DMP and Open Data training
Session 3: Metadata

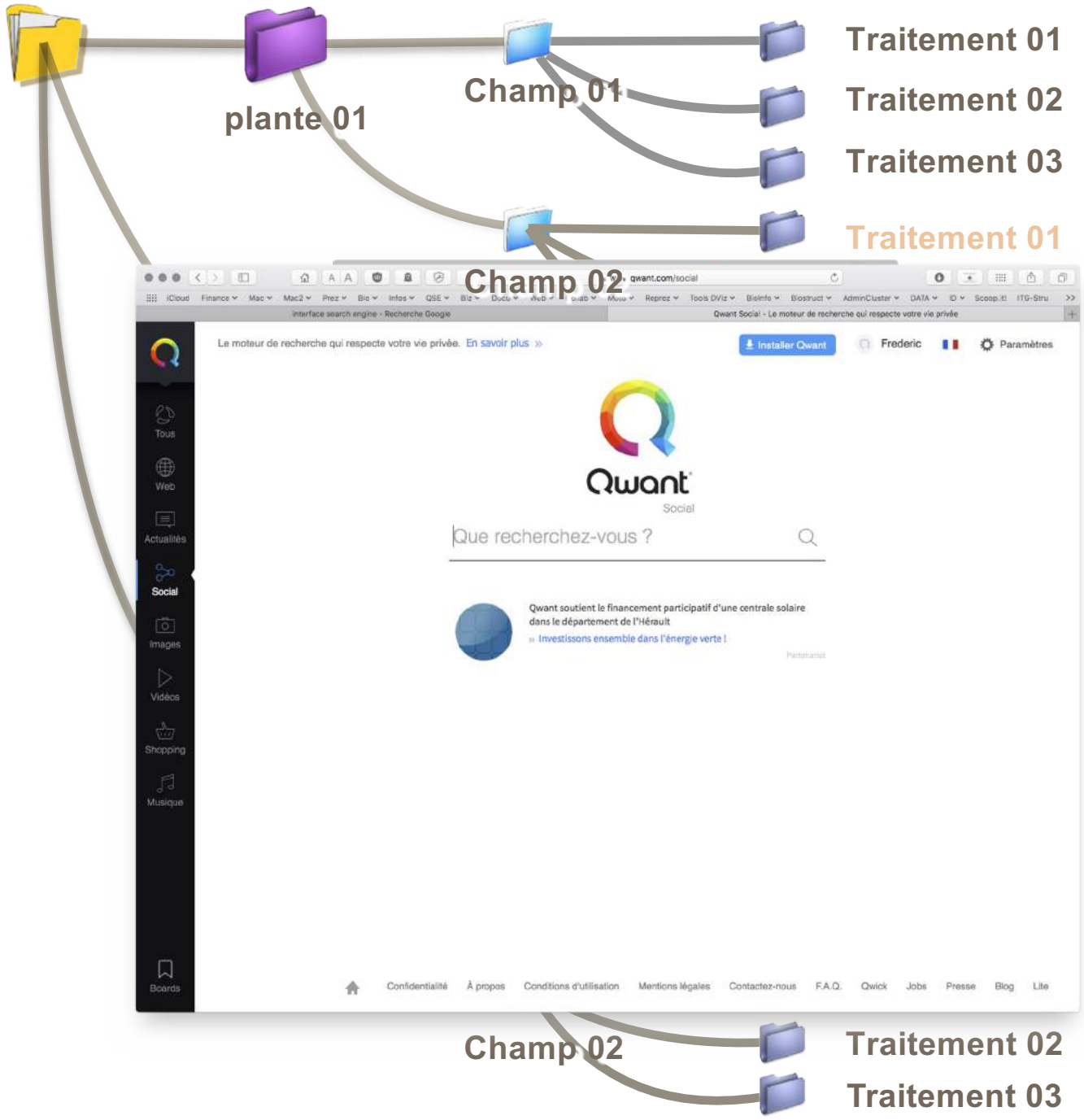
Introduction aux métadonnées

Valérie Cognat
&
Laurent Bouri

Fred de Lamotte : https://orcid.org/0000-0003-4234-1172

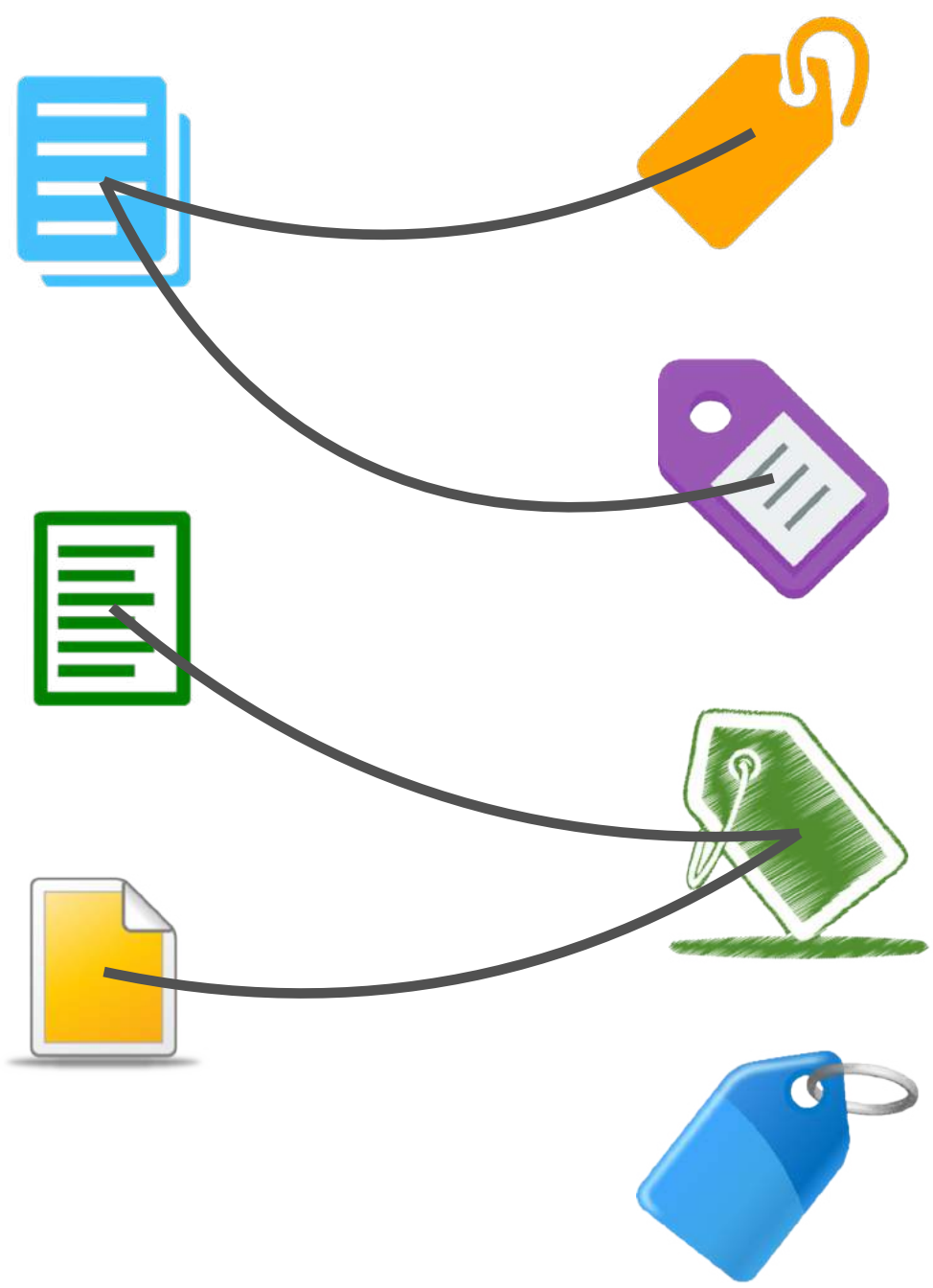
Pourquoi des métadonnées ?

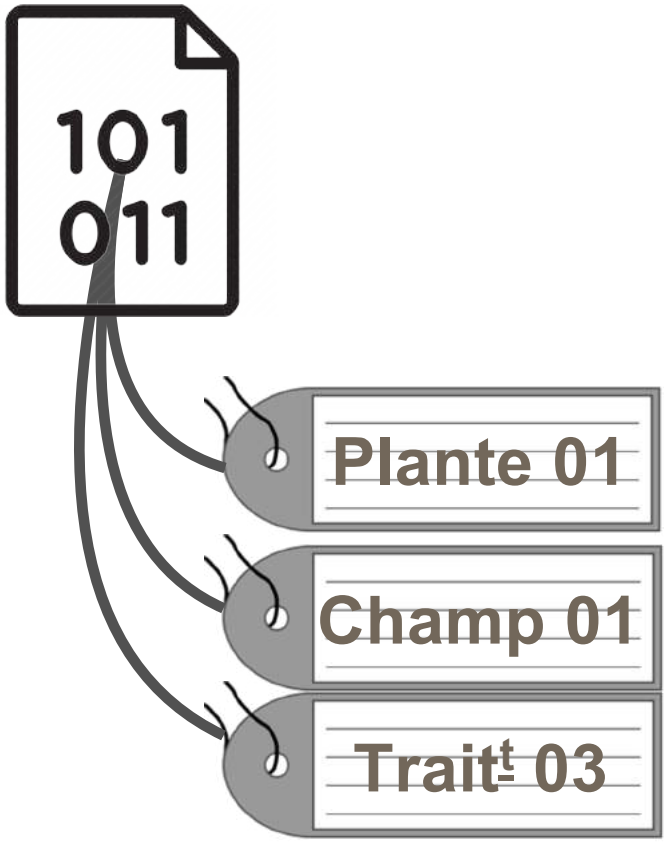
Définir une méthode commune et efficace pour retrouver nos données.


plante 01
Champ 01
Traitement 01
Traitement 02
Traitement 03
Champ 02
Traitement 01
Champ 02
Traitement 02
Traitement 03
qwant.com/social
Le moteur de recherche qui respecte votre vie privée. En savoir plus
Installer Qwant
Frederic
Paramètres
Tous
Web
Actualités
Social
Images
Vidéos
Shopping
Musique
Boards
Qwant
Social
Que recherchez-vous ?
Qwant soutient le financement participatif d'une centrale solaire dans le département de l'Hérault
Investissons ensemble dans l'énergie verte !
Confidentialité
À propos
Conditions d'utilisation
Mentions légales
Contactez-nous
F.A.Q.
Qwick
Jobs
Presse
Blog
Lite

Les métadonnées : définition

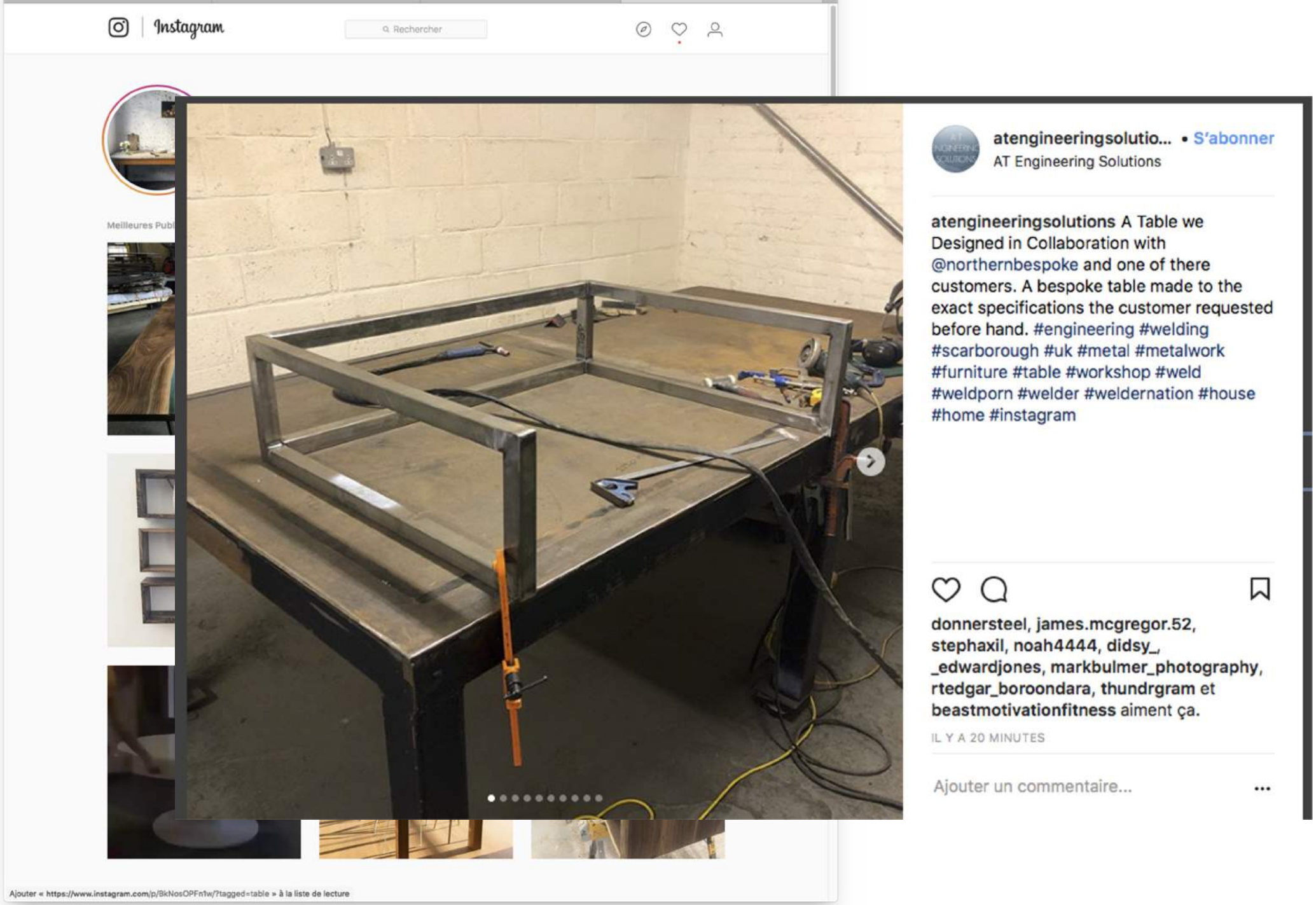
Les métadonnées



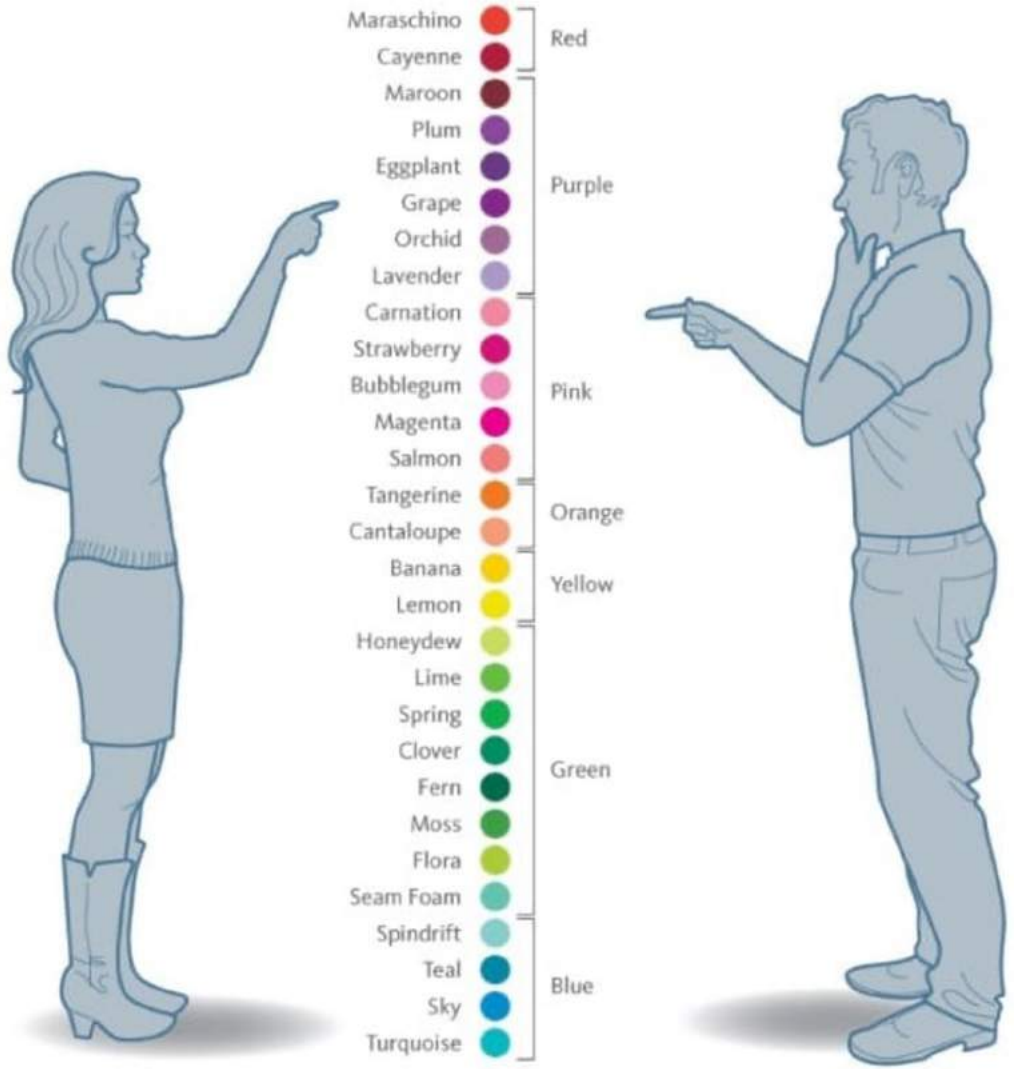


Tags
MetaData

Vocabulaire contrôlé
Défini par la communauté
Evolutif


Instagram
Rechercher
Meilleures Publ
atengineeringsolutio... • S'abonner
AT Engineering Solutions
atengineeringsolutions A Table we Designed in Collaboration with @northernbespoke and one of there customers. A bespoke table made to the exact specifications the customer requested before hand. #engineering #welding #scarborough #uk #metal #metalwork #furniture #table #workshop #weld #weldporn #welder #weldernation #house #home #instagram
donnersteel, james.mcgregor.52, stephaxil, noah4444, didsy_, _edwardjones, markbulmer_photography, rtedgar_boroondara, thundrgram et beastmotivationfitness aiment ça.
IL Y A 20 MINUTES
Ajouter un commentaire...
Ajouter « https://www.instagram.com/p/BkNosOPFn1w/?tagged=table » à la liste de lecture

Un vocabulaire contrôlé



Genre	Espece	Sous espece	Groupe	Nom
Oryza	Sativa		japonica	PENTHE BLANC
Oryza	Sativa		japonica	PENTHE NOIR
Oryza	Sativa		indica	ZOGO
Oryza	Glaberrima			GBAI-GBAI
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Sorghum	bicolor	bicolor	Dura	IS19453
Musa	acuminata	banksii	wild	Banksii H09

Germplasme	Origine	Collection
AG0003	Guinea	prospection 1979
AG0004	Guinea	prospection 1979

Type de sequençage	Taille insert	Longueur de read	Type de machine	Lieu du séquençage
illumina		1*150	HiSeq3000	Genotoul
illumina		1*150	HiSeq3000	Genotoul

Qu’est ce que les métadonnées

C’est du **data reporting** :

- WHO : Qui a créé les données
- WHAT : Quel est le contenu des données ?
- WHEN : Quand ont-elles été créées ?
- WHERE : Où ont-elles été créées ?
- HOW : Comment ont-elles été créées ?
- WHY : Pourquoi ont-elles été créées ?

