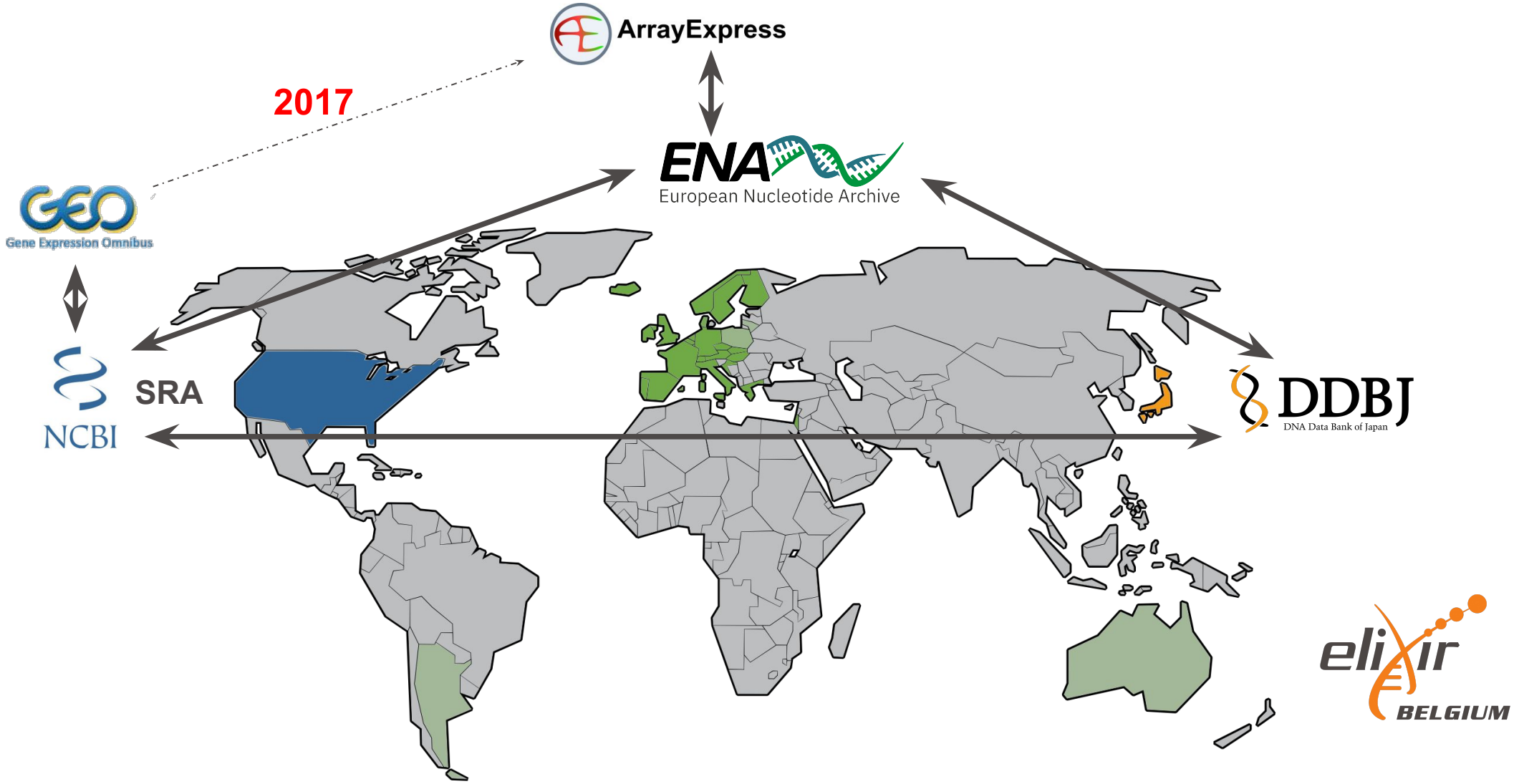
INSDC covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

Databases

Data type	DDBJ	EMBL-EBI	NCBI
Next Generation reads	Sequence Read Archive	European Nucleotide Archive	Sequence Read Archive
Assembled Sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

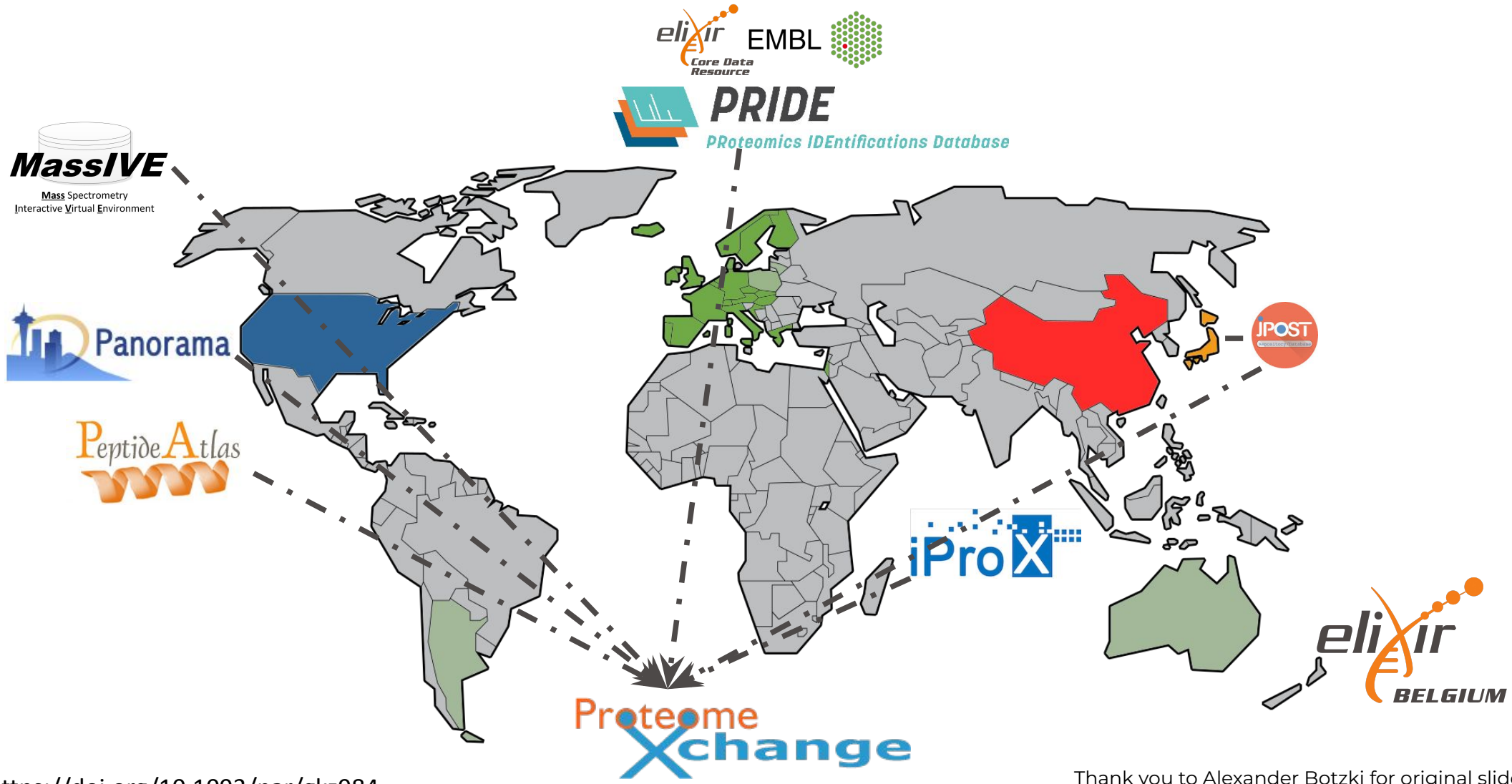
https://www.insdc.org/

Multiple repositories with a similar data type - Genomics



Thank you to Alexander Botzki for original slide material

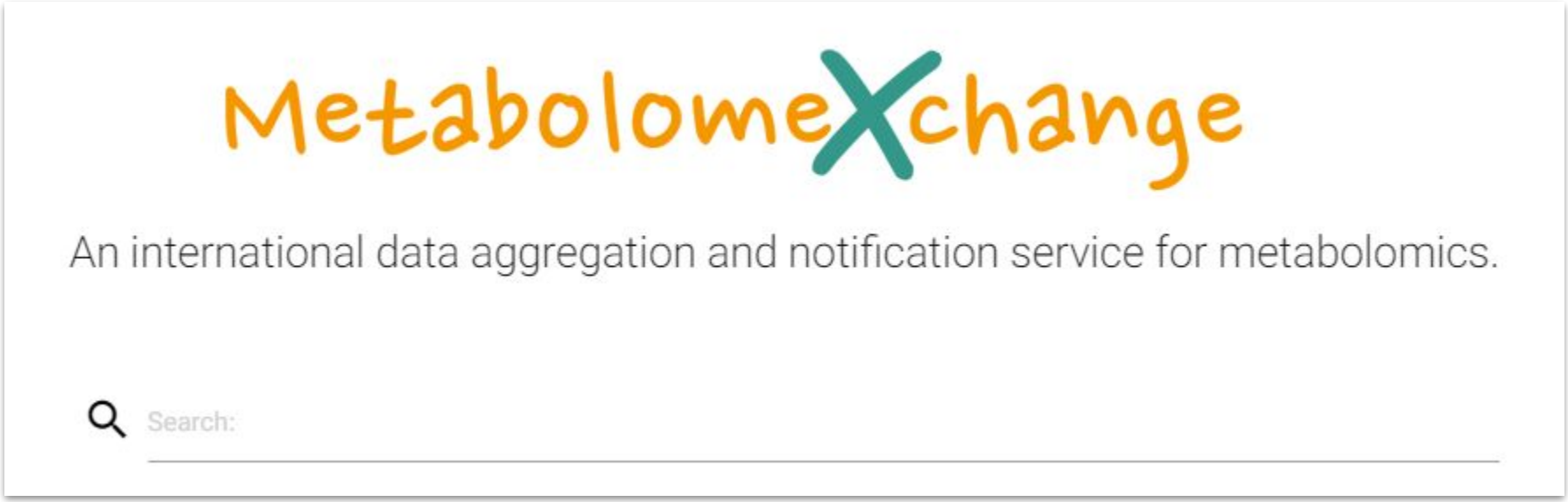
Multiple repositories with a similar data type - Proteomics