Un standard de métadonnées

PROGRAMME NATIONAL DE NUMÉRISATION
ET DE VALORISATION DES CONTENUS CULTURELS

RECOMMANDATIONS TECHNIQUES POUR LES MÉTADONNÉES ET STANDARDS

VERSION N°1 – 2017

Le regard métier

Un jeu de métadonnées ne sera pas formalisé de la même manière selon les standards employés. Un même objet peut ne pas être décrit de la manière selon la perspective « métier » portée sur lui. Le regard « métier » structure la donnée. A titre d'exemple le traitement documentaire appliqué à une collection de cartes postales ne sera pas le même selon que celui-ci est opéré par un musée ou un service d'archives. Les archivistes s'attacheront à retrouver les toponymes là où les musées relèveront plutôt des détails ayant trait à l'histoire de l'art. (mais aussi les divergences entre les 2 communautés « climat »)

Ouverture et interopérabilité

Le traitement documentaire doit s'inscrire dans des logiques d'ouverture et d'interopérabilité. La qualité des données et métadonnées conditionne les réutilisations possibles, il en est de même pour le degré d'ouverture des ressources et de leurs métadonnées.
Le choix des métadonnées qui seront produites dans le cadre d'un projet de numérisation peut répondre à des usages clairement identifies en amont du projet. Le fait de s'appuyer sur des standards favorise l'interopérabilité et peut permettre des usages autres que ceux attendus.

Approches participatives

Une approche participative peut venir compléter le traitement documentaire. Cette approche participative peut prendre diverses formes : collecte, enrichissement de métadonnées, annotations, transcriptions collaboratives... Il est préférable d'envisager cette approche collaborative en amont ou en parallèle du projet.
Le porteur de projet doit entre conscient des risques induits pour la qualité et la fiabilité des données et la nécessité de gérer et animer les communautés d'utilisateurs selon le type d'approche choisi.

Formation “Science Ouverte & PGD”
Module 3: Metadata

Standards de métadonnées en Sciences de la Vie

Valérie Cognat, IBMP, Strasbourg
https://orcid.org/0000-0001-9337-2767

&

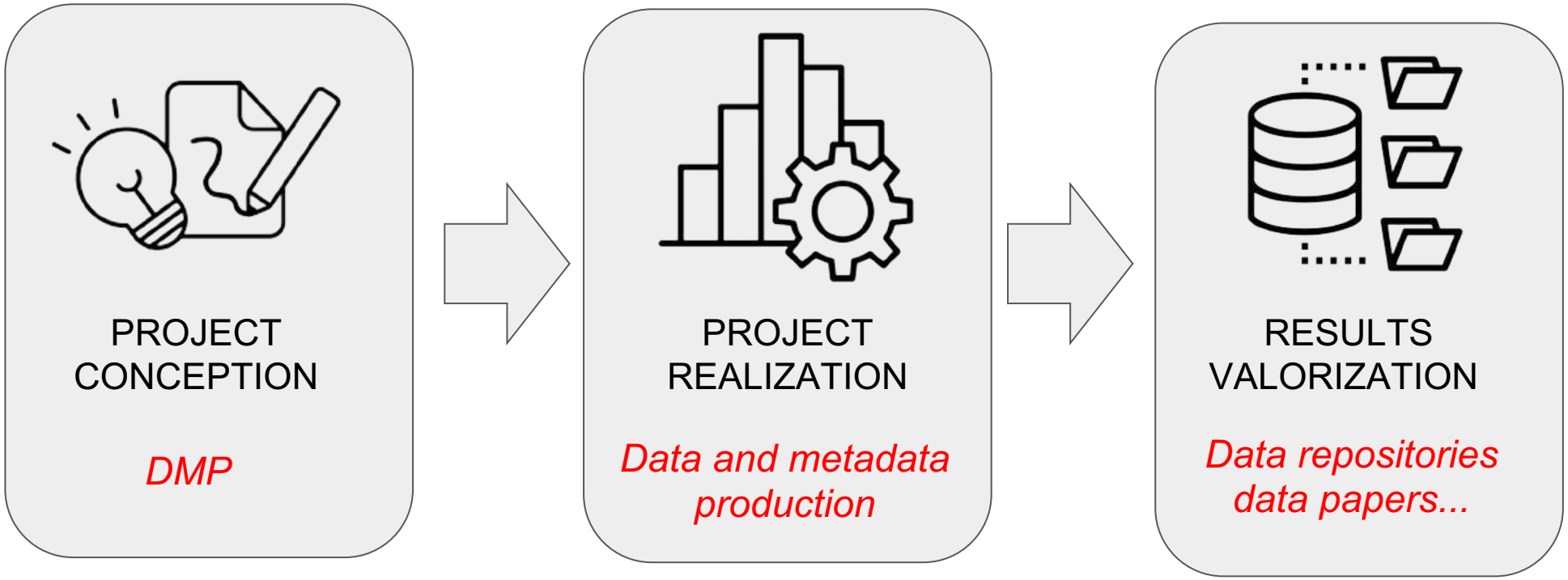
Laurent Bouri, IFB, IGBMC Strasbourg
https://orcid.org/ 0000-0002-2297-1559

Hélène Chiapello, IFB, INRAE Jouy-en-Josas
https://orcid.org/0000-0001-5102-0632

Thomas Denecker, IFB, CNRS Paris
https://orcid.org/0000-0003-1421-7641

INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Metadata during the project



Metadata concern all steps of a scientific project !

How do I produce metadata?

How do you describe the data?

With a set of metadata

	A	B
1	Titre	
2	Auteur	
3	Date	
4	Résumé	
5	Mots-clés	
6	Identifiant	
7	Format	
8	Contexte de création	

How do you ensure you don't forget certain metadata?

With a **metadata standard**

Disciplinary standard

General standard

Source: https://www.pasteur.fr/fr/file/20615/download

Question: Do you know any standard in life sciences ?

5 minutes to find an example of metadata standard and write a note in

Definition of a standard

In essence, a standard is an **agreed way of doing something**.

A standard provides the **requirements**, **specifications**, **guidelines** or **characteristics** that can be used for the **description**, **interoperability**, **citation**, **sharing**, **publication**, or **preservation** of all kinds of **digital objects** such as data, code, algorithms, workflows, software, or papers.

source: https://fairsharing.org/educational/

Example of standard in biology : Gene Ontology

The standards concern both data and metadata

Why do I have to use a **data standard**?

- To analyse, compare and exchange data
- To publish datasets in international resources

And a **metadata standard**?

- To describe data richly and accurately, with the same vocabulary as the rest of your scientific community
- To make your metadata interoperable and to allow other systems to exploit them

The Gene Ontology is a **metadata** standard

Generic and specific standards for metadata

Two kinds of standard descriptors

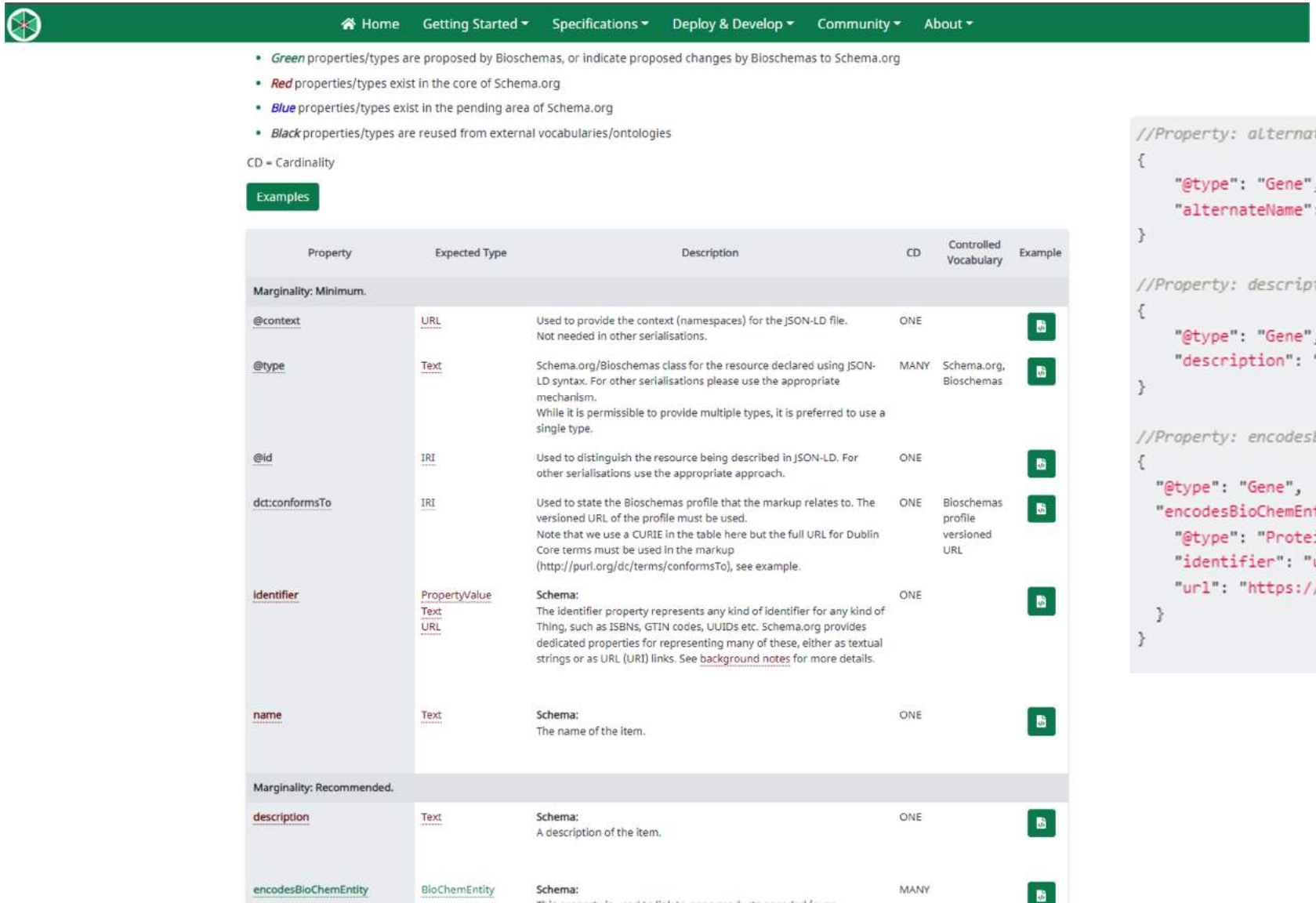
- Generic descriptors:
  - Dublin core for description of numerical resources
  - bioschema.org for description of life science resources (datasets, softwares, training material,...)
- Specific dataset descriptors:
  - MIAME (Minimum Information About a Microarray Experiment)

Metadata standards often depend on the repository you will use to publish data

> It is helpful to decide at the beginning of the project what are the recommended repositories for your data types

> You can view ELIXIR repositories here: https://elixir-europe.org/platforms/data/elixir-deposition-databases

Bioschemas profile specifications for “Gene”



Json-LD

```
//Property: alternateName
{
    "@type": "Gene",
    "alternateName": "AD1"
}

//Property: description
{
    "@type": "Gene",
    "description": "amyloid beta precursor"
}

//Property: encodesBioChemEntity
{
  "@type": "Gene",
  "encodesBioChemEntity": {
    "@type": "Protein",
    "identifier": "uniprotkb:P05067",
    "url": "https://www.uniprot.org/uniprot/P05067"
  }
}
```

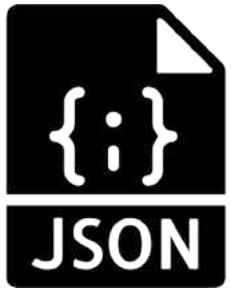
Three text formats frequently used for metadata

Comma Separated Values

```
Sample_ alias, date, source
A, 20200802, blood
B, 20200802, feces
C, 20200802, skin
```

eXtensible Markup Language

```
<SAMPLE_SET>
          <SAMPLE alias="A">

          <date>20200802</date>
          <source>blood</source>
     </SAMPLE>
          <SAMPLE alias="B">

          <date>20200802</date>
          <source>feces</source>
     </SAMPLE>
          <SAMPLE alias="C">

          <date>20200802</date>
          <source>skin</source>
     </SAMPLE>
</SAMPLE_SET>
```

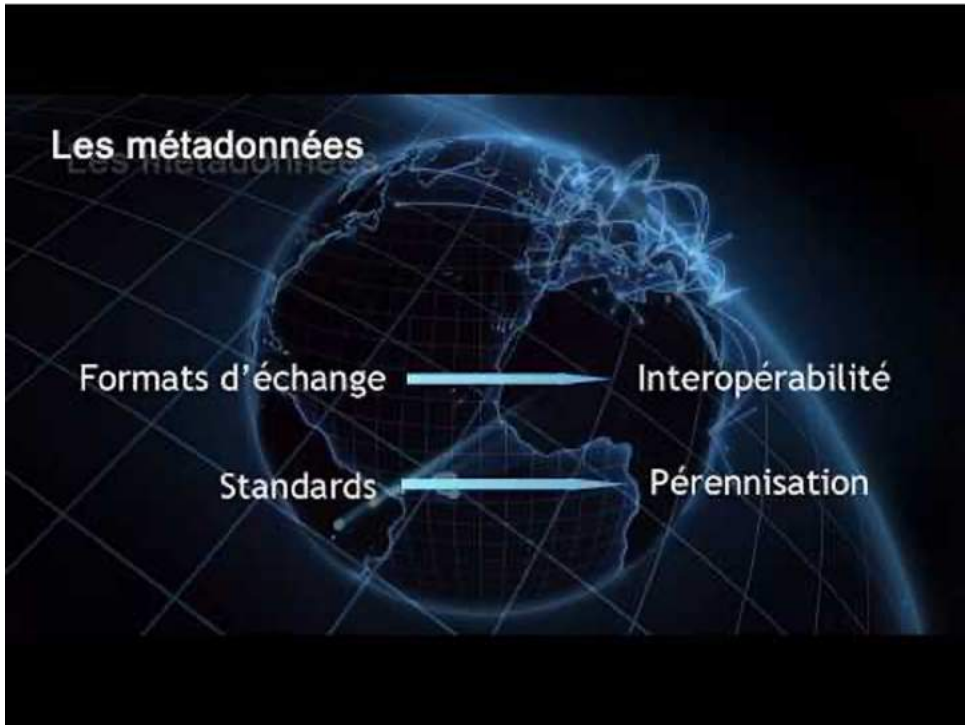


JavaScript Object Notation

```
{
"SAMPLE_SET": {
          "SAMPLE":[
     {
          "alias": "A",

          "date":"20200802",
          "source":"blood"
     },
          {
          "alias": "B",

          "date":"20200802",
          "source":"feces"
     },
  {
          "alias": "C",

          "date":"20200802",
          "source":"skin"
     }
     ]
  }
```

Vidéo: la minute métadonnées:

https://doranum.fr/metadonnees-standards-formats/

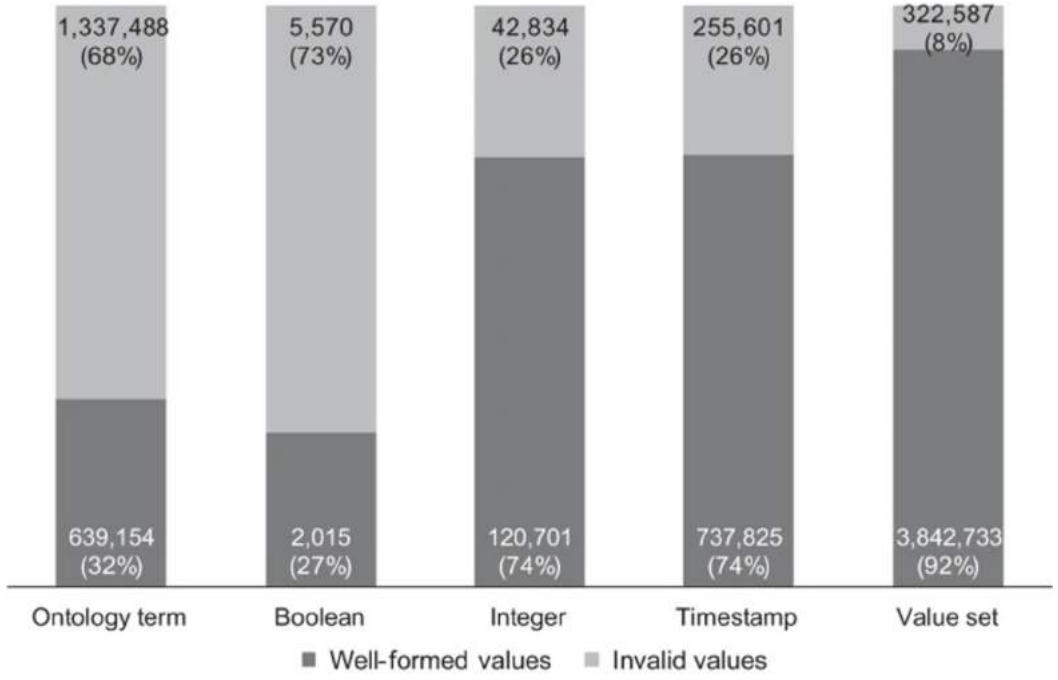


Metadata exhibit questionable quality in biology

Submission in public resources is often a complex task

Submission procedures are heterogeneous

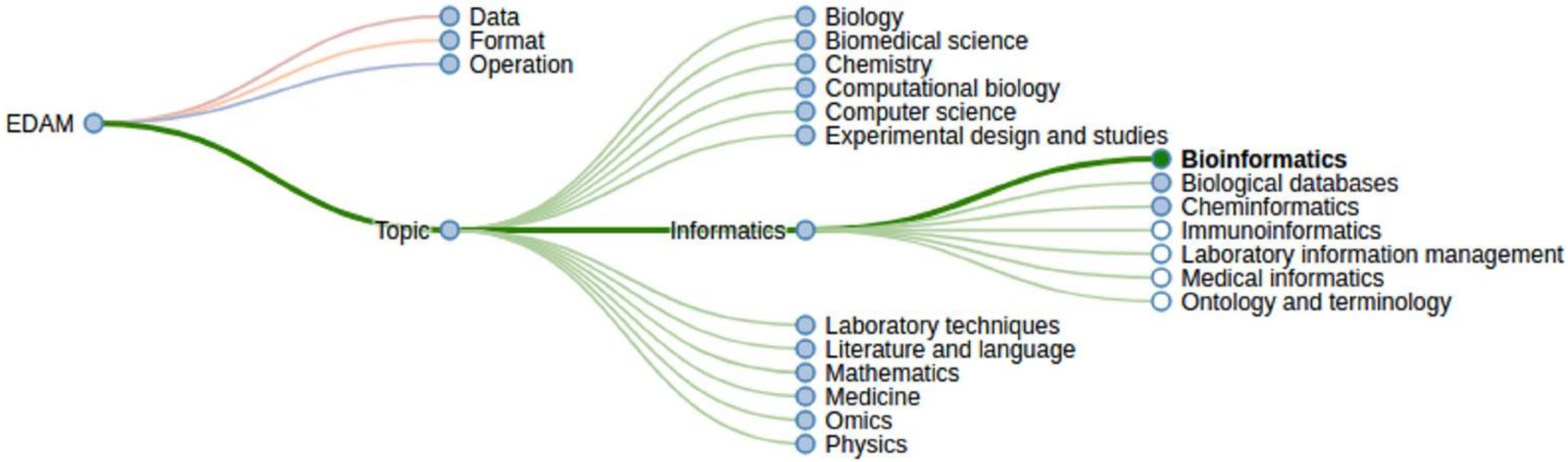
Metadata are often incomplete, inconsistent, redundant or not informative enough



Quality of dictionary attributes in NCBI BioSample according to their type, in Gonçalves et al., 2019

Ontology of bioscientific data analysis and data management

EDAM: a simple ontology of well established, familiar concepts that are prevalent within bioinformatics



https://edamontology.github.io/edam-browser/#topic_0091

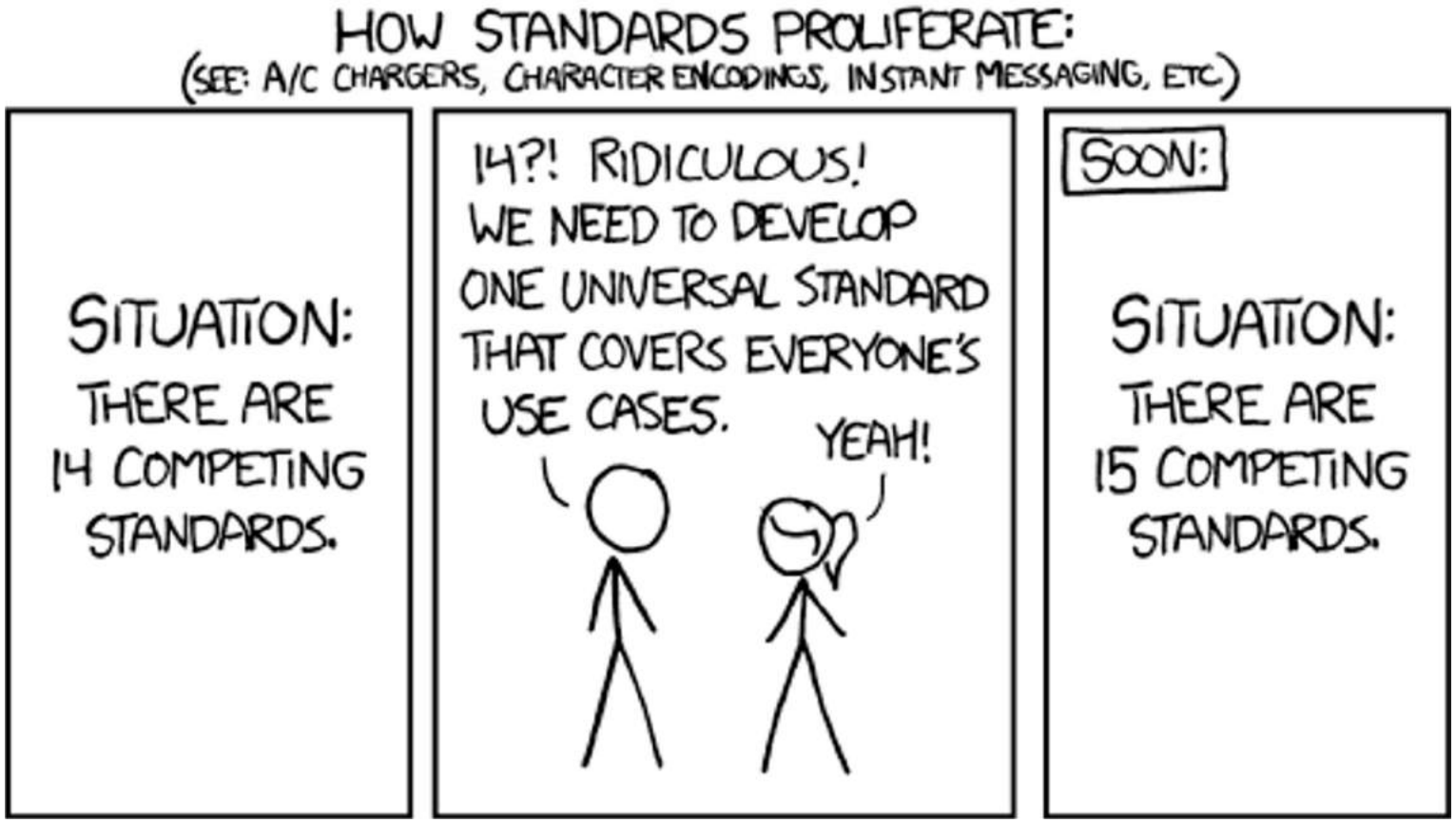
Standard adoption and perennity

- There are thousand of databases, softwares and resources in biology with an **unequal level of standard adoption**
- Is is not always easy for life scientists and bioinformaticians to identify and use the most appropriate standards

1641 databases in NAR Database 2021

Rigden *et al*, 2021

Standard adoption and perennity



Source: https://xkcd.com/927/

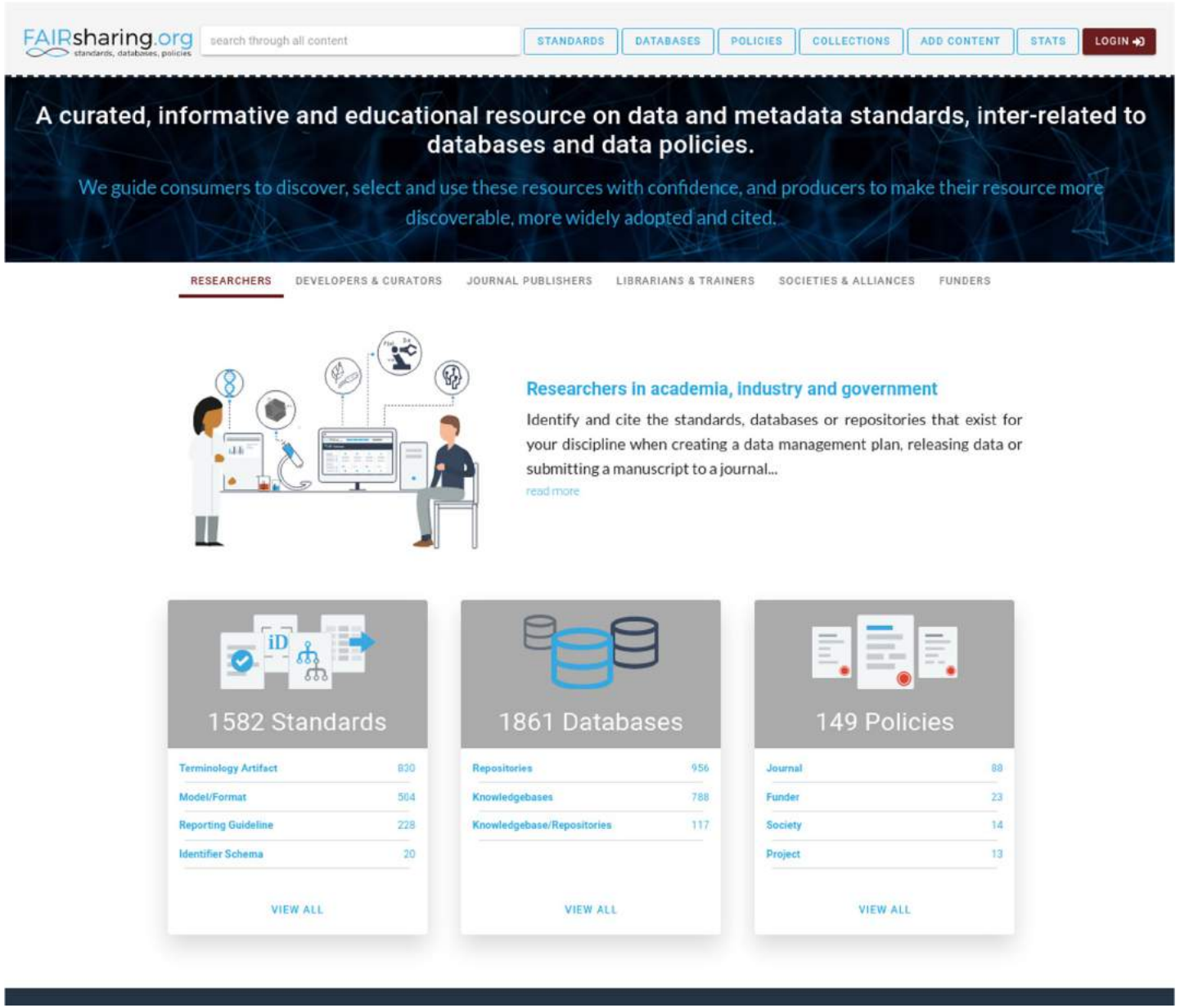
How do I find the standard I need?

The FAIRsharing portal

Sansone, *et al.* FAIRsharing as a community approach to standards, repositories and policies.