https://doi.org/10.1093/nar/gkz984

Thank you to Alexander Botzki for original slide material

Multiple repositories with a similar data type - Metabolomics



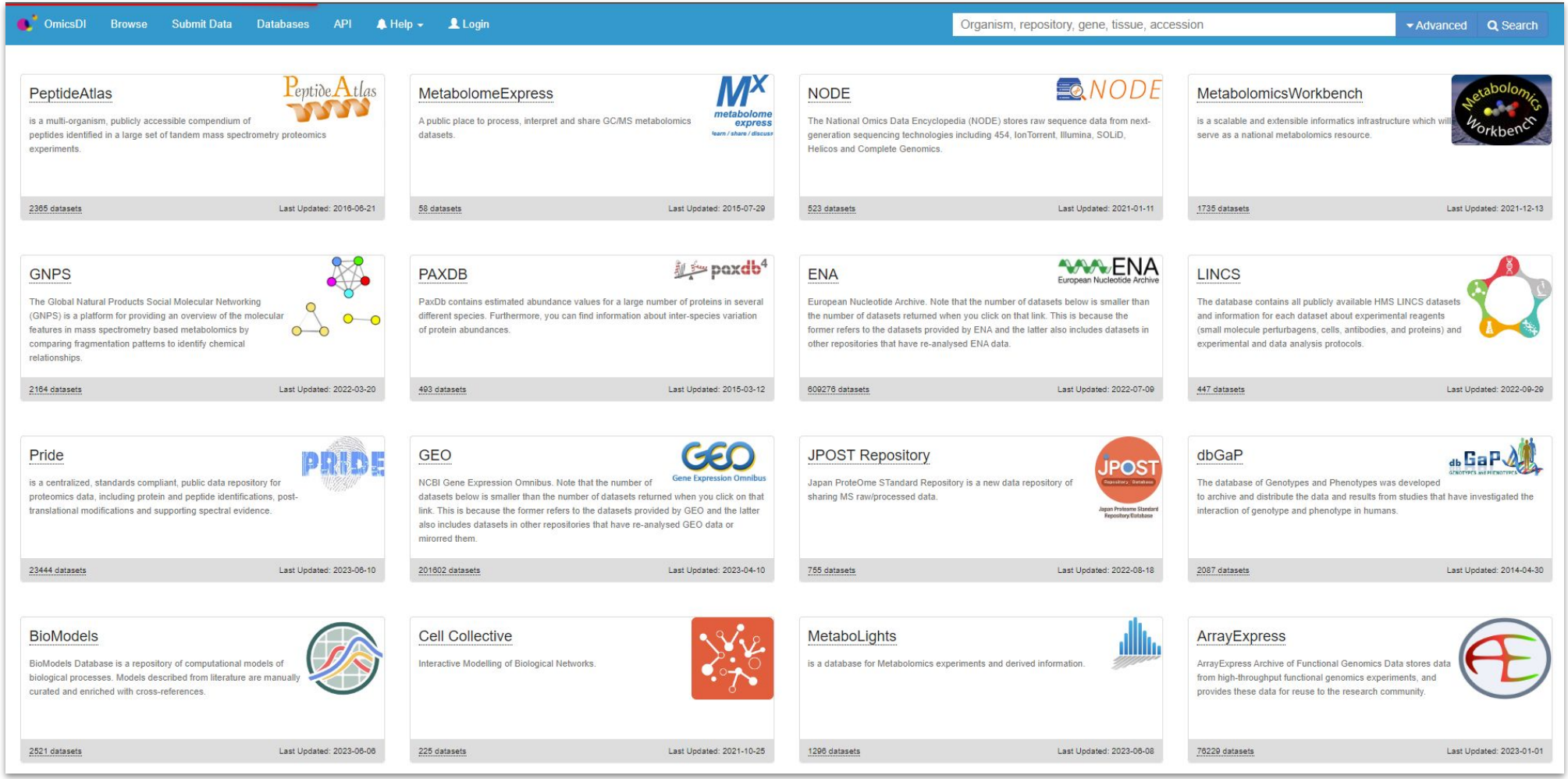
http://www.metabolomexchange.org/

Similar initiatives are being built

Multiple repositories with a similar data type - multi-omics

The Omics Discovery Index - OmicsDI

25 Resources
>3M Datasets
4 Continents



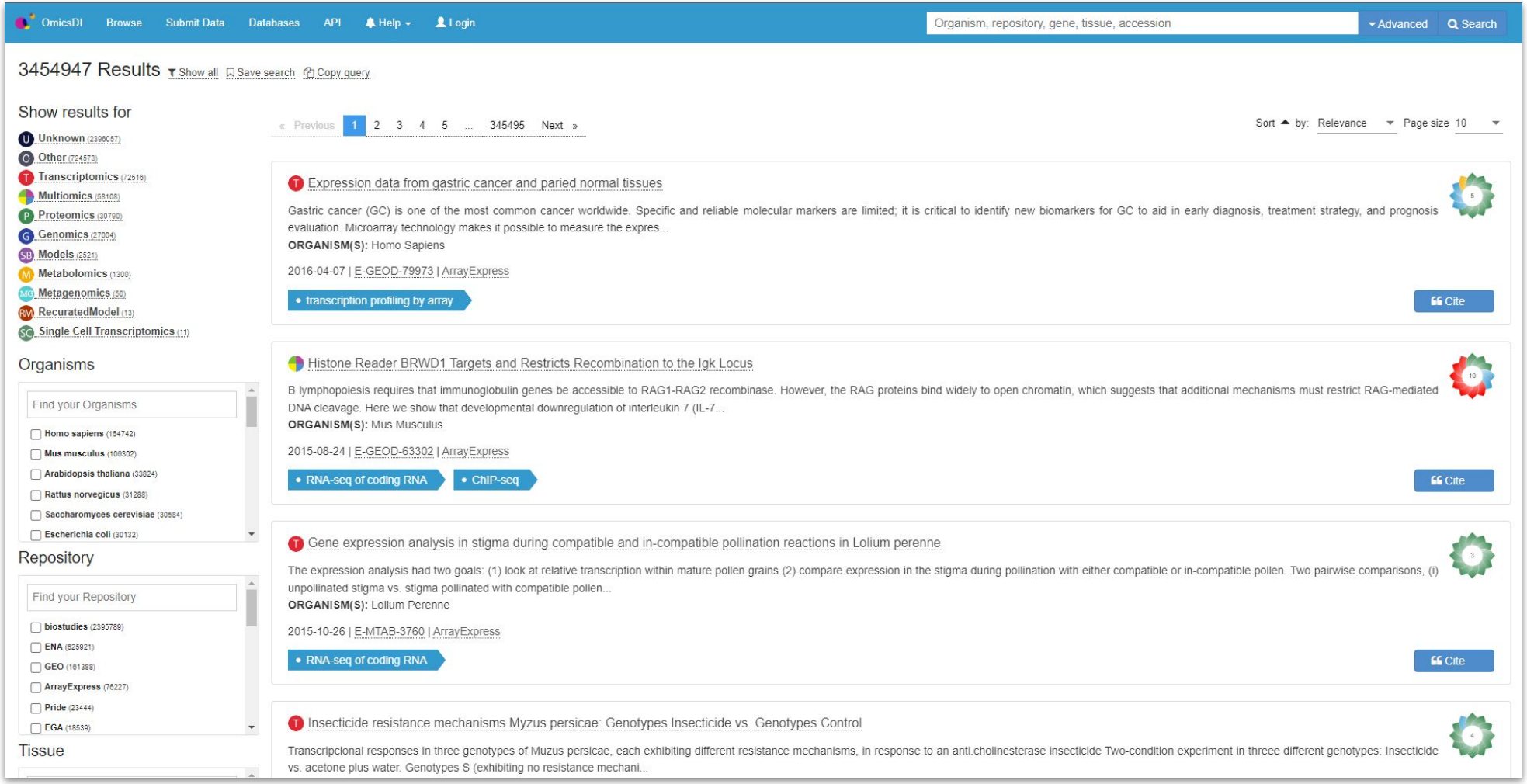
https://www.omicsdi.org/

Multiple repositories with a similar data type - multi-omics

The Omics Discovery Index - OmicsDI

Indexing metadata

The richer the metadata, the easier it is to find a dataset



https://www.omicsdi.org/

A multi-omics project

 Alice is a researcher who:

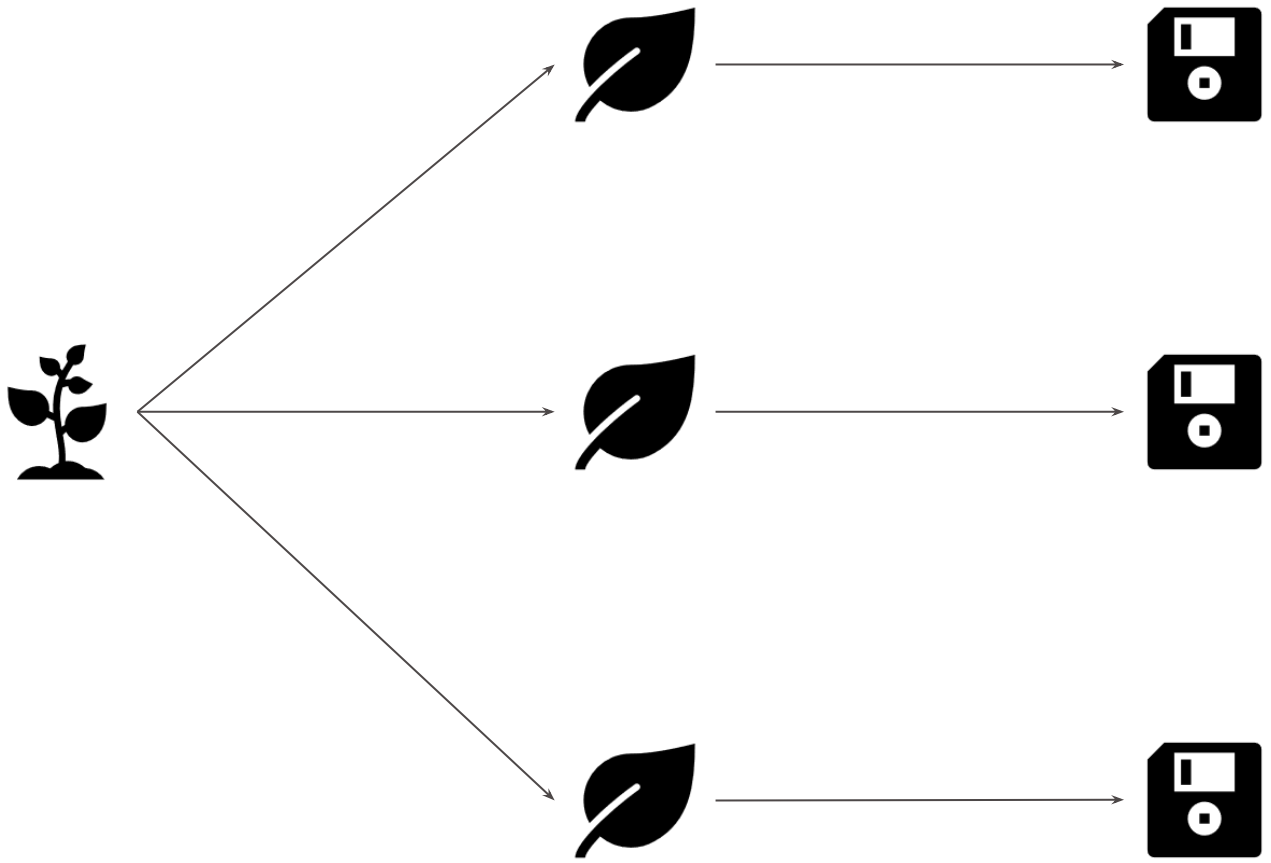
- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal
- **Is also generating proteomics data**
- **Is also generating metabolomics data**
- **Would really like that the datasets are recognized as belonging to the same project**