Nat Biotech. 2019

https://doi.org/10.1038/s41587-019-0080-8



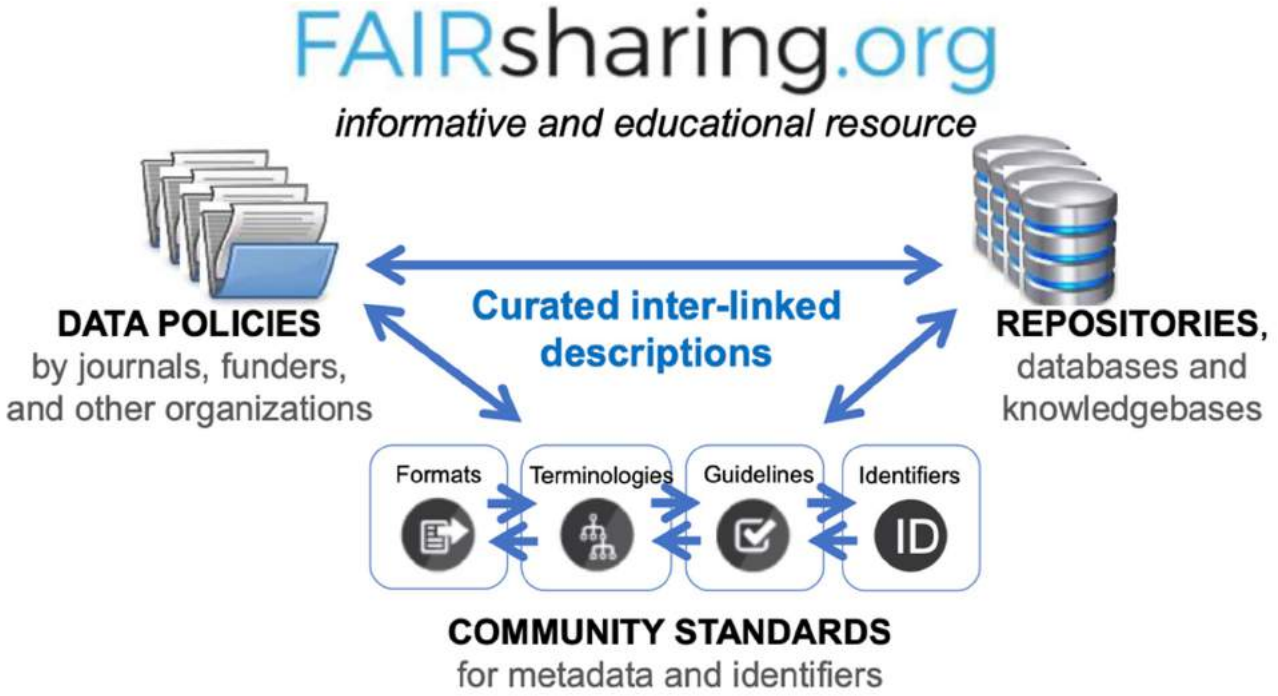
https://fairsharing.org

The FAIRsharing portal

Citable *DOI* for all records

Accessible via *API* or *web interface*

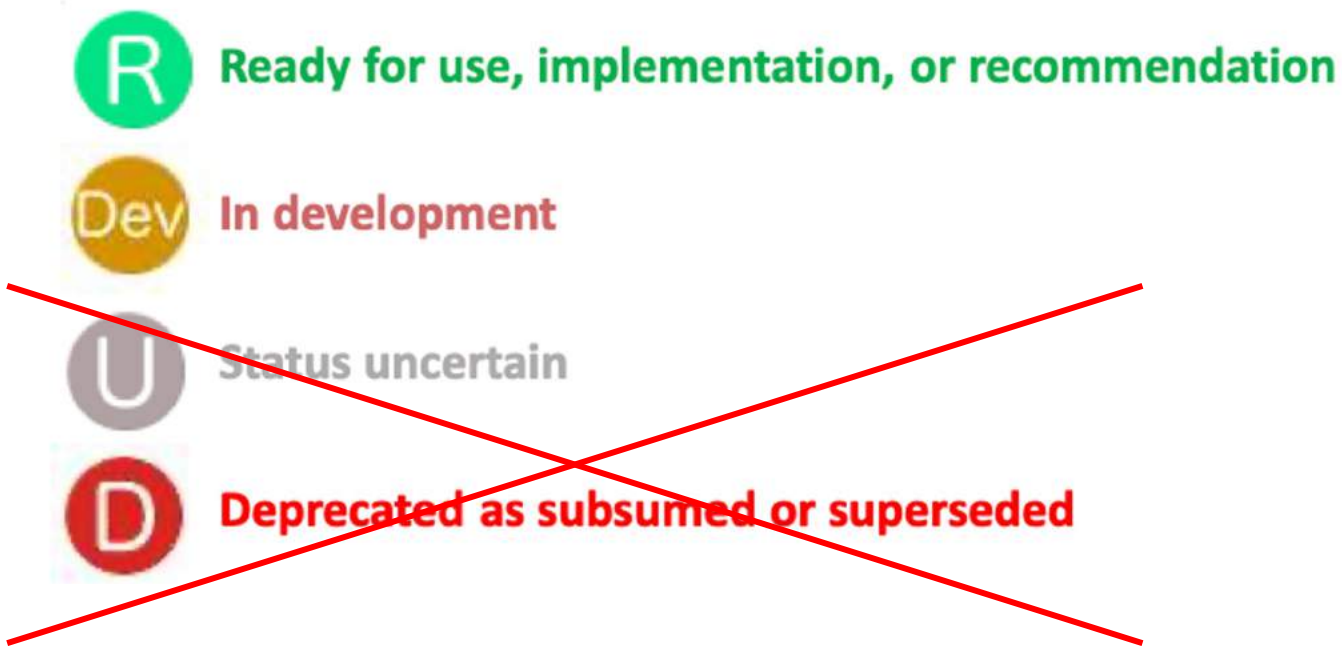
Curation



RECORD STATUS

https://fairsharing.org

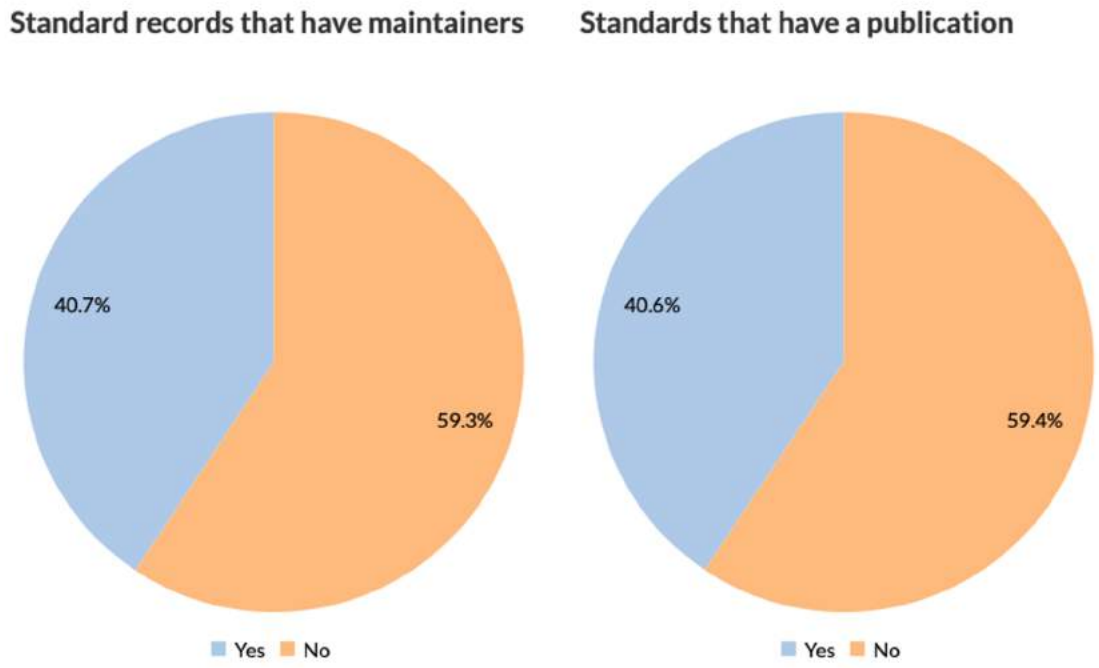
The FAIRsharing portal: record status



Please don’t use “Uncertain” or “Deprecated” standards

https://fairsharing.org

Standard maintenance is a key point



59.3 % of standards have no maintainer

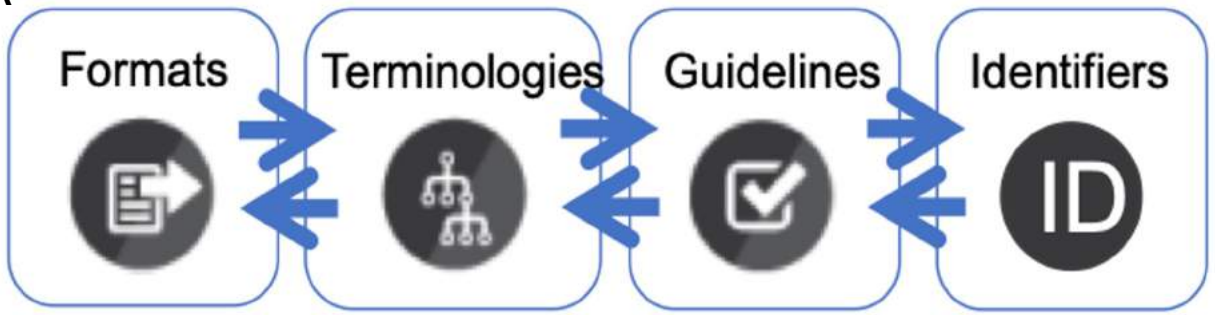
59.4% of standard has no publication

https://fairsharing.org/summary-statistics/?collection=standards

Types of data standards

Conceptual model, schema, exchange formats,etc…
e.g. SBML, FASTA

Minimum information reporting requirements, checklists…
e.g. MIAME guidelines

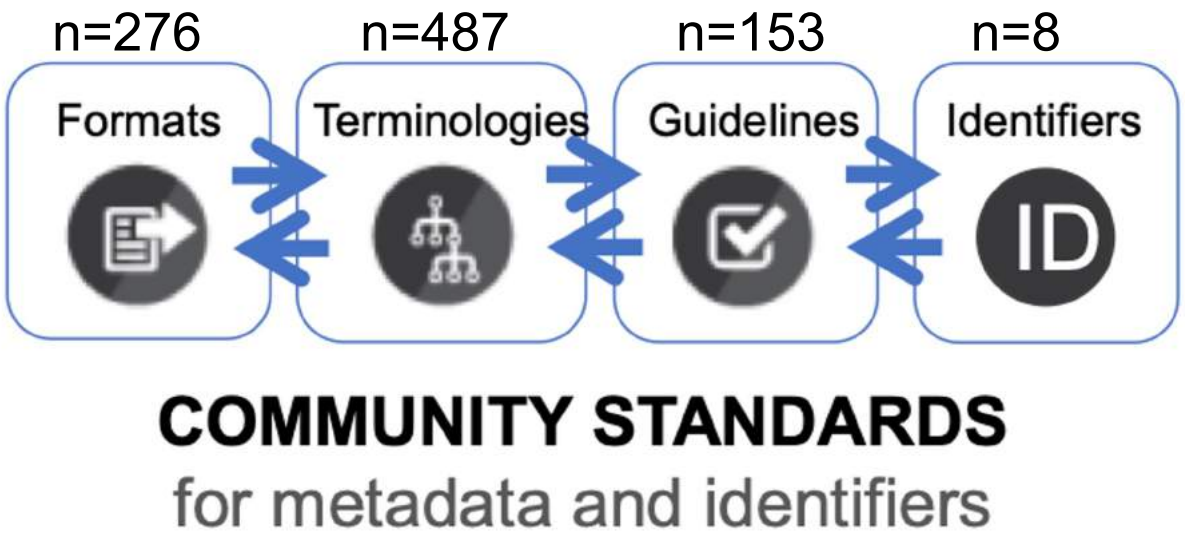


Controlled vocabularies, taxonomies, ontologies…
e.g. Gene Ontology

Formal systems for resources and digital objects that allow their identification
e.g. DOI

The landscape of standards in life sciences

FASTQ
FASTA
SBML
GFF
MIAME
Newick
EC number
BAM
VCF
MINSEQE



Source: https://fairsharing.org/standards/?q=life+sciences

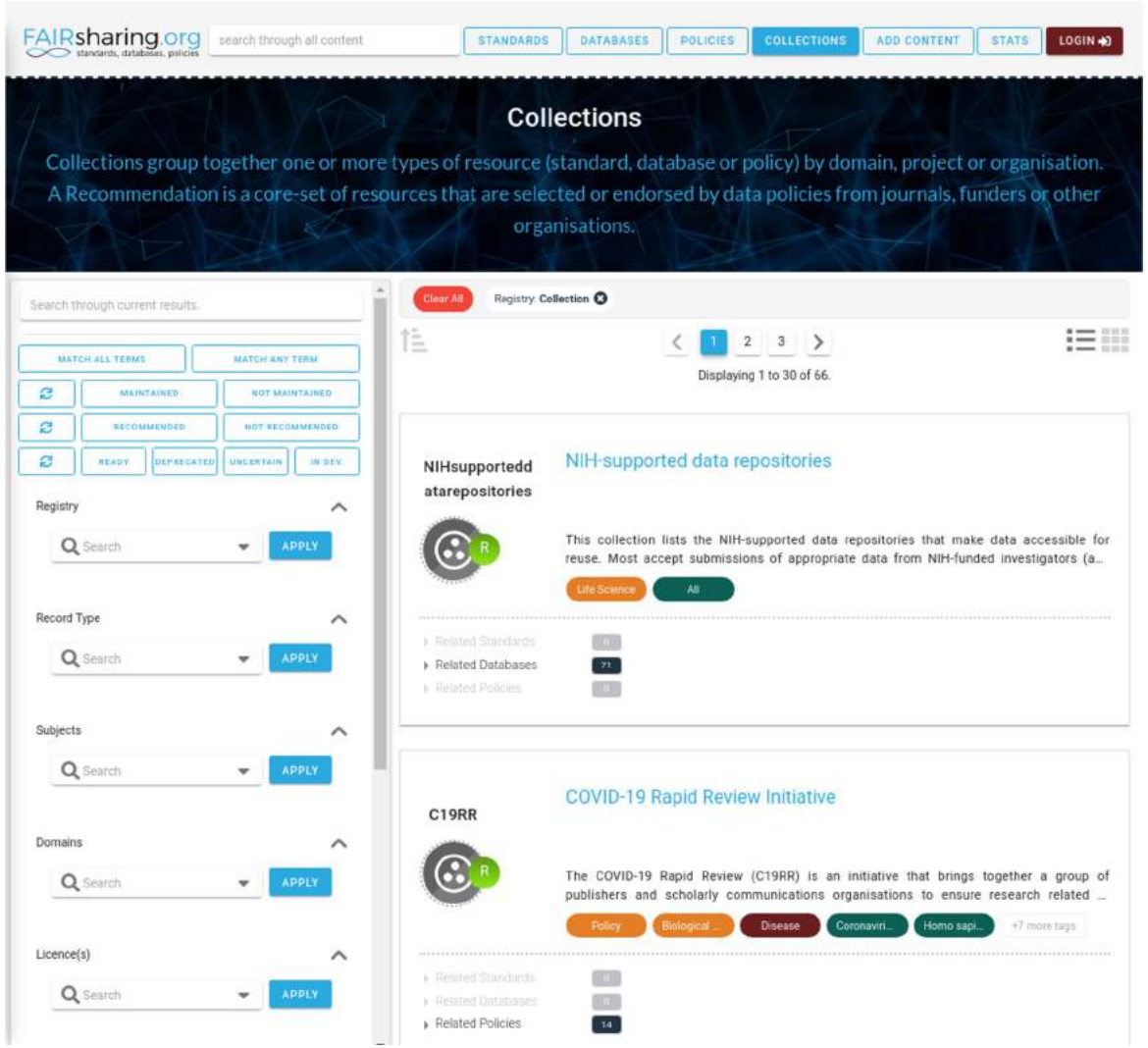
genomic STANDARDS consortium
isatools
GENEONTOLOGY Unifying Biology
Crop Ontology for agricultural data
DISEASE ONTOLOGY

Collections in the FAIRsharing portal

A *collection* includes standards and/or databases *grouped by domain, species or organization*

Graph view to visualize relationship links between resources

https://fairsharing.org/collections/

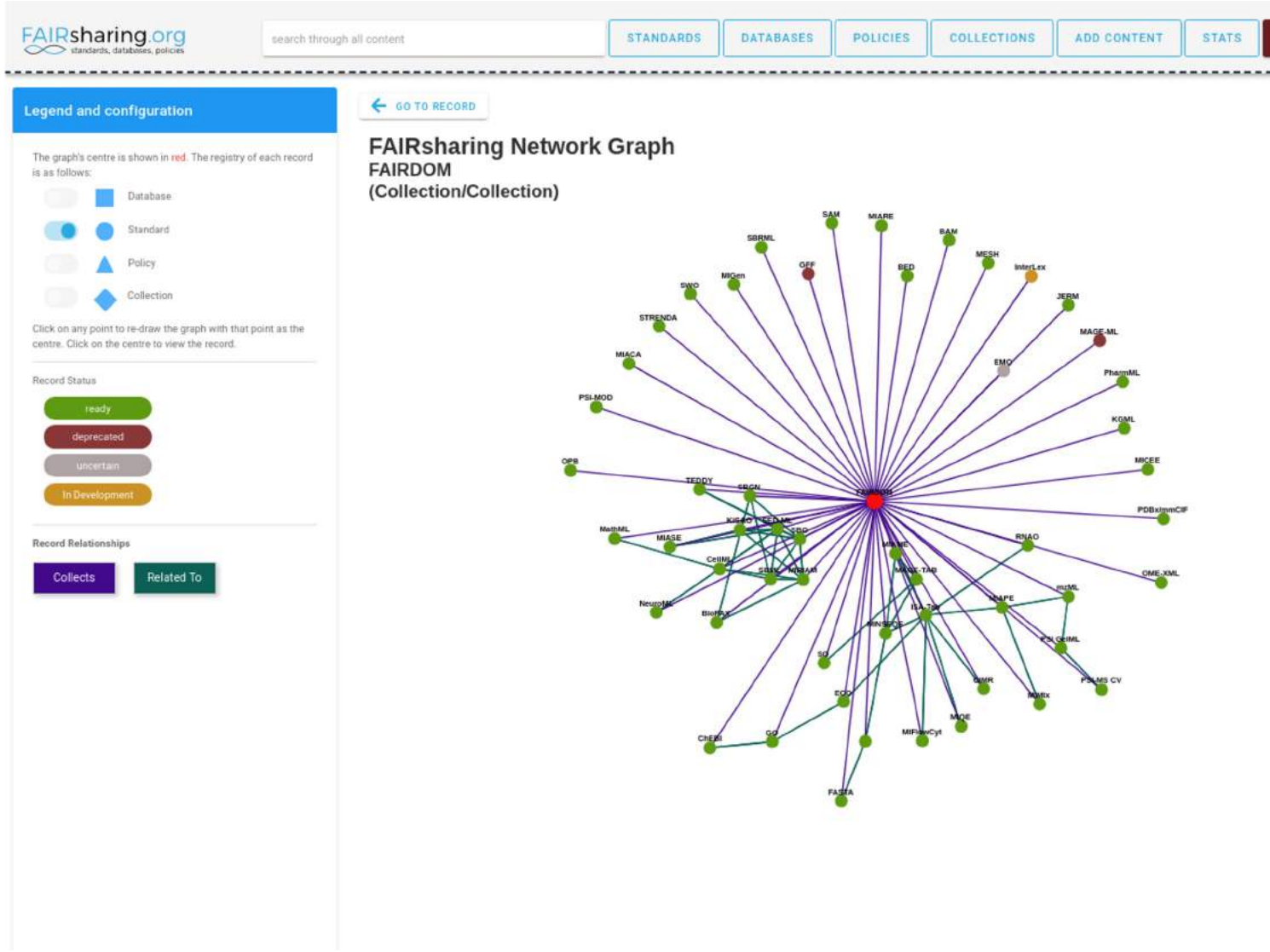


Collections in Life Sciences

52 collections related to Life Science standards in FAIRsharing

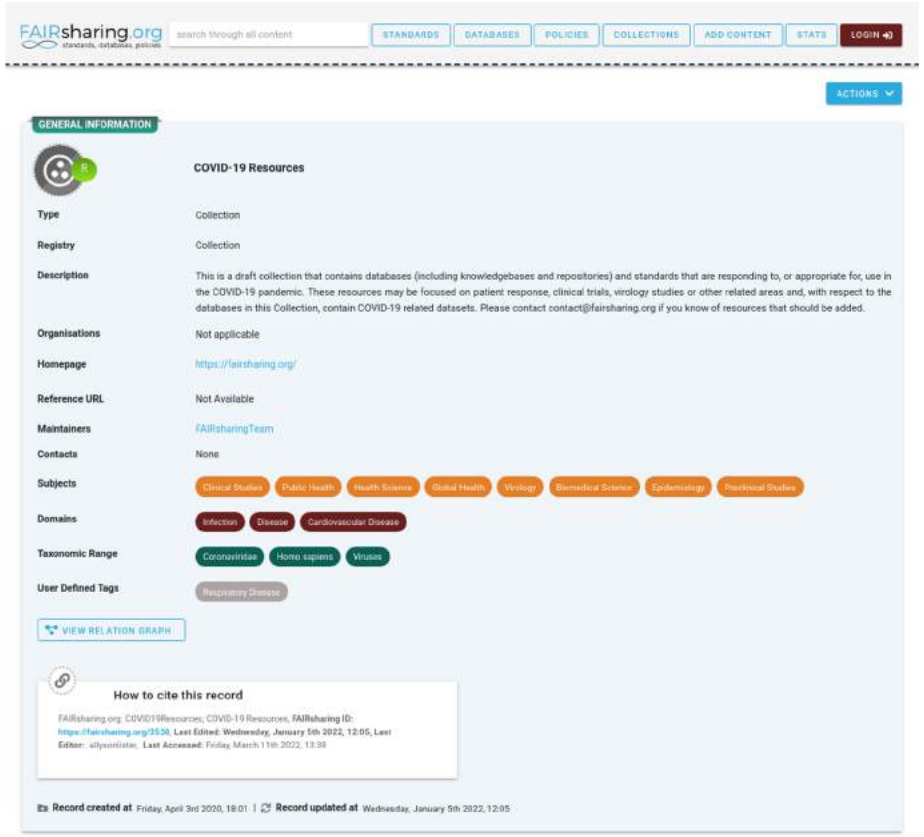
Example 1: the *FAIRdom community Standards collection* (System biology)

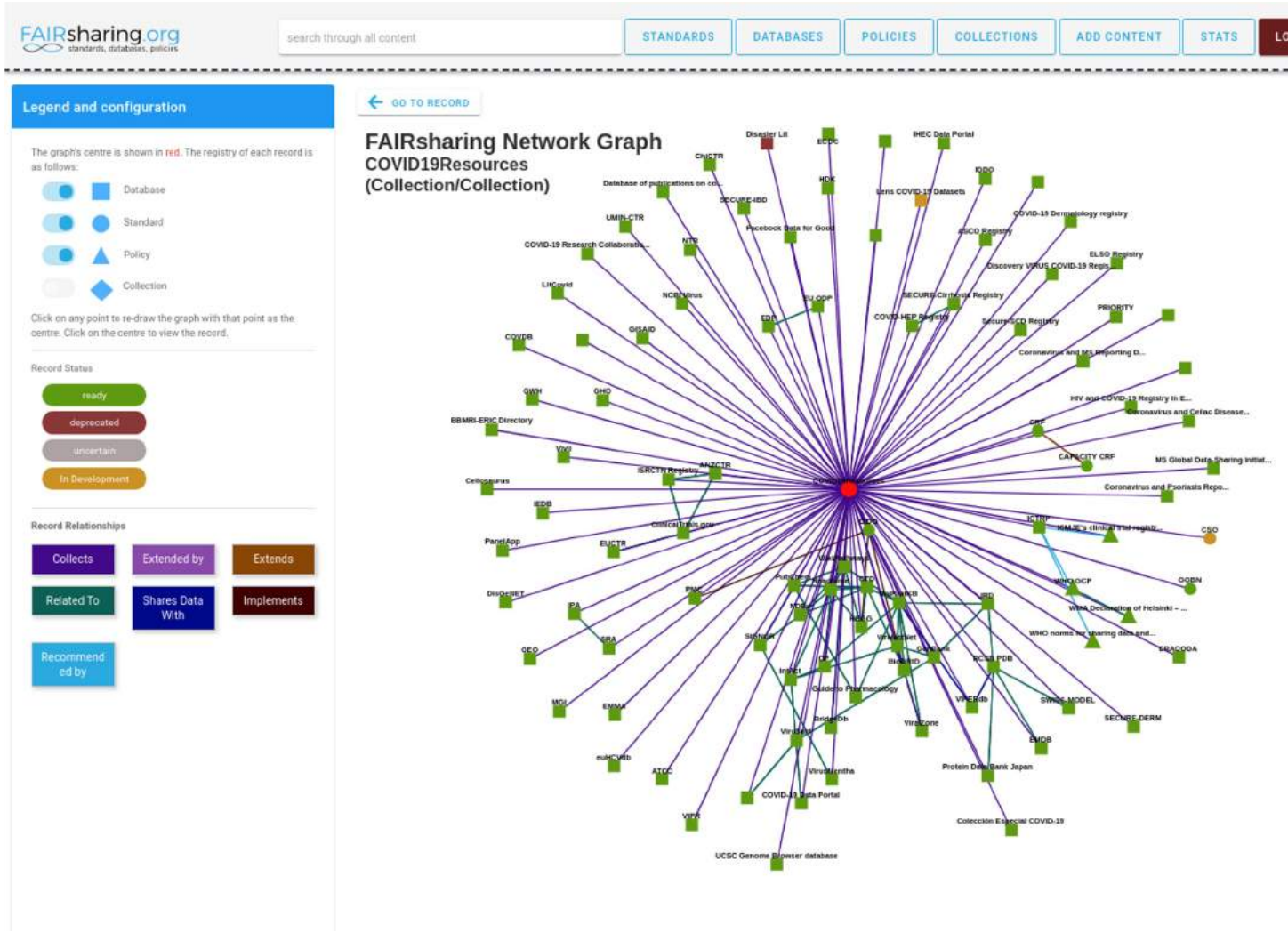
https://fairsharing.org/collection/FAIRDOM



Some collections are recent

Example 2: The *Covid-19* collection



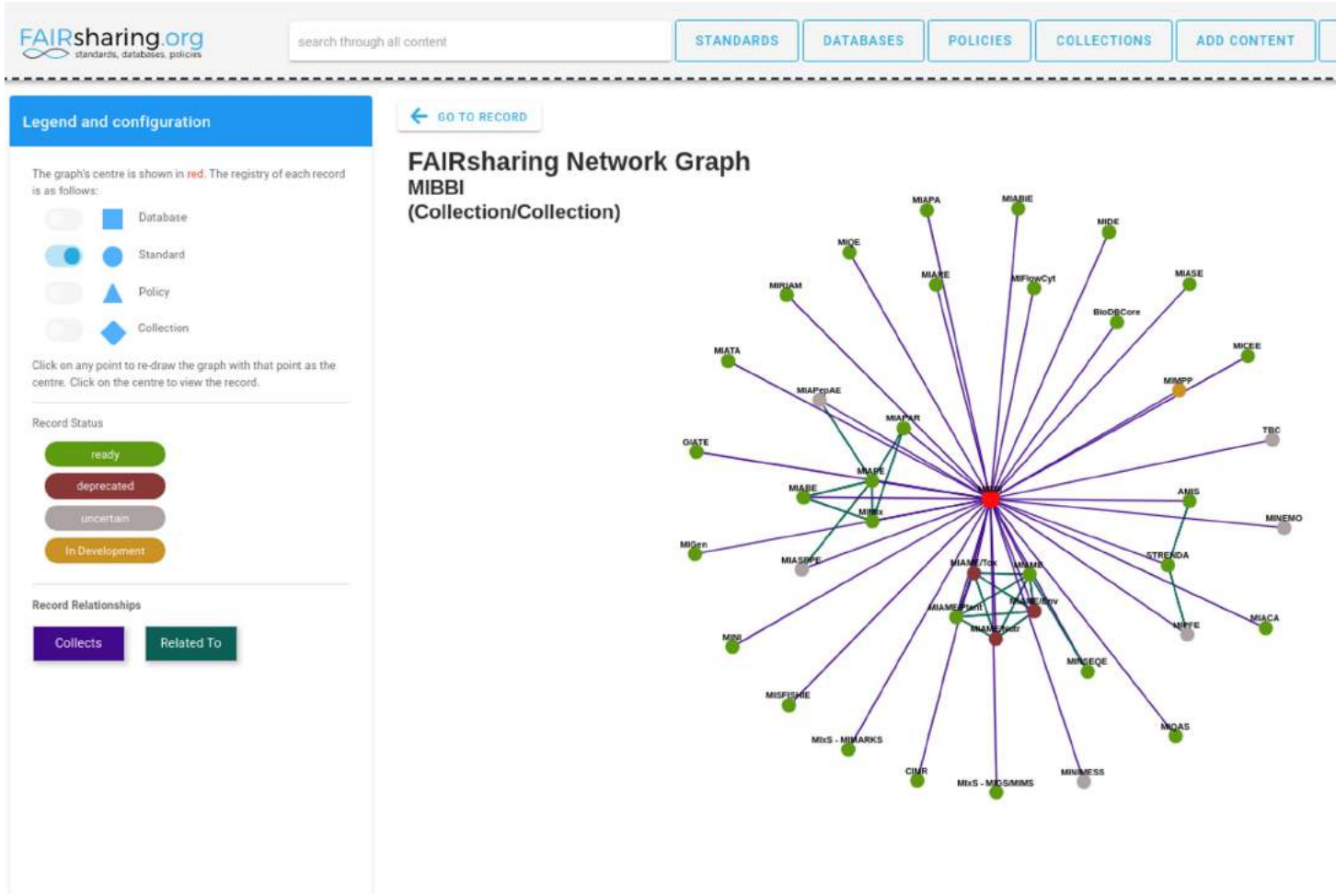


https://fairsharing.org/collection/COVID19Resources

https://fairsharing.org/graph/3538

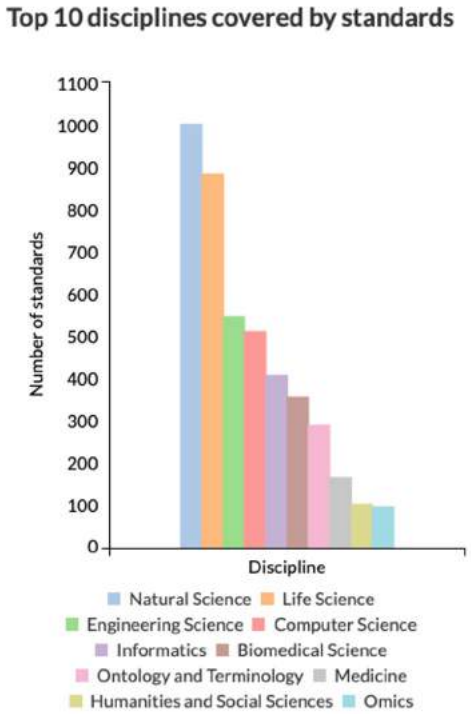
What about the minimum required metadata in biology?

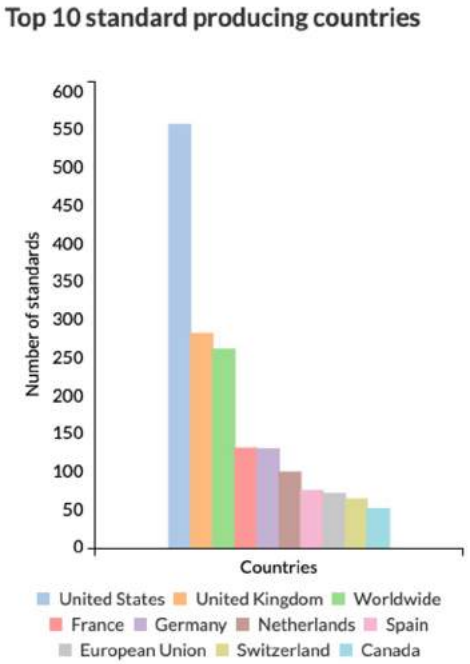
Example 3: the *Minimum Information for Biological and Biomedical Investigations* collection

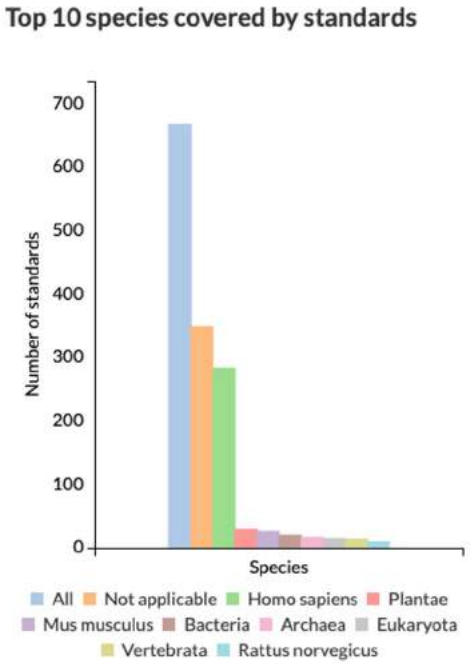


https://fairsharing.org/collection/MIBBI

Summary statistics about standards







Life Science is one of the best covered discipline

US and UK are the main standards producers

Human species is the best covered species

https://fairsharing.org/summary-statistics/?collection=standards

https://fairsharing.org/browse/subject

Practice

Find the *Genomic Standards Consortium (GSC) used by both ENA and SRA databases* in the **FAIRsharing collections**

Use both the record summary and the Graph visualization to interpret and answer the questions in zoom:

1. How many records (*i.e.* standards) are associated to the *GSC* ?
2. What type of standard is *Minimum Information about any (x) Sequence (MiXS)* ?
3. What is the record status of the *GAZ* record ?

Source: https://gensc.org

Practice => Answers

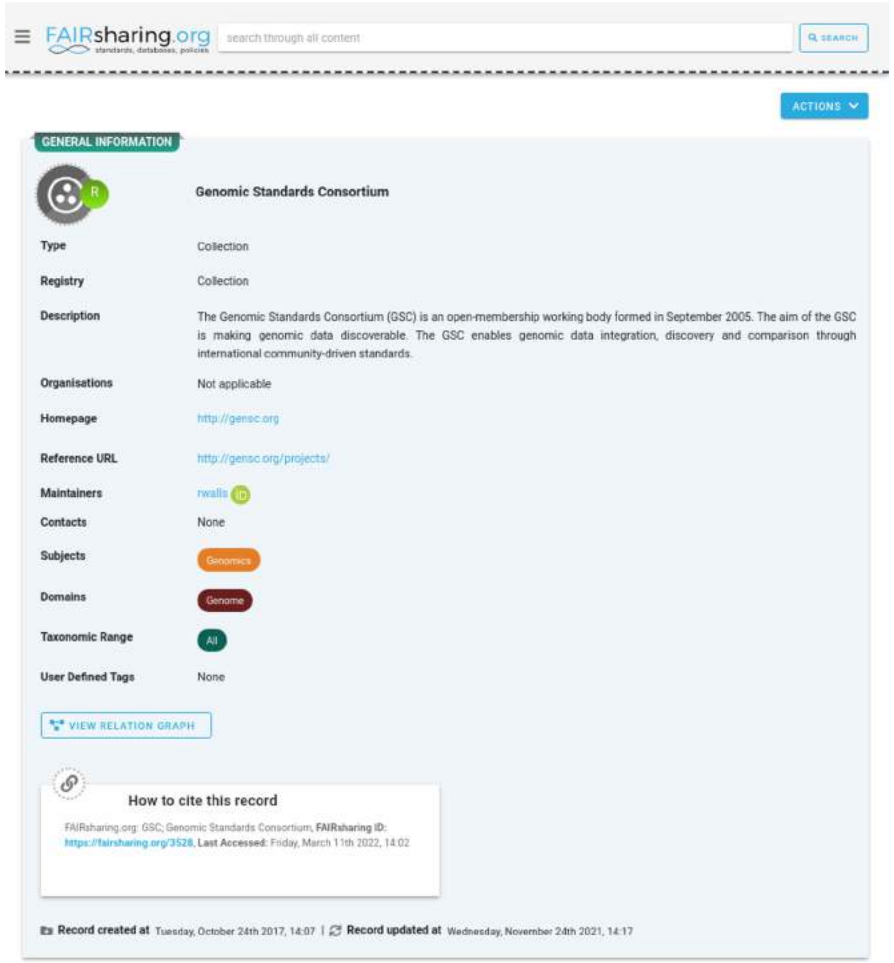
Find the *Genomic Standards Consortium (GSC) used by both ENA and SRA databases* in the **FAIRsharing collections**

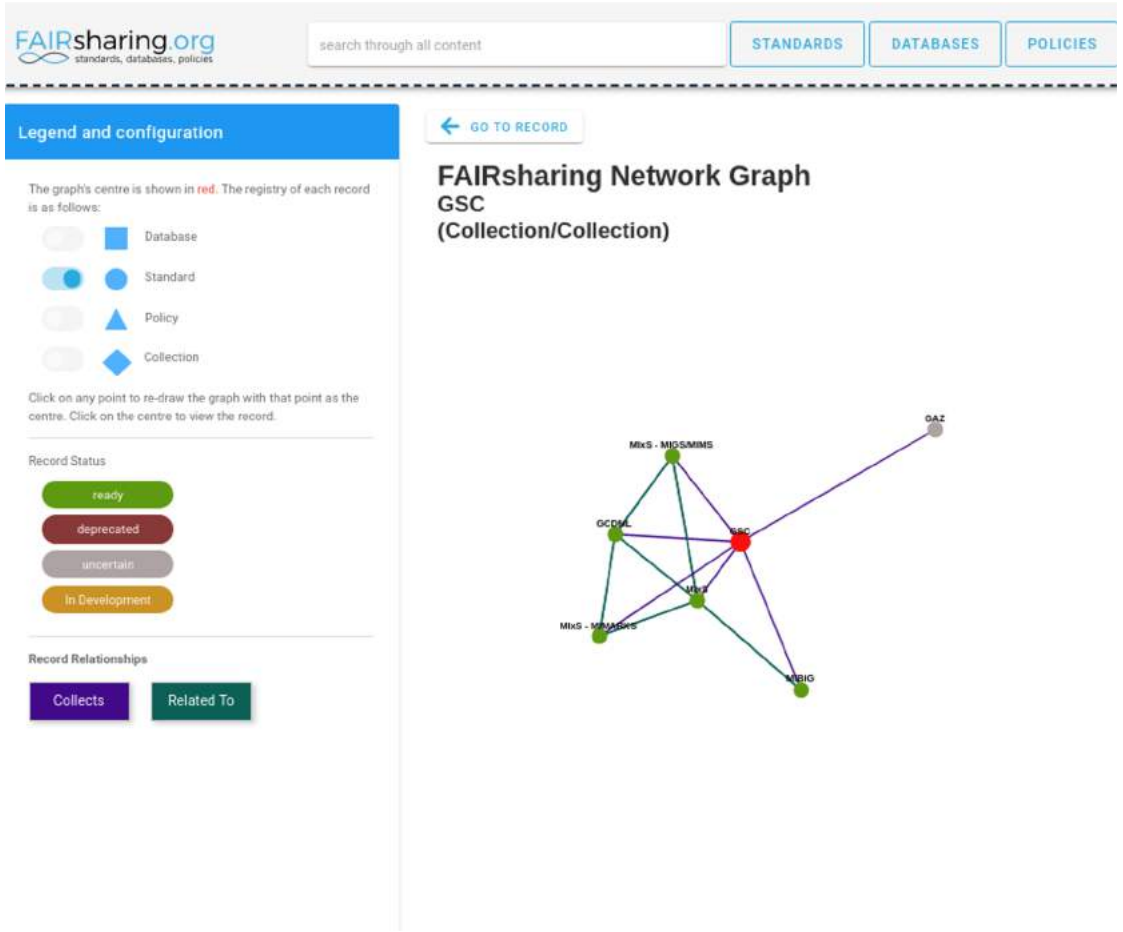
Use both the record summary and the Graph visualization to interpret and answer the questions in zoom:

1. How many records (*i.e.* standards) are associated to the *GSC* ? => 6
2. What type of standard is *Minimum Information about any (x) Sequence (MiXS)* ? => Reporting guideline
3. What is the record status of the *GAZ* record ?=>Uncertain

Source: https://gensc.org

The Genomic Standards Consortium (GSC)





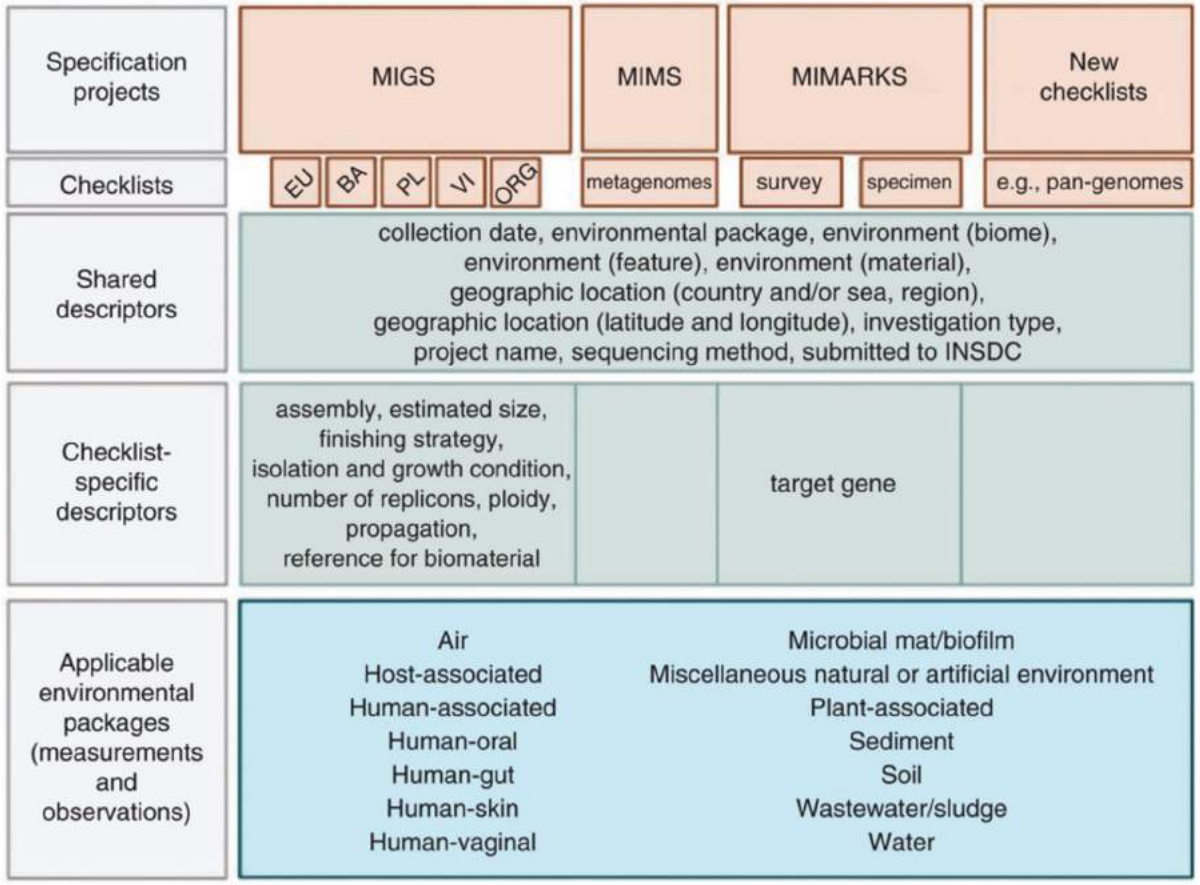
https://fairsharing.org/collection/GSC

https://fairsharing.org/graph/#/collection/bsg-c000040

The Genomic Standards Consortium (GSC)

- An international community-driven standard in **Genomics** producer of the ***MIxS: Minimum Information Standards about any(X) Sequence***

- MIxS includes **technology-specific checklists** (MIGS, MIMS, MIMARKS,...) and also allows **annotation of sample data** using environmental packages