Identify

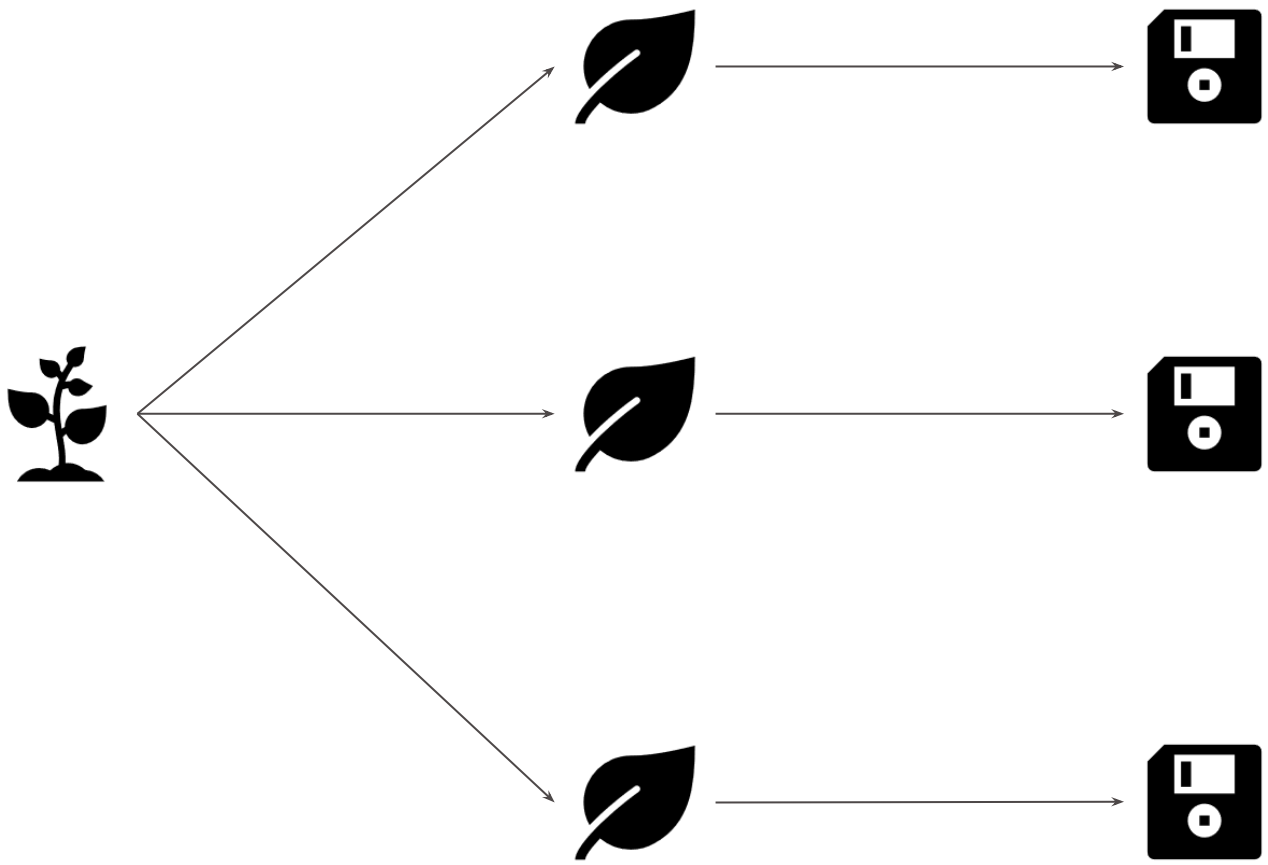
- Suitable, domain specific, repositories
- A strategy (repository) to help connect the datasets

elixir BELGIUM

Multi-omics



Multi-omics

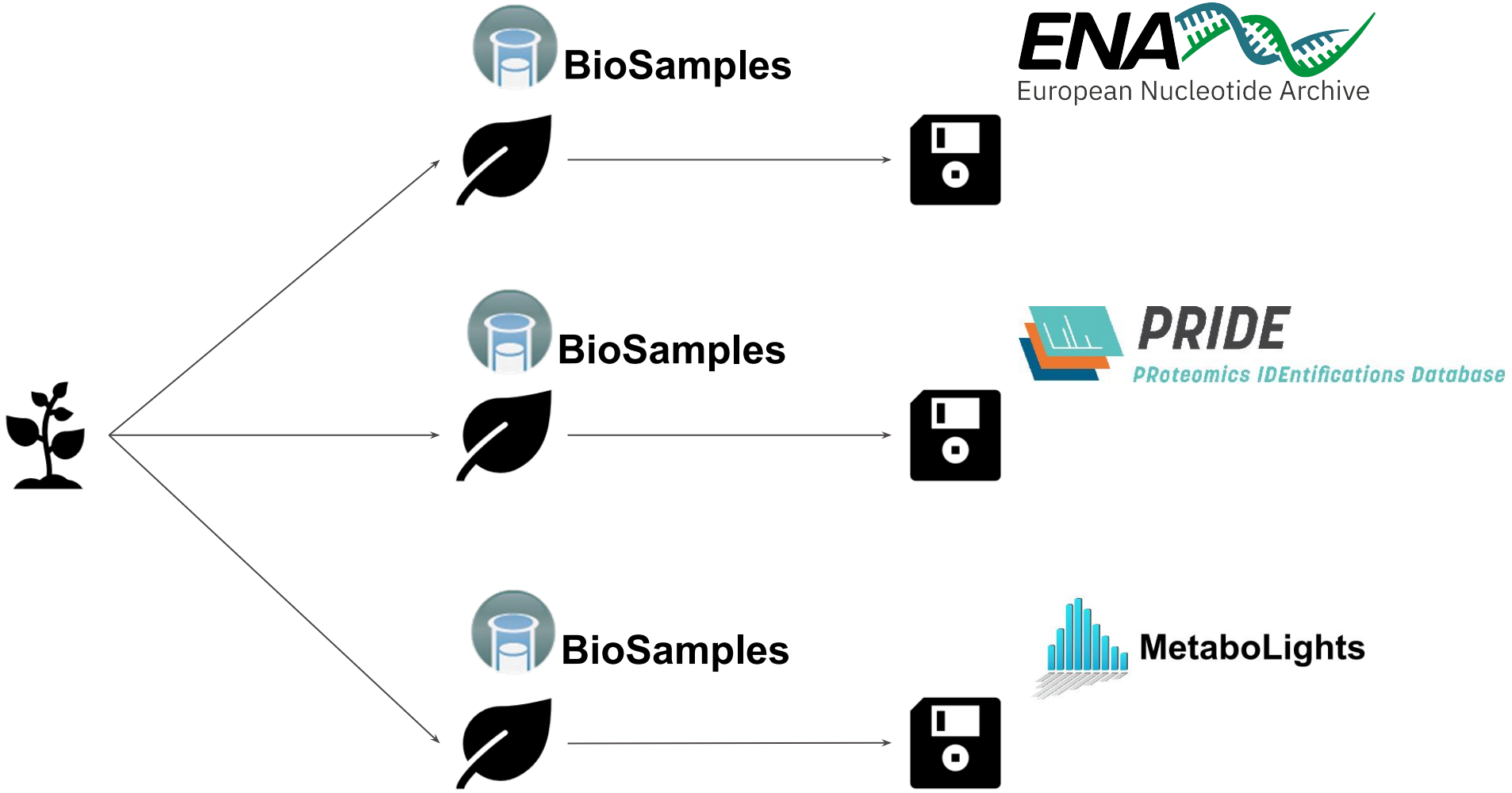


ENA
European Nucleotide Archive

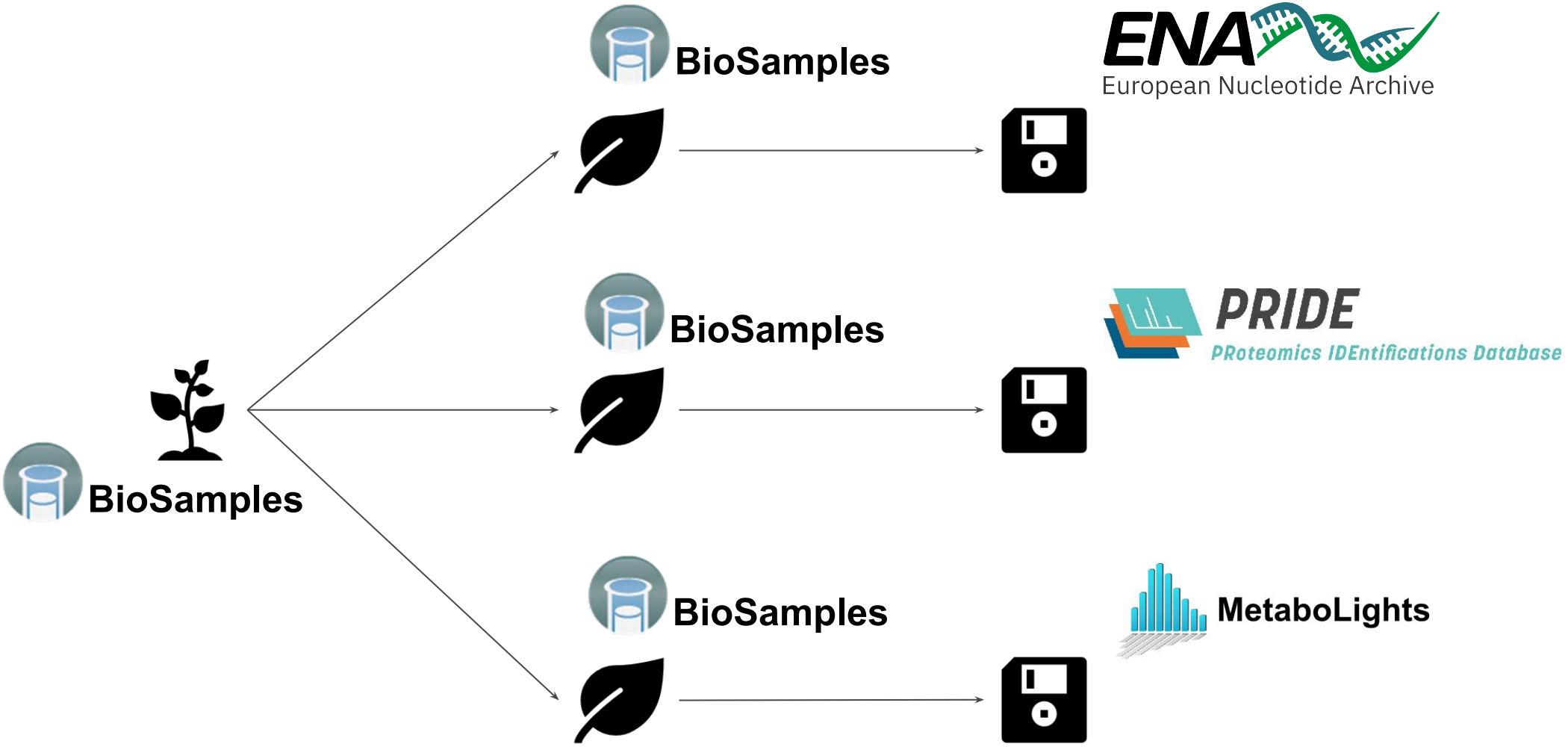
PRIDE
PRoteomics IDEntifications Database