Yilmaz et *al*, 2011

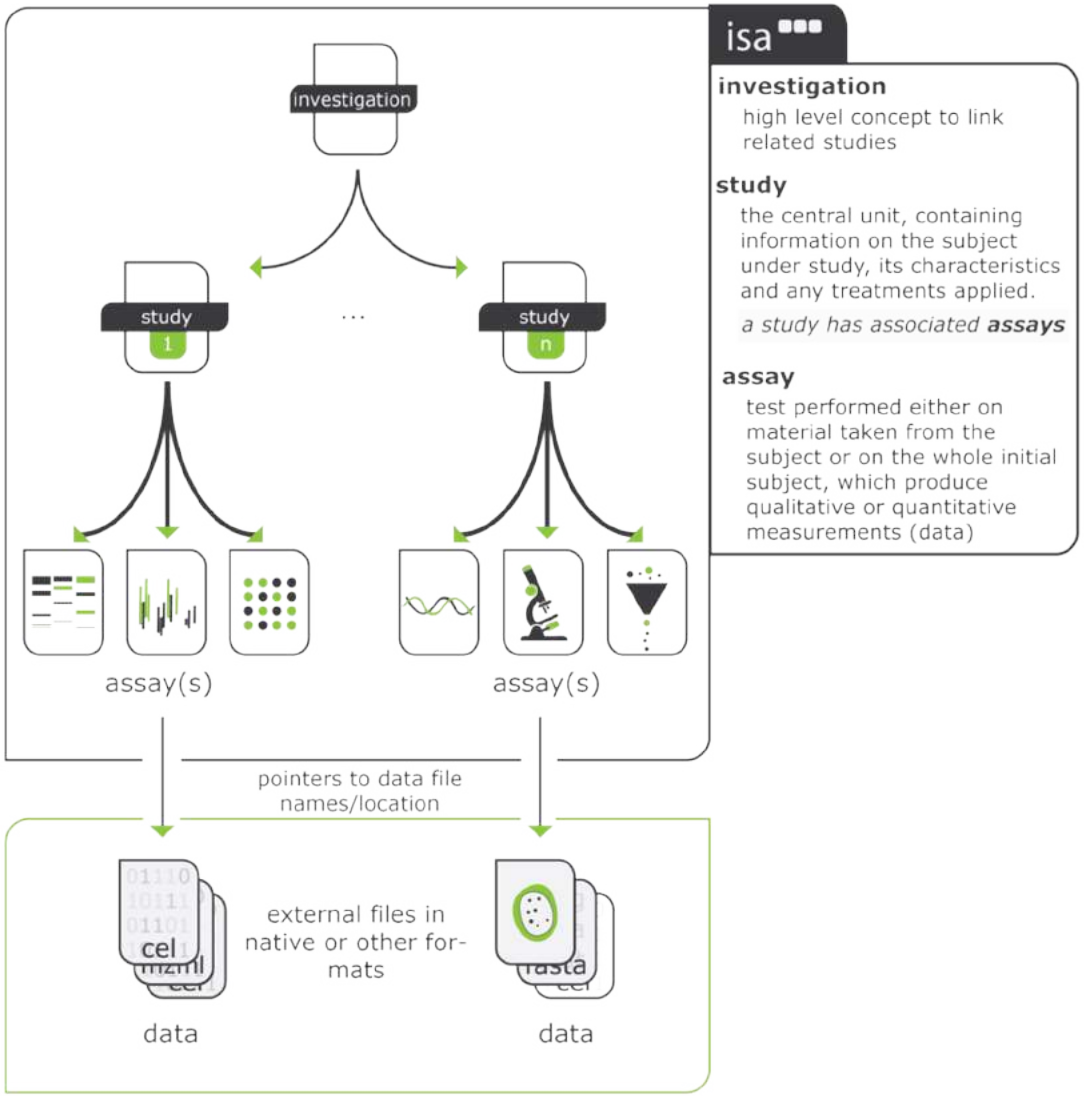
Source: https://gensc.org

The ISA model

A standard for Life ScienceData

A **model** to capture **experimental metadata** through **3 core entities**:

- **Investigation**: the project context
- **Study**: an experimentation in one location
- **Assay:** a specific measurement that targets a trait with a method and a scale

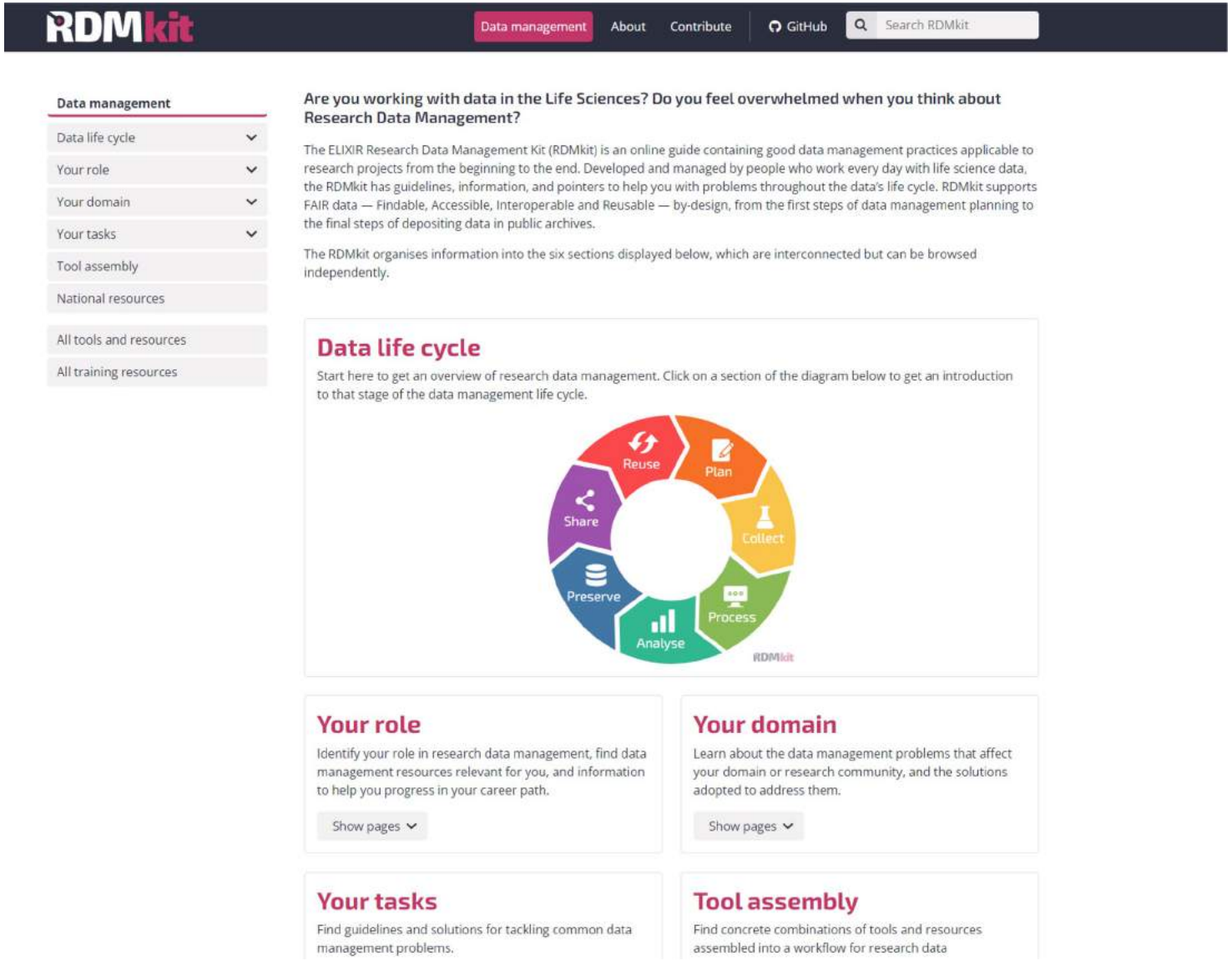


ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Rocca-Serra P et al. **Bioinformatics 2010**. https://doi.org/10.1093/bioinformatics/btq415

Sources: https://isa-tools.org and : https://isa-specs.readthedocs.io/en/latest/isamodel.html

RDMkit

The ELIXIR Research Data Management Kit (RDMkit) is an online guide containing **good data management practices** applicable to research projects from the beginning to the end.



To conclude: sources & useful links

Description	Name	URL
A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies.	FAIRsharing portal	https://fairsharing.org
Investigation, Study, Assay (ISA) ressource: A standard model an a set of tools to capture experimental data in life sciences	ISAtools	https://isa-tools.org
Genomics Standard Consortium (GSC): An international consortium developing standards and checklists in genomics	GSC	https://gensc.org
RDMkit: Documentation and metadata	RDMkit documentation and metadata	https://rdmkit.elixir-europe.org/metadata_management.html

Thanks

Paulette Lieby

Jean-François Dufayard

Frédéric de Lamotte

Hélène Chiapello

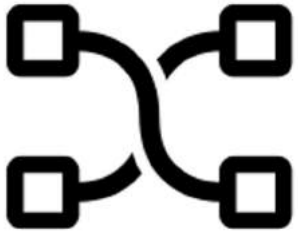
Thomas Denecker

Supplementary slides

Standard for data and metadata

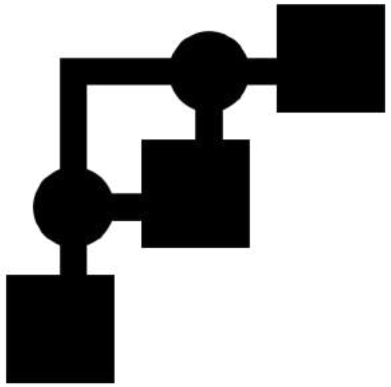
Guidelines or checklists

Ex: the GSC checklist



Models or schemas

Ex: ISA model



Terminology artefacts, ontology

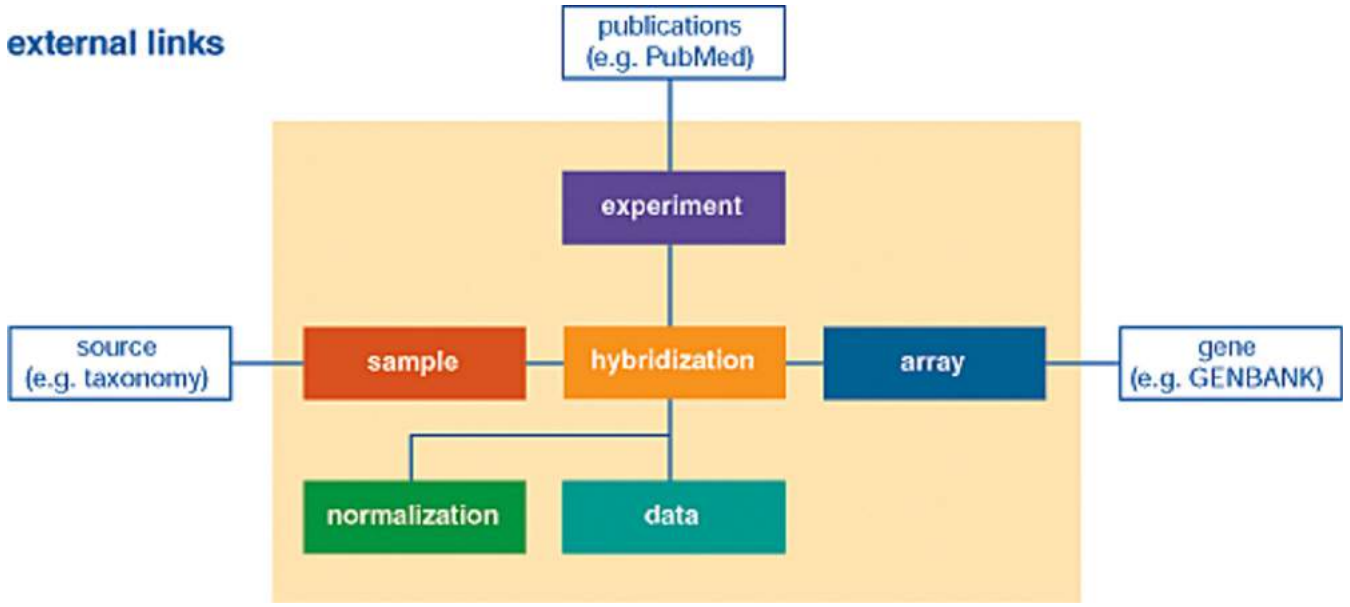
Ex: The Gene Ontology

Identifier schemata

https://fairsharing.org

The Minimum information standard initiative

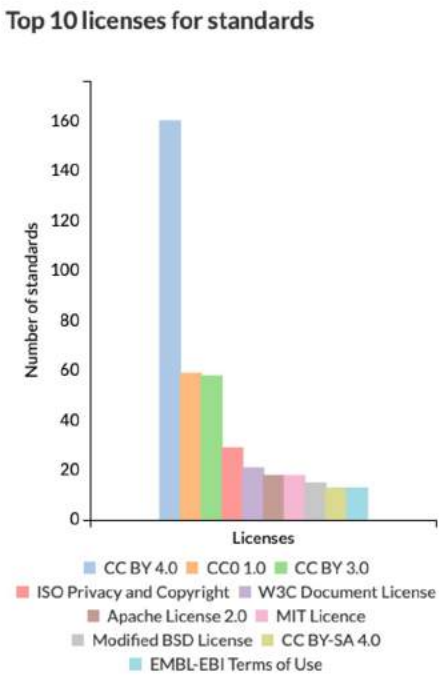
- A set of **guidelines** for reporting data derived by relevant methods in biosciences.
- Example : the **Minimum Information About a Microarray Experiment (MIAME)**

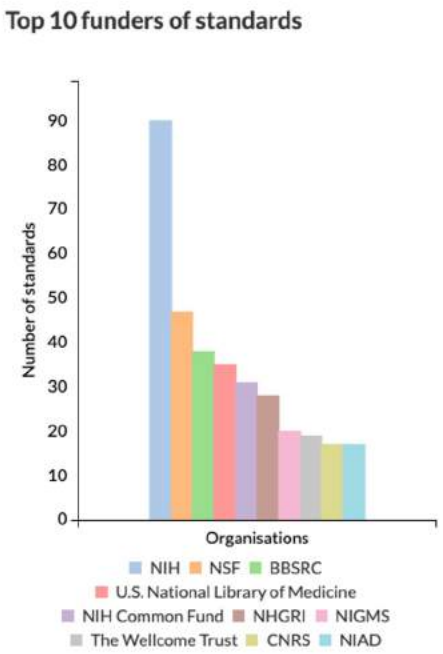


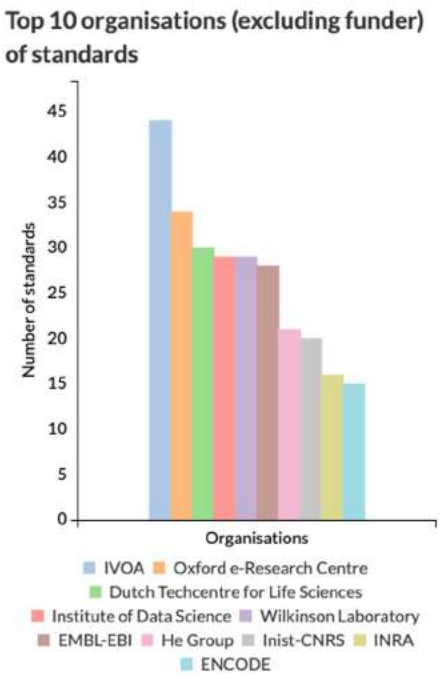
A schematic representation of six components of a microarray experiment.

https://en.wikipedia.org/wiki/Minimum_information_standard
10.1038/ng1201-365

Summary statistics about standards







https://fairsharing.org/summary-statistics/?collection=standards

The CC by 4.0 licence is the most adopted

US and UK National institutes are the most important funders

Worldwide Research Organisations produce standards

DMP and Open Data training
Session 3: Metadata

Retour d'expérience de soumission en banque de données internationales

valerie.cognat@ibmp-cnrs.unistra.fr

&

laurent.bouri@igbmc.fr

Hélène Chiapello : https://orcid.org/0000-0001-5102-0632
Thomas Denecker : https://orcid.org/0000-0003-1421-7641

Pourquoi soumettre mes données ?

- Open science
- La reproductibilité des expériences
- Donner accès à mes données
- Archiver mes données
- Publication d’articles
- Analyser mes données

3 bases de données


ENA
European Nucleotide Archive

Qui a déjà soumis à l’ENA ?

C’était facile ?

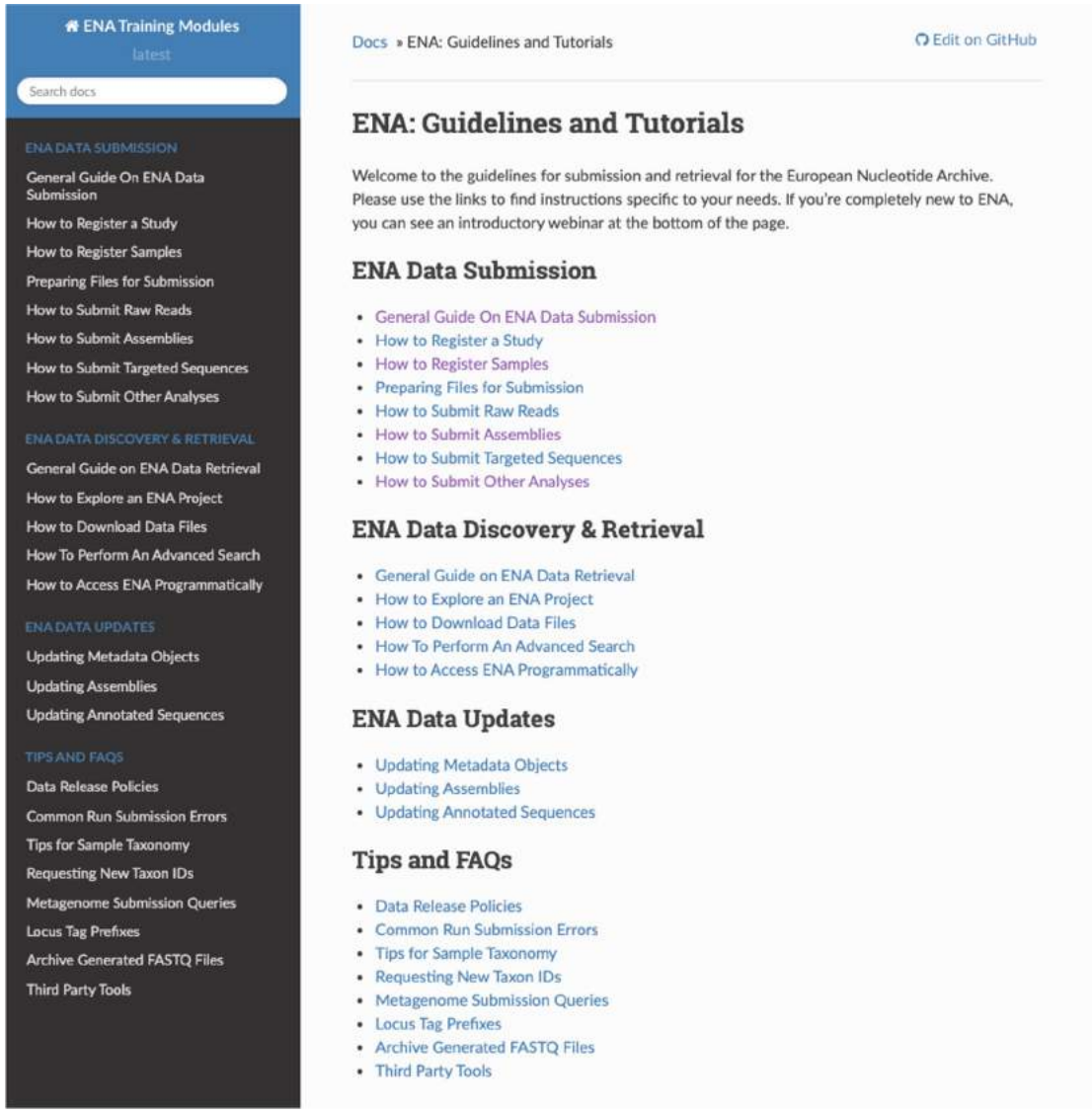
La base de données

Plateforme ouverte pour la gestion, le partage, l'intégration, l'archivage et la diffusion des données de séquençage.

Connecté avec UniProt, RNAcentral, EBI Metagenomics, Ensembl, Ensembl Genomes, ArrayExpress, ...

Des données variées: génomique animale, la biotechnologie marine, la biodiversité, la surveillance des agents pathogènes et la biologie des cellules souches

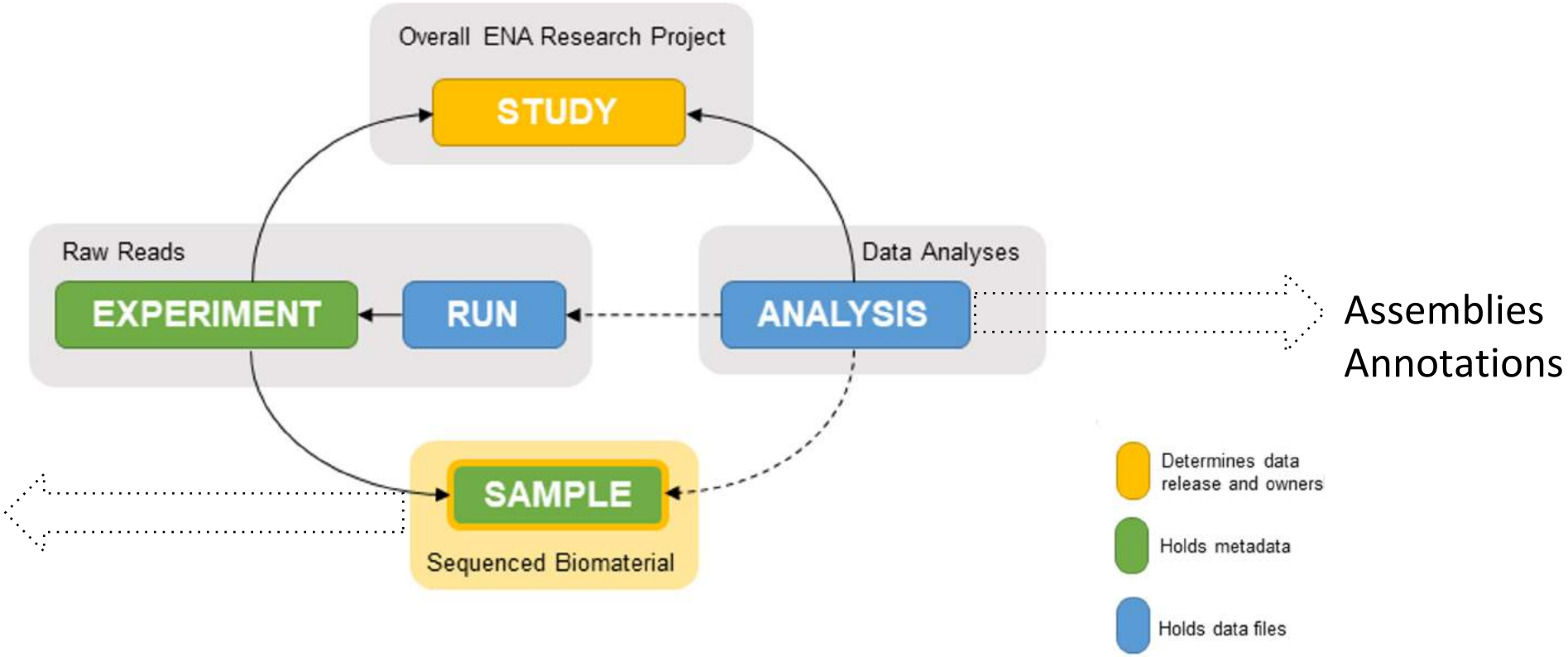
La documentation



https://ena-docs.readthedocs.io/en/latest/

Modèle des métadonnées

ISA compliant !



All **samples** submitted to ENA must conform to a **Checklist**

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

Description des expériences et validation

Metadata validation

Permitted values for platform

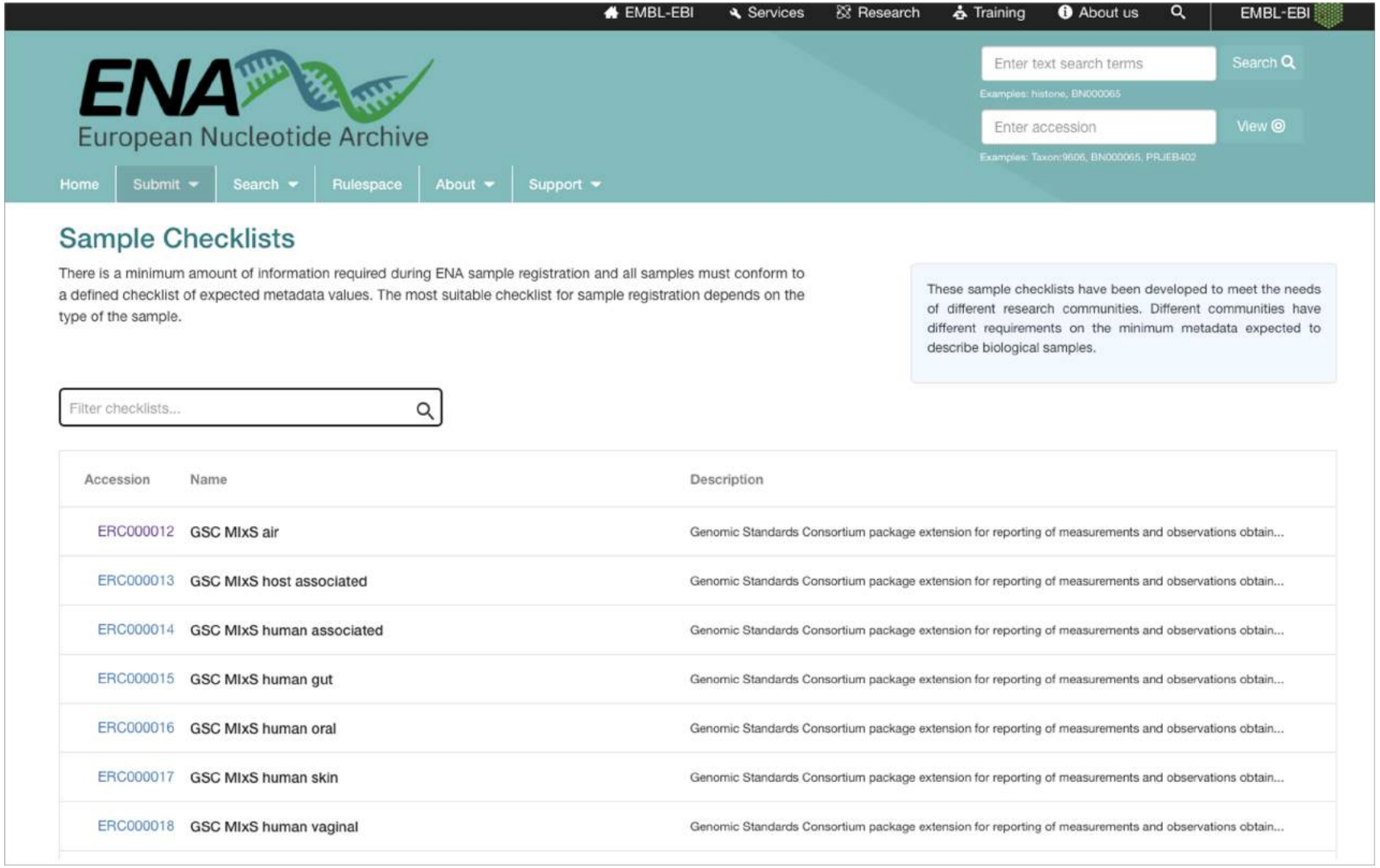
- LS454: 454 technology use 1-color sequential flows
- ILLUMINA: Illumina is 4-channel flowgram with 1-to-1 mapping between basecalls and flows
- PACBIO_SMRT: PacificBiosciences platform type for the single molecule real time (SMRT) technology.
- ION_TORRENT: Ion Torrent Personal Genome Machine (PGM) from Life Technologies.
- CAPILLARY: Sequencers based on capillary electrophoresis technology manufactured by LifeTech (formerly Applied BioSciences).
- OXFORD_NANOPORE: Oxford Nanopore platform type. nanopore-based electronic single molecule analysis.
- BGISEQ
- DNBSEQ

https://ena-docs.readthedocs.io/en/latest/submit/reads/webin-cli.html?permitted-values-for-instrument

Les checklists de l'ENA pour les "samples"

- A **checklist** defines the **minimum and optional metadata** expected to describe biological samples
- ENA are based on the **Genomic Standards Consortium (GSC)** recommandations
- The **most suitable checklist** depends on the type of the sample: https://www.ebi.ac.uk/ena/browser/checklists
- All ENA checklist are defined by an **access number** like ERCxxx (Ena R Checklist xxx)
  - example: GSC MIxS plant associated https://www.ebi.ac.uk/ena/browser/view/ERC000020

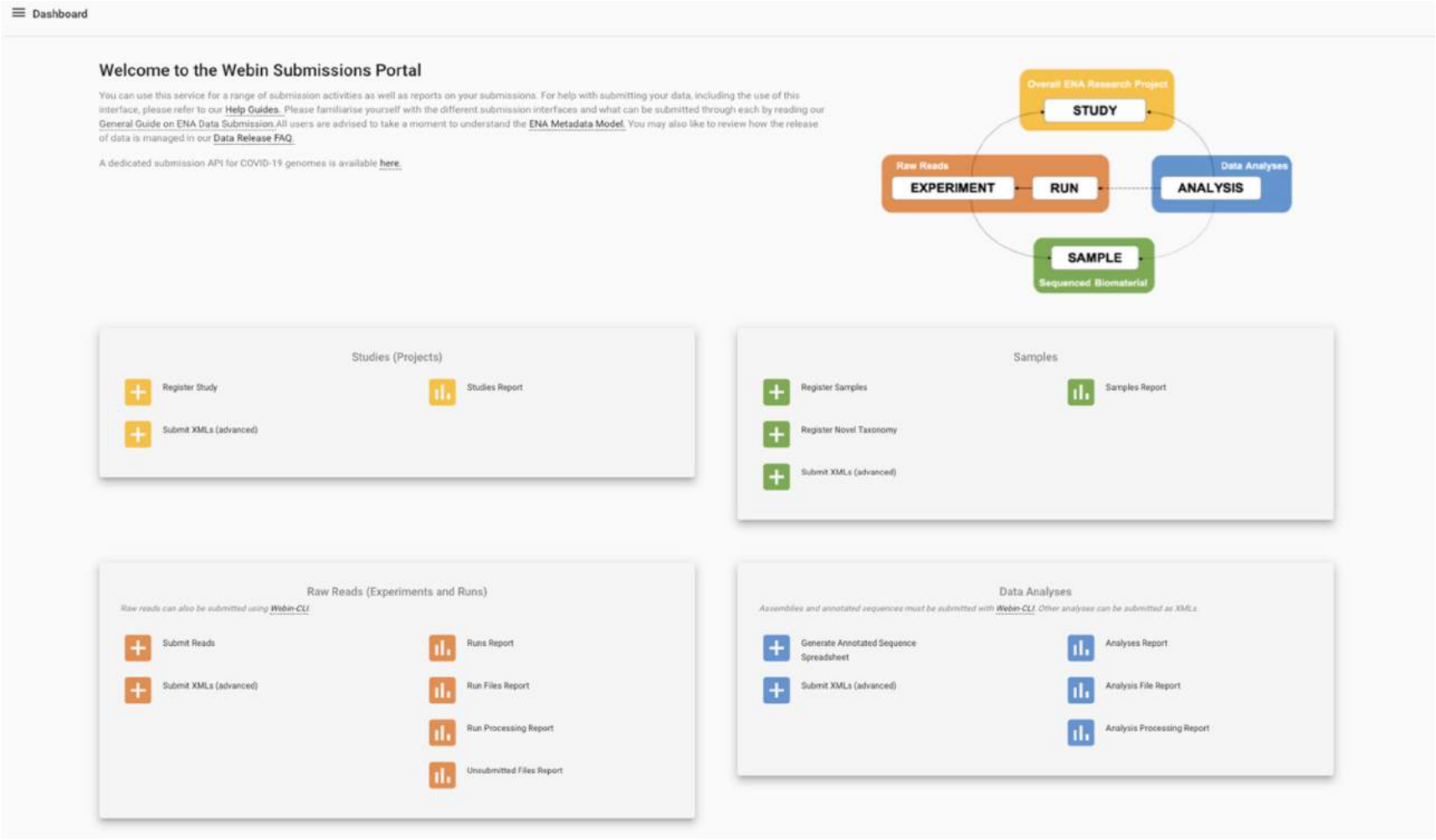
Listes des checklists pour les “Sample”



Méthodes de soumission

	Interactive	Webin-CLI	Programmatic
Study	Y	N	Y
Sample	Y	N	Y
Read data	Y	Y	Y
Genome Assembly	N	Y	N
Transcriptome Assembly	N	Y	N
Template Sequence	N	Y	N
Other Analyses	N	N	Y

Interactive



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

Web-Cli

v4.2.1 Latest

Compare

Rajkumar-D released this 26 days ago v4.2.1 0d34c7a

- sequence context: Added support for BioSample accessions, SRA Sample accessions and SRA Sample aliases in the ORGANISM field in addition to the already supported NCBI taxonomy names and IDs.

Assets 4

webin-cli-4.2.1-sources.jar	109 KB
webin-cli-4.2.1.jar	61.5 MB
Source code (zip)	
Source code (tar.gz)	

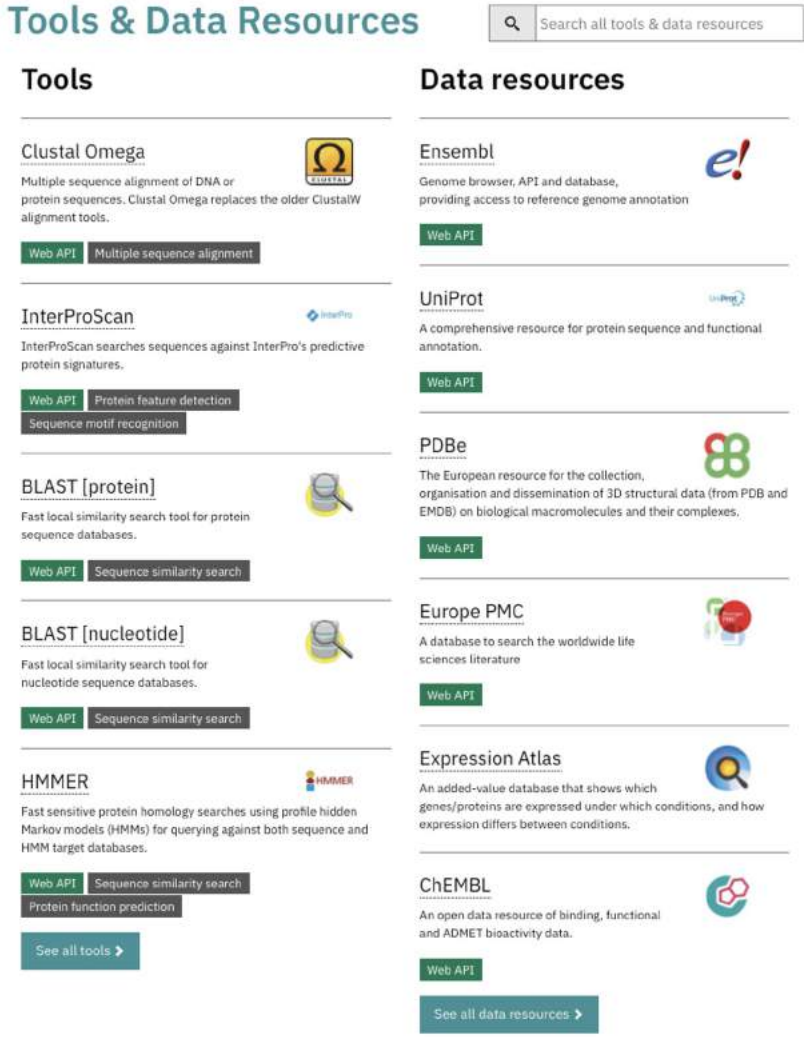
Programmatic

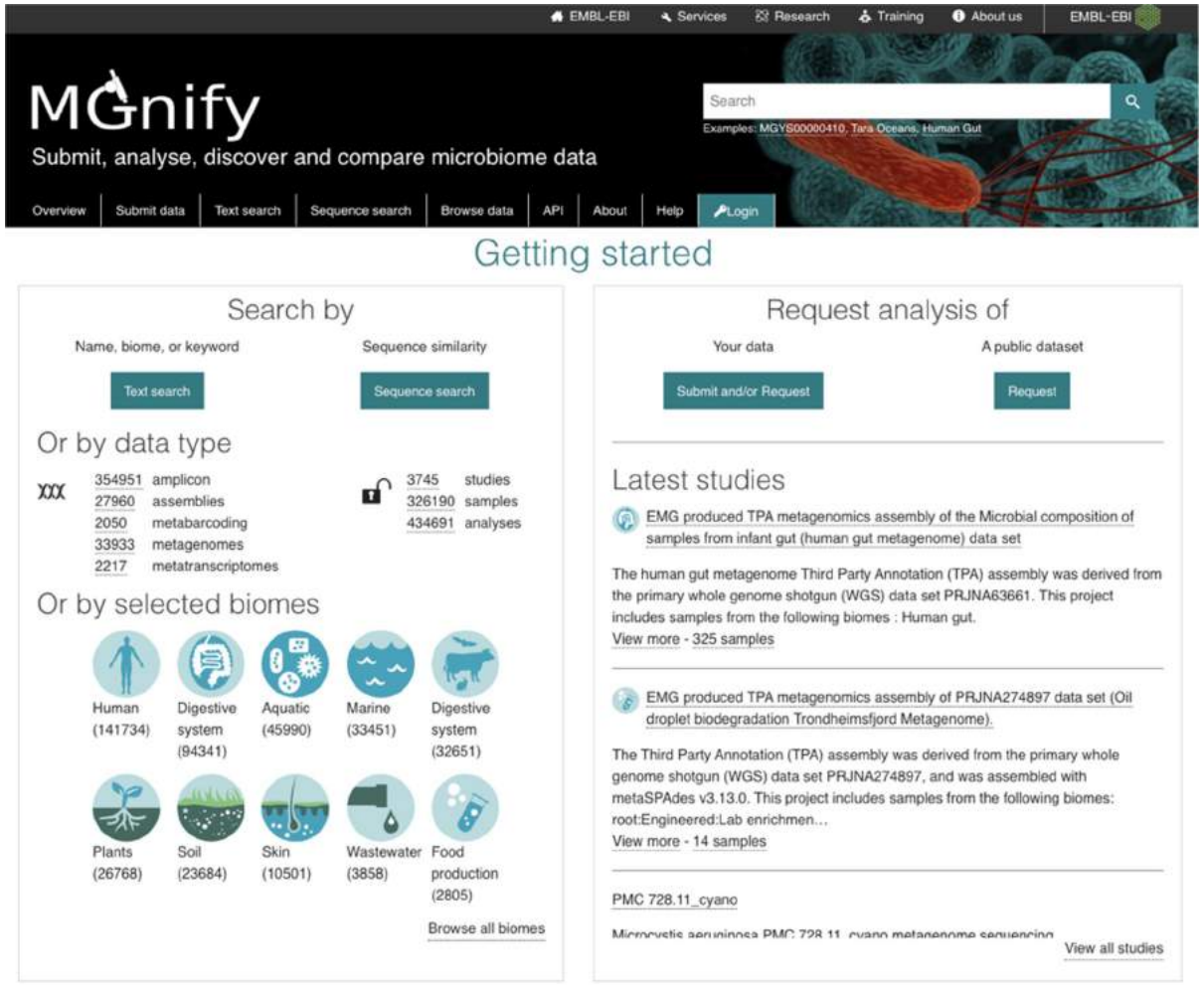
- SUBMISSION (XML Schema)
- STUDY (XML Schema)
- SAMPLE (XML Schema)
- EXPERIMENT (XML Schema)
- RUN (XML Schema)
- ANALYSIS (XML Schema)
- DAC (XML Schema)
- POLICY (XML Schema)
- DATASET (XML Schema)
- PROJECT (XML Schema)

Exemple : submission.xml

```
<SUBMISSION>
     <ACTIONS>
          <ACTION>
               <ADD/>
          </ACTION>
    </ACTIONS>
</SUBMISSION>
```

Les outils complémentaires





https://www.ebi.ac.uk/services/all


GEO
Gene Expression Omnibus

Qui a déjà soumis à GEO ?