MetaboLights

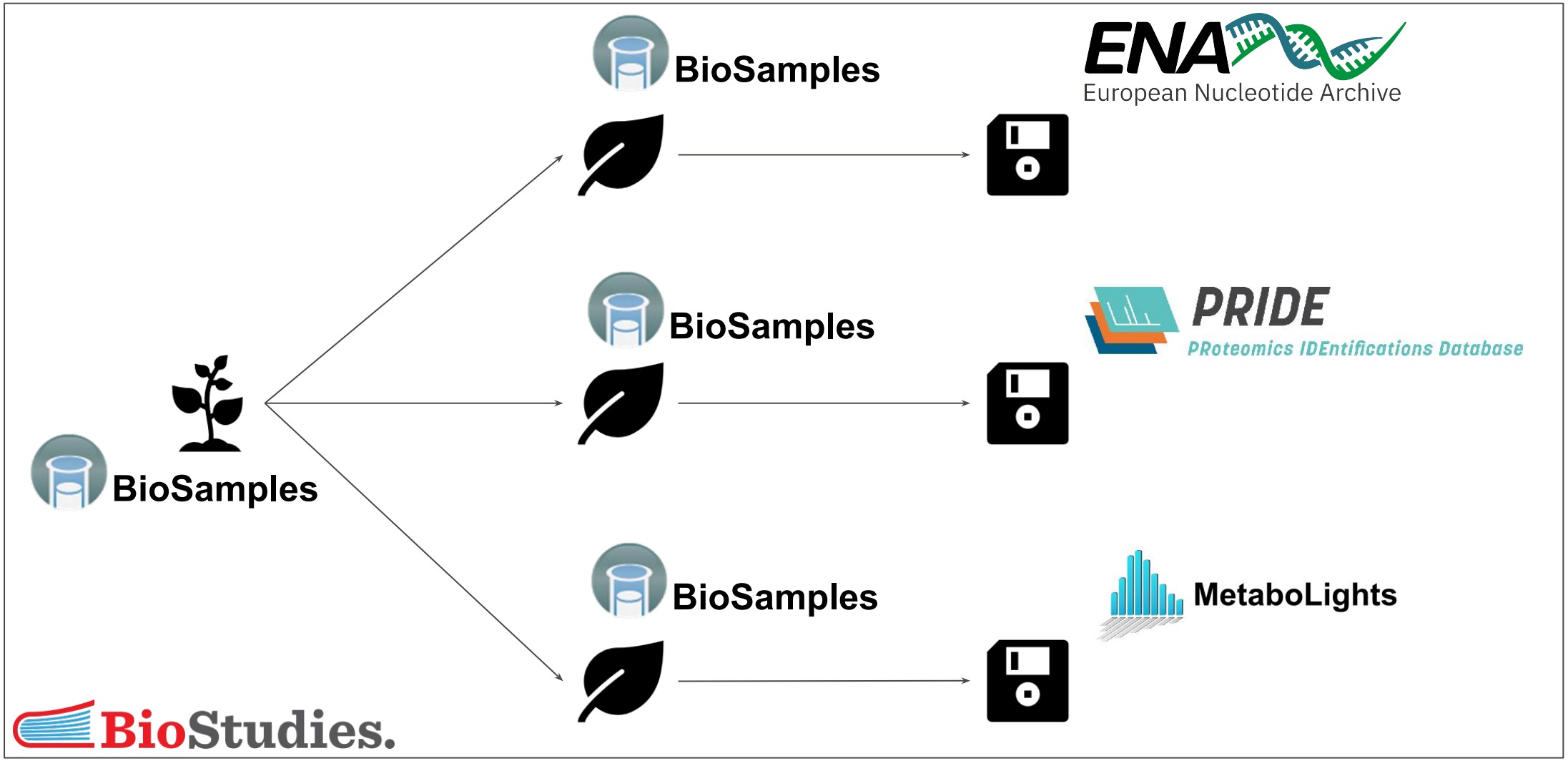
Multi-omics


BioSamples
BioSamples
BioSamples
ENA
European Nucleotide Archive
PRIDE
PRoteomics IDEntifications Database
MetaboLights

Multi-omics


BioSamples
BioSamples
BioSamples
BioSamples
ENA
European Nucleotide Archive
PRIDE
PRoteomics IDEntifications Database
MetaboLights

Multi-omics


BioSamples
ENA
European Nucleotide Archive
BioSamples
PRIDE
PRoteomics IDEntifications Database
BioSamples
BioSamples
MetaboLights
BioStudies.

What about human sensitive data?

"The European Genome-phenome Archive (EGA) is a service for permanent archiving and sharing of personally identifiable genetic, phenotypic, and clinical data generated for the purposes of biomedical research projects or in the context of research-focused healthcare systems."

EUROPEAN GENOME-PHENOME ARCHIVE

https://ega-archive.org/

- Central metadata access
- Secure storage of sensitive data
- Data access defined, including Data Access Committees (DAC)
- Different datasets may have different DACs

elixir BELGIUM

Access types

- **Open access:** Data is shared publicly. Open-access is a **rarely used** access mode for the sharing of human data. To use open-access researchers need to **ensure** that the **data** needs is **anonymised**, which is difficult in practice.

- **Registered access:** Data is shared with researchers, who have been vouched for by their institution and who **agree** to abide by **data usage policies**. Datasets that are shared via registered-access would typically have no restrictions besides the condition that data is to be used for research.

- **Controlled access:** Access is **reviewed** and **approved** by a data access **committee** (DAC). Typically researchers who were involved in the primary collection of data will form the DAC. There can be a **multitude of usage conditions**, including allowed research topics, geographical regions, and recipients e.g. non-profit organisations.

elixir BELGIUM

Modified from https://rdmkit.elixir-europe.org/human_data#sharing-and-reusing-of-human-data

Submitting data to the EGA

Quick Guide

This is a quick guide to submit data to the EGA. Please select data type to display the right detailed instructions. Please note - Submissions to EGA can take approximately one month, so please, allow plenty of time for the submission and archiving processes.

Array-based data | **Sequence data**

Get your submission account

Fill the **submission form** and provide details of the data type(s), used platform(s) and estimated size of your submission.
If you are affiliated to an existing consortium, such as the International Cancer Genome Project (ICGC), please add this information in the comments.

Register Study/DAC

Use **Submitter Portal** to register your study, samples, Data Access Committee (DAC) and Policy.

Upload data

Encrypt your data files using the **EgaCryptor** or **encrypt your files localy** and upload it using default FTP clients or Aspera.

Register experiments and runs

Associate each data file to a registered sample and study by **Linking files to samples** . Details of the experimental procedure you followed must be provided.

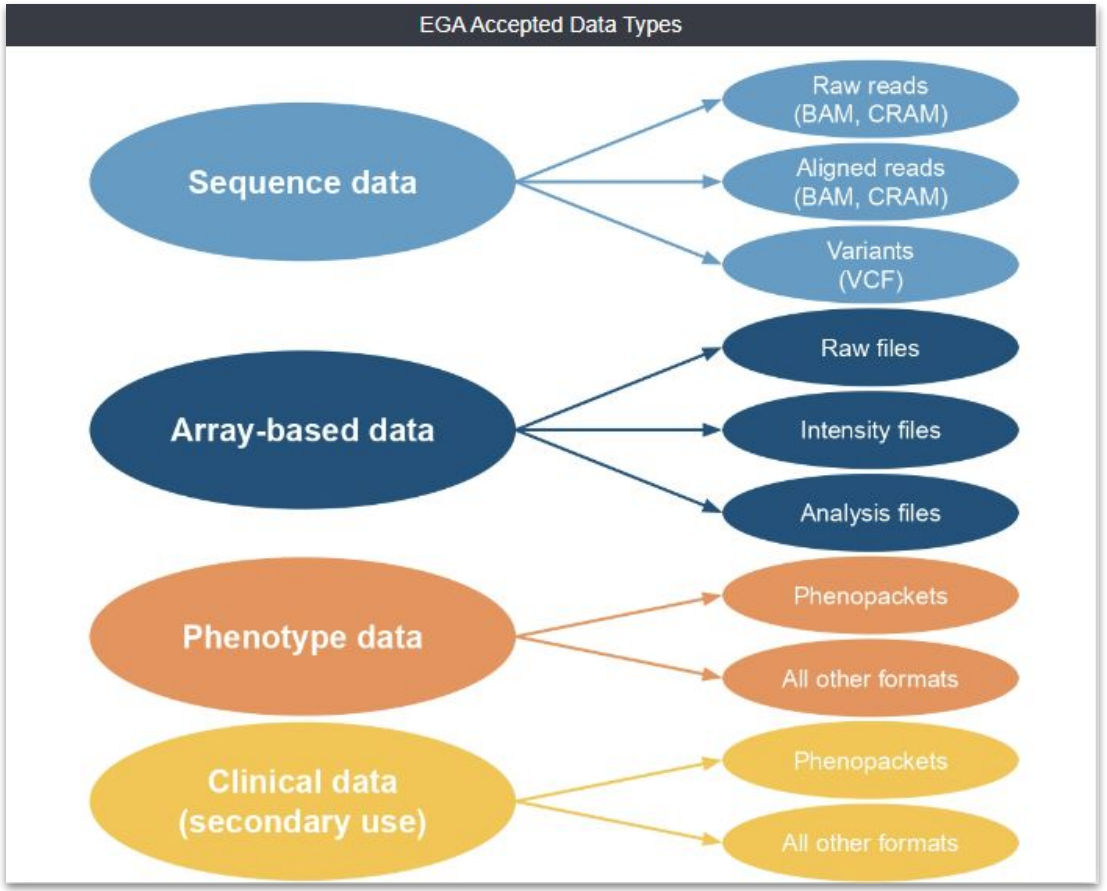
Finish your submission

Group your runs/analysis into datasets and link them to your new or existing DAC and policy using **Submitter Portal** or an **XML based programmatic** submission. Data request are done at a dataset level, thus files within a datasets must share release conditions.

Release and admin

Instruct our Helpdesk to release your study. All registered studies are automatically placed on hold until the named submission or DAC contact instructs our Helpdesk for the study to be released.

https://ega-archive.org/submission/quickguide



https://ega-archive.org/submission

What if the data cannot leave my country?

Federated EGA

https://ega-archive.org/federated

"Federated EGA strives to support the discovery of and secure access to human data globally, while respecting national data protection regulations, with the goal of accelerating disease research and understanding and improving human health."

https://doi.org/10.7490/f1000research.1118988.1

- Launched in September 2022
- Data does not leave the Local EGA
- Dataset discovery through the Central EGA

elixir BELGIUM

Federated EGA

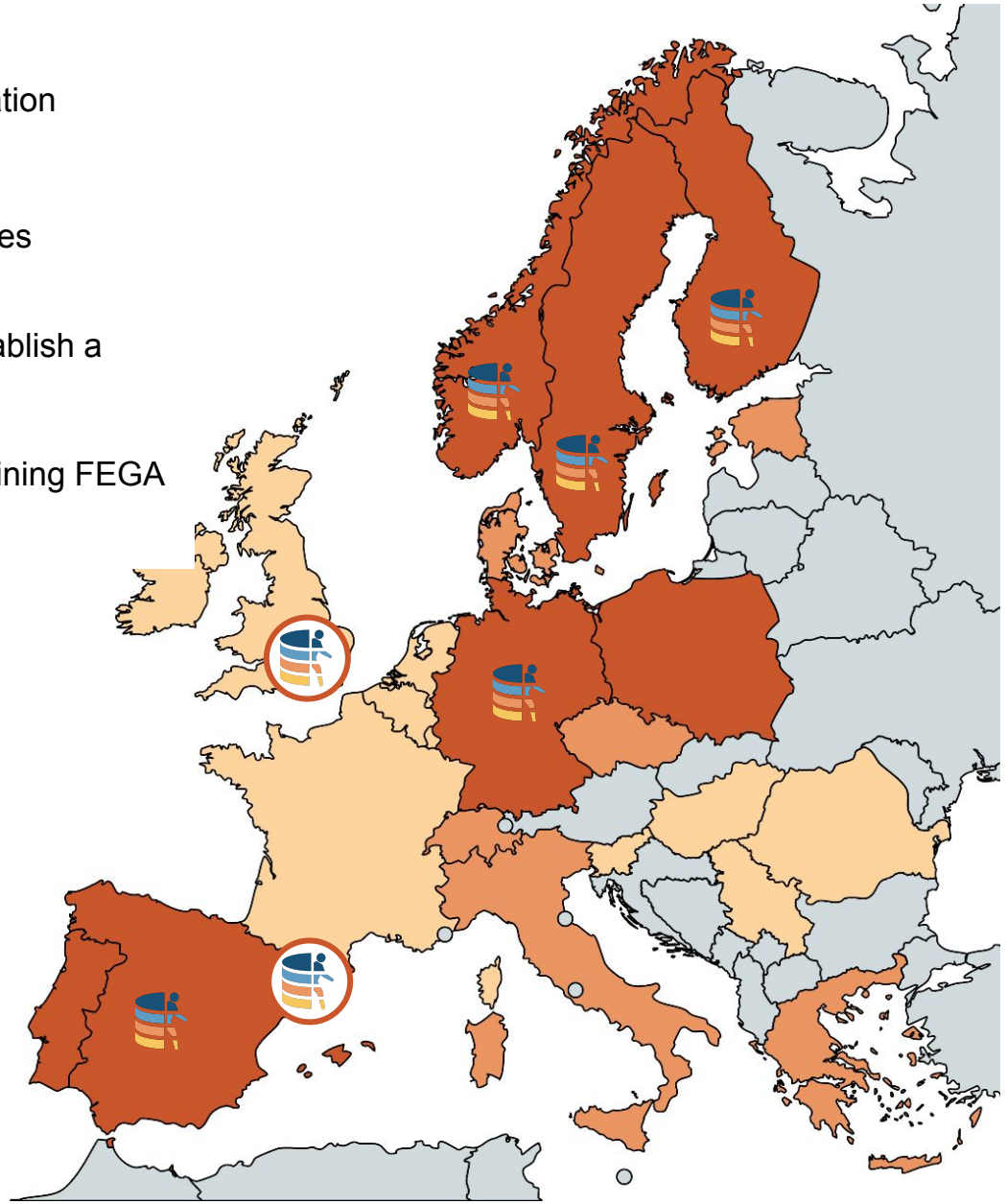
Central EGA

Signed FEGA Collaboration Agreement

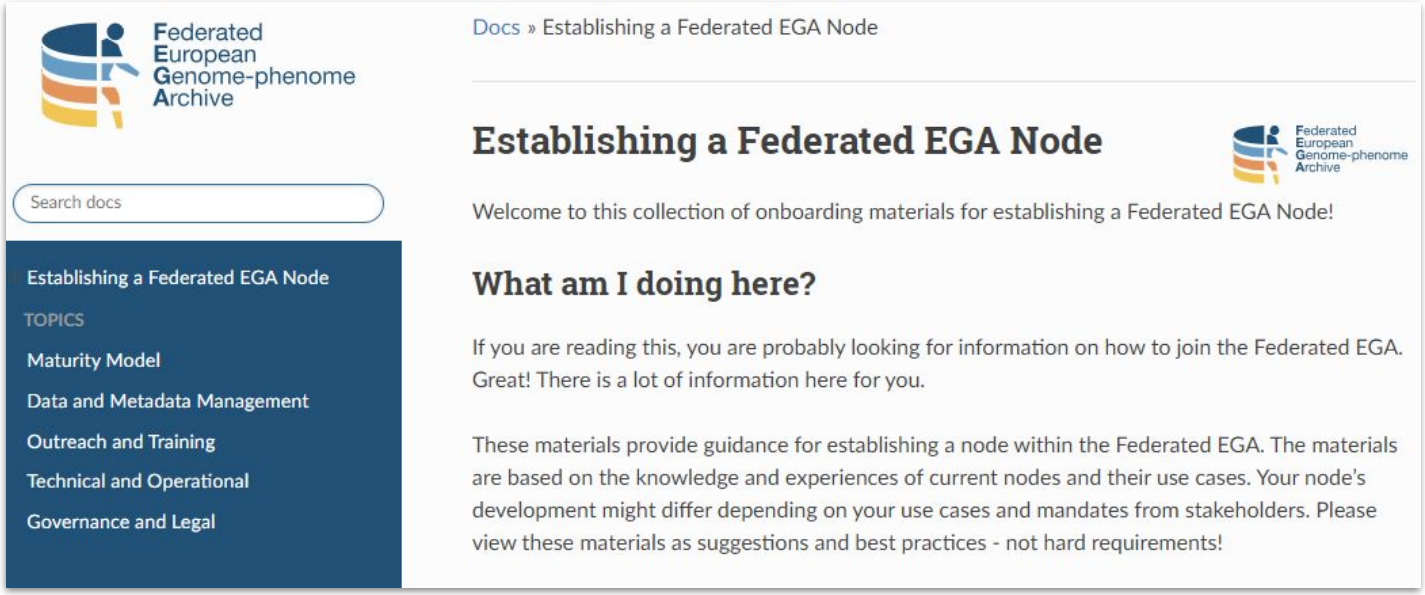
Joined FEGA Committees

Engaged in work to establish a FEGA node

Expressed interest in joining FEGA Network



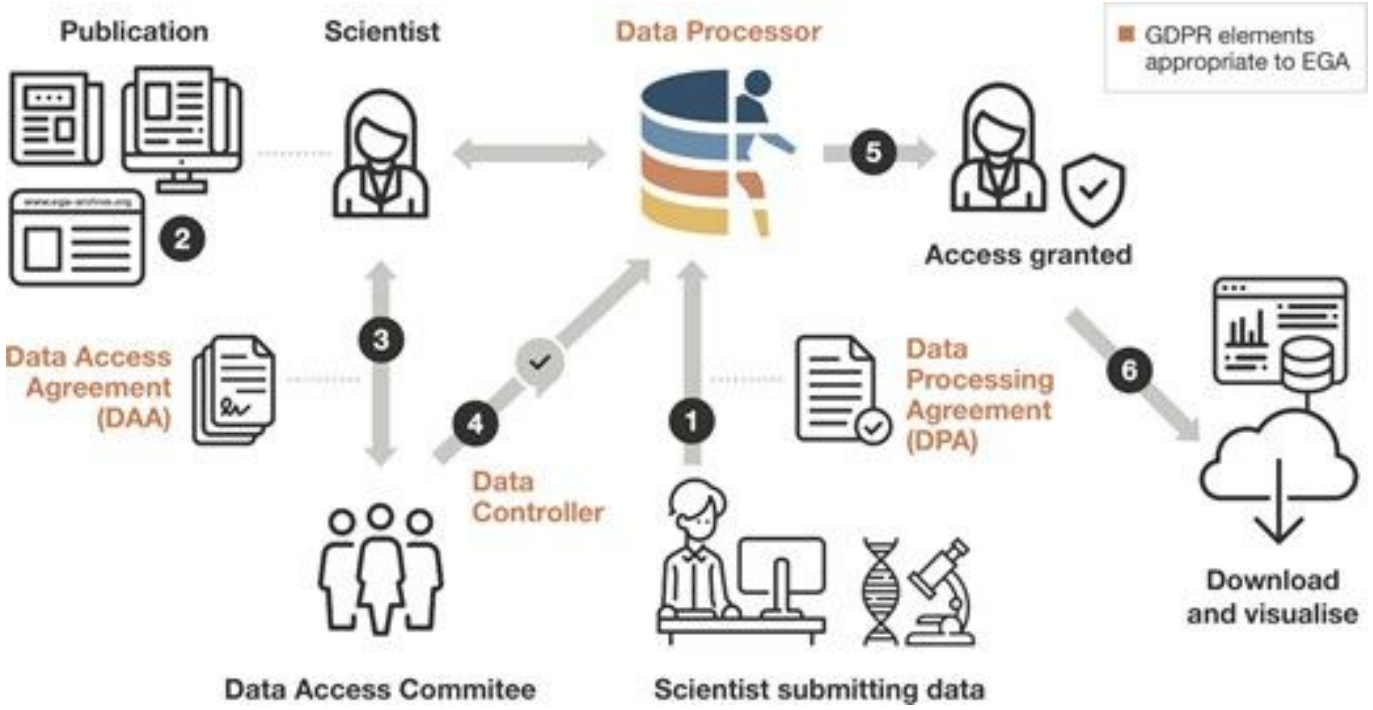
Modified from https://doi.org/10.7490/f1000research.1119478.1



https://ega-archive.github.io/FEGA-onboarding/

Re-using data from the EGA

1. Submission after a Data Processing Agreement
2. A scientist discovers the dataset (e.g. through a publication or by navigating the EGA)
3. The DAC is contacted and a DAA is defined
4. EGA is informed of access permission
5. Scientist can access the data
6. Data is downloaded and re-used



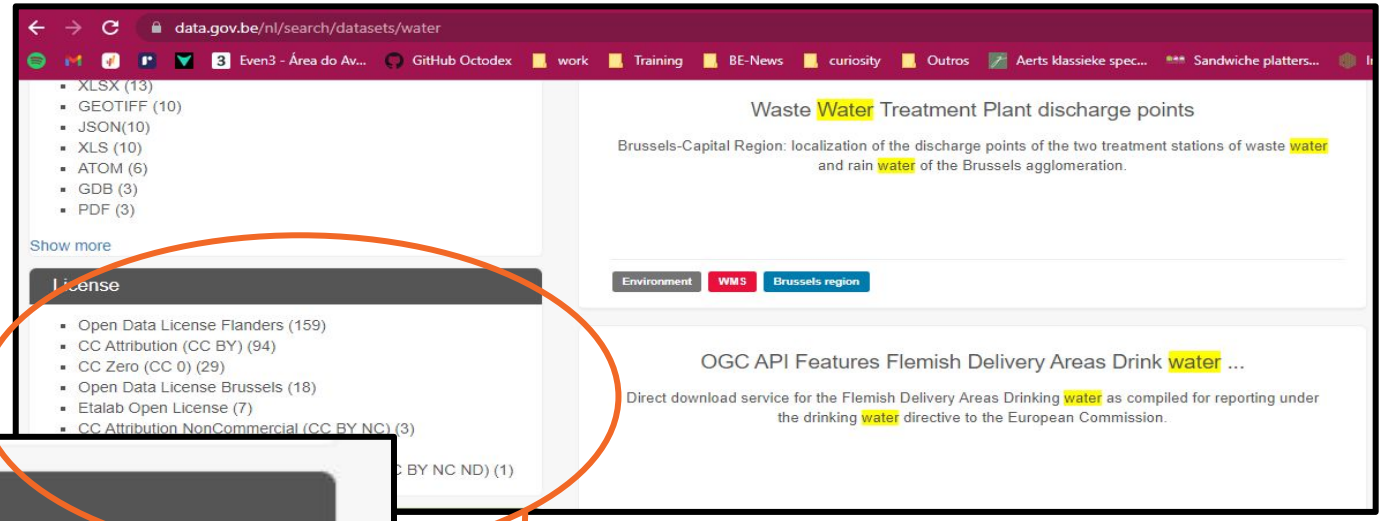
https://doi.org/10.1093/nar/gkab1059

Re-using data from repositories

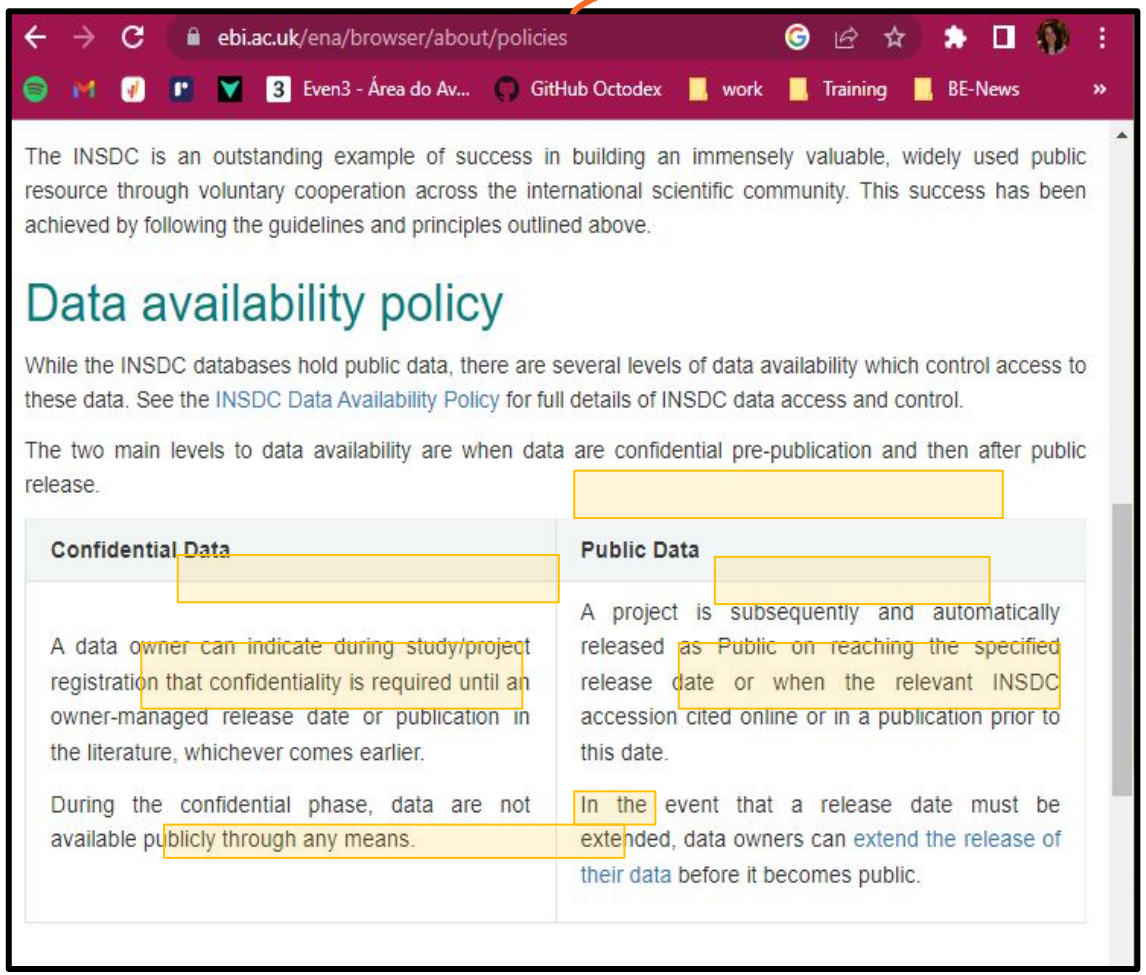
1. Look into Licenses and Policies
    a. For Reuse
    b. Before depositing

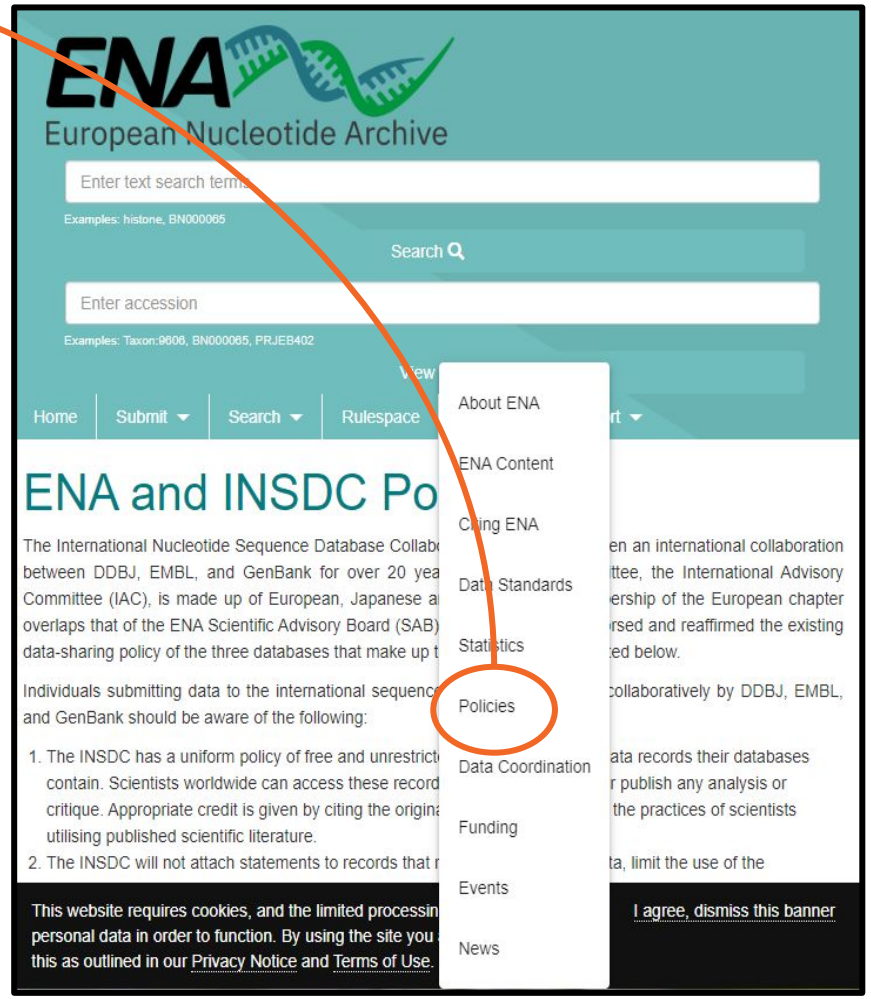
Re-using data from repositories

1. Look into Licenses and Policies
   a. For Reuse
   b. Before depositing

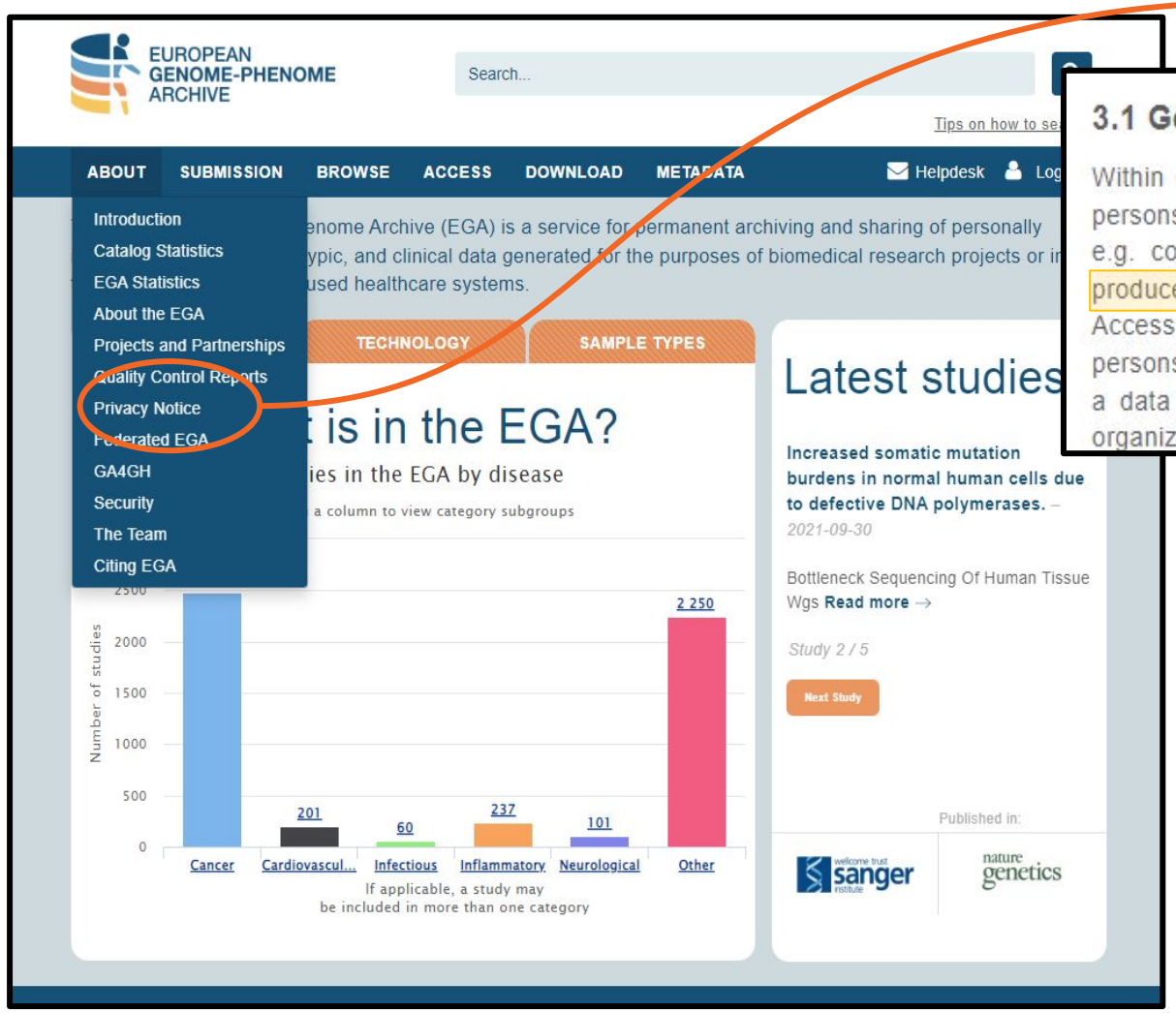


Re-using data from repositories





Re-using data from repositories



3.1 Genetic and phenotypic data

Within GDPR, there are two main actors: data controllers and data processors. Data controllers are persons or entities which determine the purposes and means that the personal data may be processed, e.g. companies, researchers, or universities. For EGA, the data controller is ultimately the data producer and the submitter(s) who submit the data to EGA. The data controller also creates a Data Access Committee (DAC) who will decide on data access permissions at EGA. Data processors are the persons or entities which process the data on behalf of a data controller. With regard to GDPR, EGA is a data processor as it processes data as instructed by the data controller. GDPR applies to any organization which accesses personal data from an individual within the EU. Under GDPR, personal

Re-using data from repositories

1. Look into Licenses and Policies
    a. For Reuse
    b. Before depositing

https://creativecommons.org/

No conditions stated! Most open.

Credit for the original creators

New creations must be shared under the same therms

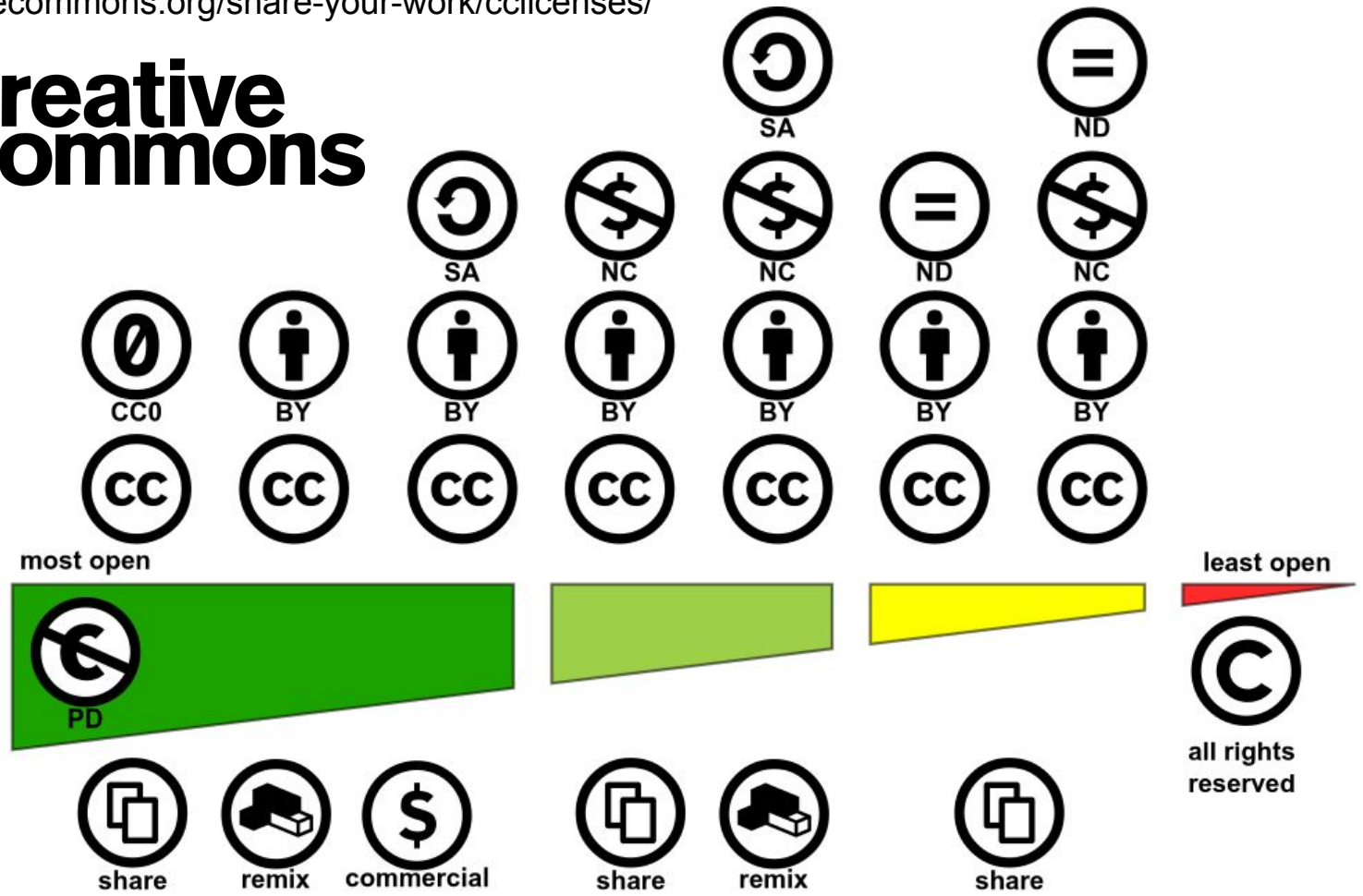
Only non-commercial use

Non derivatives or adaptations are allowed, only re-use in original form

elixir BELGIUM

Re-using data from repositories

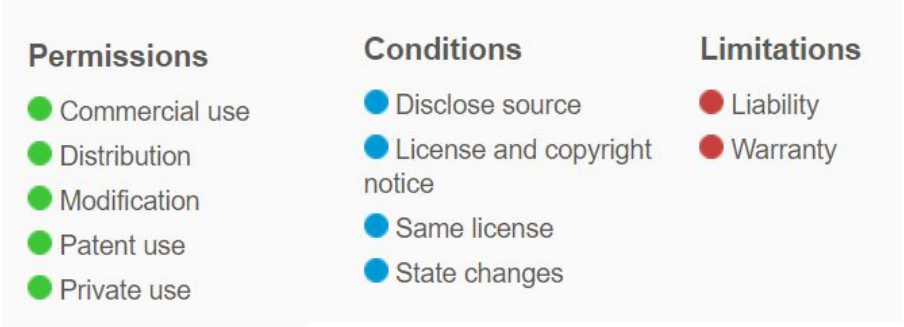
https://creativecommons.org/share-your-work/cclicenses/



https://creativecommons.org/choose

Re-using data from repositories

- GNU GPLv3:



- GNU AGPLv3: Complete source code modifications need to be available
- GNU LGPLv3: Large modifications can be distributed under different license.
- MIT license: No attribution license
- Apache 2.0 ~ GNU PLv3

https://opensource.org/licenses
https://choosealicense.com/

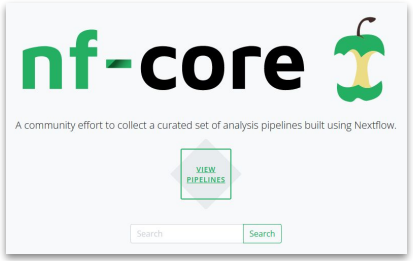
A short side note on computational workflows

Where can I find workflows?

Platform/System specific repositories

Galaxy PROJECT

IWC - Intergalactic Workflow Commission



Snakemake workflow catalog

The Git based repositories

GitHub

Generalist data repositories

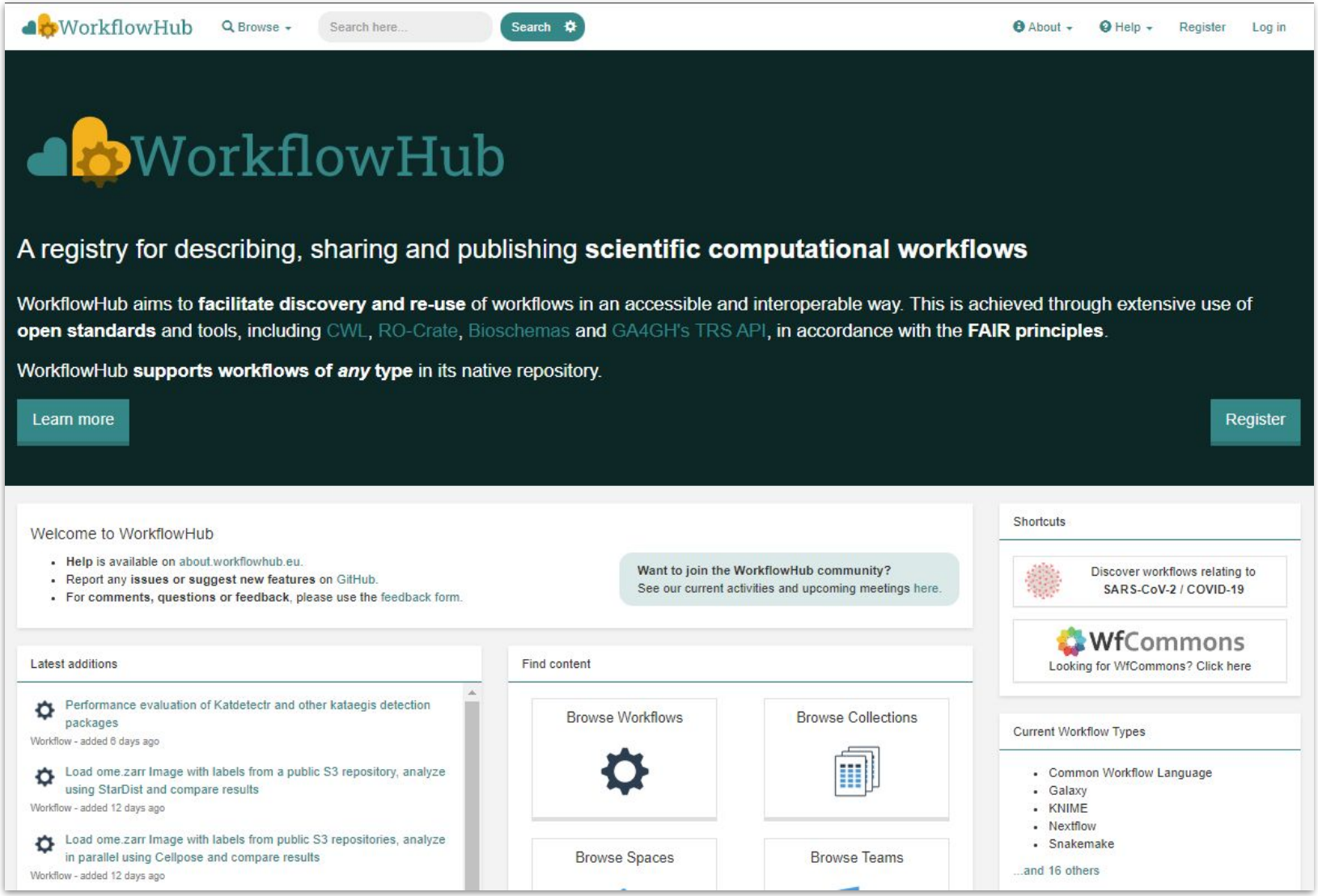
zenodo

OpenAIRE

elixir BELGIUM

Using a registry to overcome fragmentation

Start: April 2020
Launch: Sept 2022

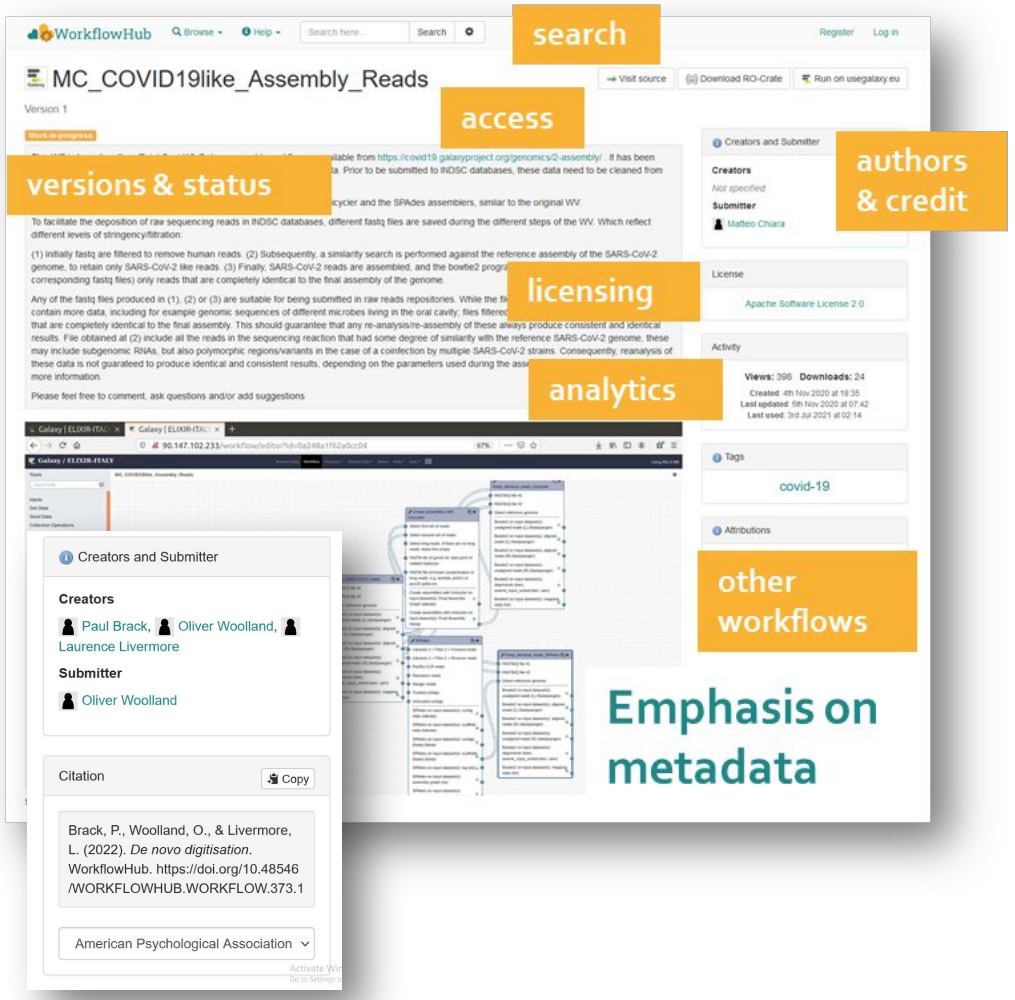


https://workflowhub.eu/

WorkflowHub

Key features

- System agnostic
- Native repository support
- Git integration
- Versioning
- Links to external files (docs, test and reference data)
- Metadata extraction and standards
- Integration with Bio.tools
- Author credit
- Citable
- DOI minting

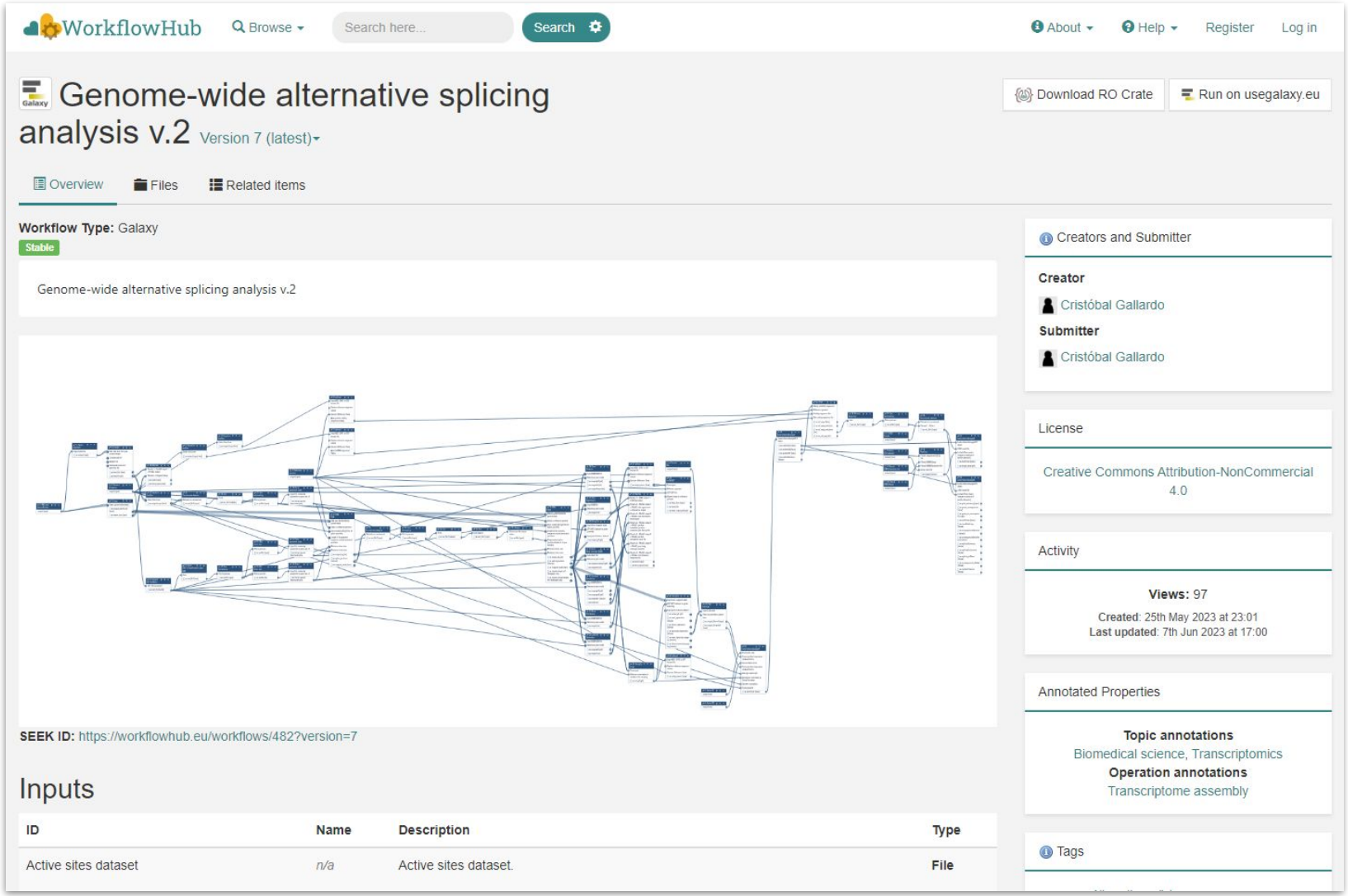


A sandbox is available

https://dev.workflowhub.eu/

Modified from slide of Carole Goble at ELIXIR All Hands 2023

Integration with different platforms



DOI: 10.48546/workflowhub.workflow.482.3

Galaxy EUROPE

https://usegalaxy.eu/

Connecting workflow discovery to analysis

elixir BELGIUM

Get in touch with us.

info@elixir-belgium.org

https://www.elixir-belgium.org