C’était facile ?

La base de données

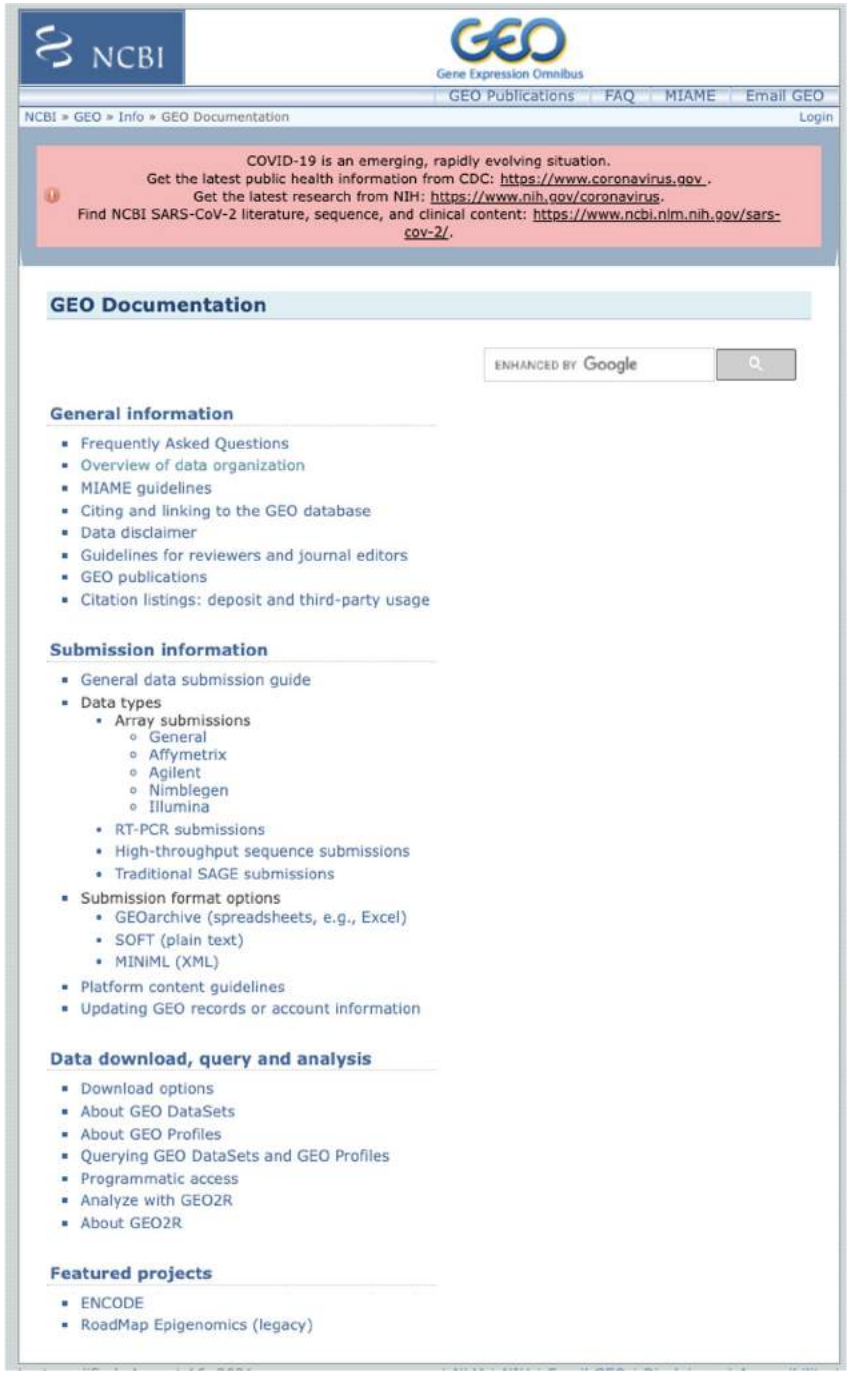
GEO est un dépôt public international qui archive et distribue librement des données de:

- microarray ;
- de NGS ;
- et d'autres formes de données de génomique fonctionnelle à haut débit .

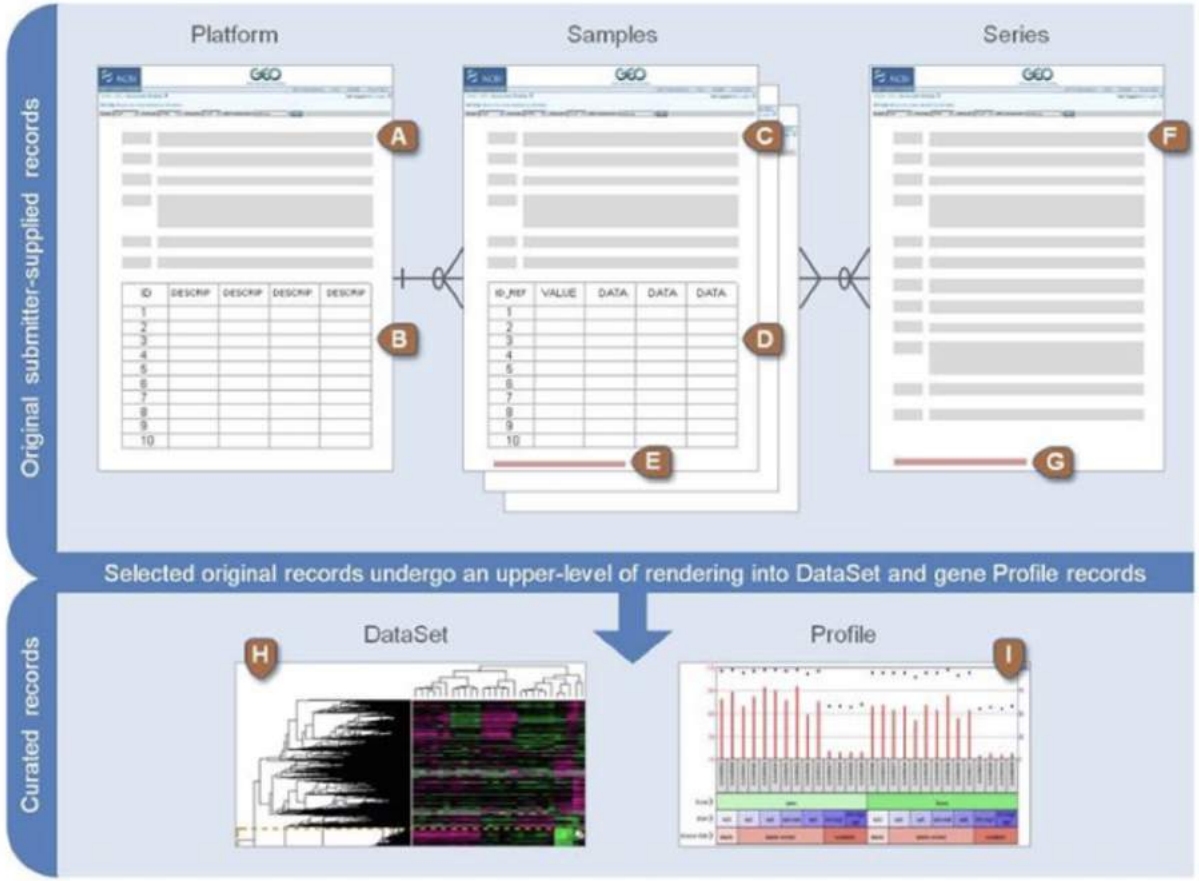
soumises par la communauté des chercheurs.

Documentation

https://www.ncbi.nlm.nih.gov/geo/info/



Organisation des données



Platform

Platform records are supplied by submitters

A Platform record is composed of a summary description of the array or sequencer and, for array-based Platforms, a data table defining the array template.Each Platform record is assigned a unique and stable GEO accession number (GPLxxx). A Platform may reference many Samples that have been submitted by multiple submitters.
Example Platform record »

Sample

Sample records are supplied by submitters

A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it. Each Sample record is assigned a unique and stable GEO accession number (GSMxxx). A Sample entity must reference only one Platform and may be included in multiple Series.
Example Sample record »

Series

Series records are supplied by submitters

A Series record links together a group of related Samples and provides a focal point and description of the whole study. Series records may also contain tables describing extracted data, summary conclusions, or analyses. Each Series record is assigned a unique and stable GEO accession number (GSExxx).
Example Series record »

A	**Text description of the array or sequencer**
B	**Text tab-delimited table of the array template**
C	**Text description of the biological sample and protocols to which it was subjected**
D	**Text tab-delimited table of processed hybridization result** (may optionally include raw data columns)
E	**Original raw data file, or processed sequence data file**
F	**Text description of the overall experiment**
G	**Tar archive of original raw data files, or processed sequence data files**

https://www.ncbi.nlm.nih.gov/geo/info/overview.html

Fichiers

GEOarchive format

GEOarchive is a flexible spreadsheet-based submission format useful for batch deposit of experiments. GEOarchive submissions can be created in any spreadsheet software, usually Microsoft Excel.

A GEOarchive submission consists of several parts as follows:

Metadata spreadsheet	'Metadata' refers to descriptive information and protocols for the overall experiment and individual Samples. This information is supplied by completing all fields of the appropriate metadata spreadsheet template which can be downloaded from the GEOarchive templates and examples section below.
Matrix table	The matrix table is a spreadsheet containing the final, normalized values that are comparable across rows and Samples, and preferably processed as described in any accompanying manuscript. A complete data matrix should be supplied, not a summary subset. It is possible to include additional data columns in the table, for example, Affymetrix Detection calls and P-values, or background or flag columns. See the Affymetrix template for an example.
Raw data files	In addition to the normalized data provided in the Matrix table, submitters are required to provide raw data, usually in the form of supplementary raw data files. This facilitates the unambiguous interpretation of the data and potential verification of the conclusions as described in the MIAME and MINSEQE standards. Affymetrix submissions must include CEL files. Non-Affymetrix GEOarchive submissions should include the original software-generated scan quantification files, for example, GenePix GPR files. Next-generation sequence submissions must include files containing reads and quality scores.
Platform	If your experiments are performed using a commercial array (e.g., Affymetrix GeneChip) or other array already deposited in GEO, please use the FIND PLATFORM tool to find the GEO accession number (GPLxxxx) for inclusion in the '*platform*' column in the *SAMPLES* section of the metadata spreadsheet. If your array does not already exist in GEO, please include a *PLATFORM* section in your metadata spreadsheet and include Platform annotation columns in your matrix table. The Platform data must include meaningful, trackable, sequence identifiers (e.g. GenBank/RefSeq accessions, locus tags, clone IDs, oligo sequences, chromosome locations, etc - see the Platform content guidelines for full list). References to in-house databases or top BLAST hits are not sufficient. Platform submission is not necessary for SAGE or next-generation sequence submissions.

Bundle all parts (Excel file containing the metadata spreadsheet and matrix spreadsheet, raw data files) together into a .zip, .rar, or .tar archive using a program like WinZip, and transfer to GEO using the 'Transfer files to GEO with web form' option on the Submit to GEO page. Incomplete submissions will result in processing delays.

Submit

GEOarchive templates and examples

The first step in creating your GEOarchive submission is to download the appropriate template (Excel spreadsheet) from the list below. Each Excel file consists of several worksheets, including a metadata template, and examples of metadata and matrix tables. Click the tabs at the bottom of the worksheet window to switch between worksheets. Mouse over field names in the templates to view content guidelines.

Microarray

For the following microarray vendors, please download templates from the vendor-specific instructions pages:

- Affymetrix submissions
- Agilent submissions
- Nimblegen submissions
- Illumina submissions

For microarrays not from the vendors above, please use a 'Generic' template. For generic microarray submissions where the Platform is already deposited in GEO, please download the most appropriate template:

- Generic single channel submission template
- Generic dual channel submission template
- Generic merged dye-swap submission template
- Generic tiling ChIP-chip submission template

For generic microarray submissions where the Platform is not deposited in GEO, please download the most appropriate template:

- Generic single channel submission template, including Platform
- Generic dual channel submission template, including Platform
- Generic merged dye-swap submission template, including Platform
- Generic tiling ChIP-chip submission template, including Platform

To submit only a Platform, please download the following template (this option is appropriate only if you have no hybridization or sequence data to deposit):

- Platform-only template

High-throughput sequencing

For high-throughput sequence submissions, please refer to full instructions at:

- High-throughput sequence submissions

Other data types

For NanoString submissions, please use one of the 'Generic single channel' templates as appropriate:

- Generic single channel submission template
- Generic single channel submission template, including Platform

For high-throughput RT-PCR submissions, please refer to full instructions at:

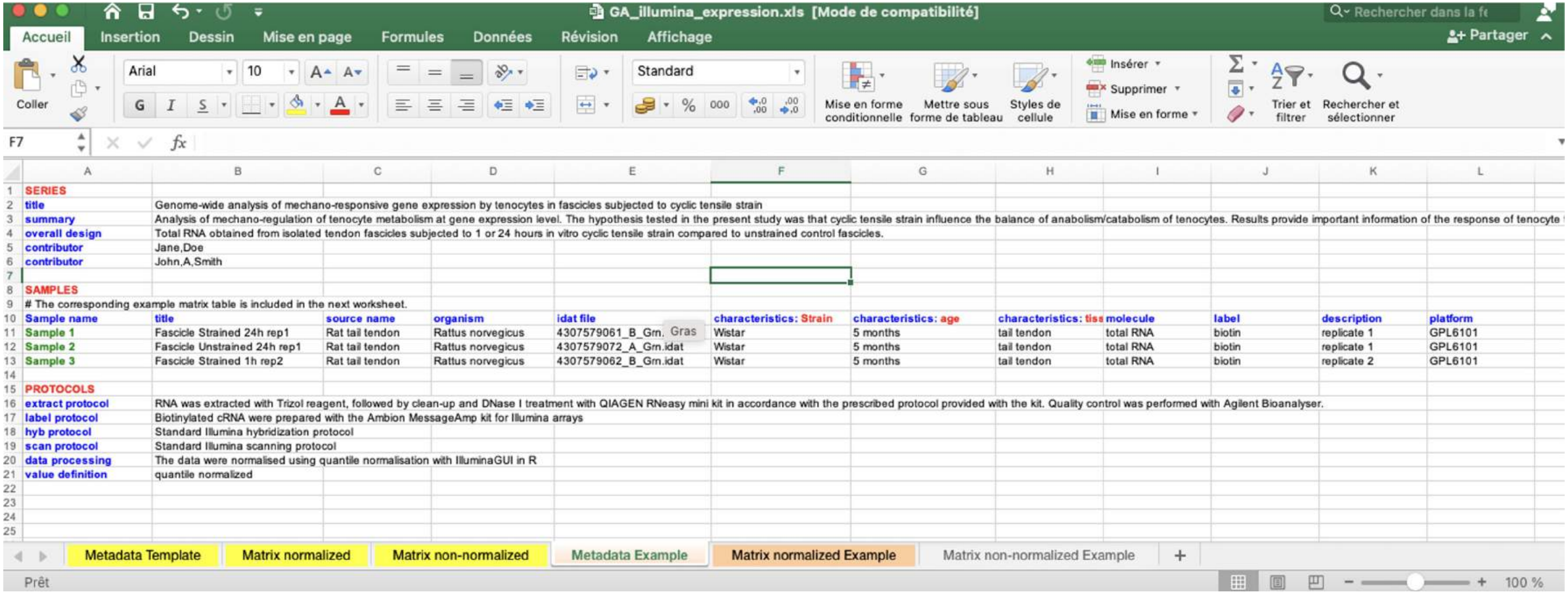
- RT-PCR submissions

For traditional SAGE submissions, please refer to full instructions at:

- Traditional SAGE submissions

https://www.ncbi.nlm.nih.gov/geo/info/spreadsheet.html#GAtemplates

Exemple Excel Illumina



GA_illumina_expression.xls [Mode de compatibilité]

	A	B	C	D	E	F	G	H	I	J	K	L
1	SERIES											
2	title	Genome-wide analysis of mechano-responsive gene expression by tenocytes in fascicles subjected to cyclic tensile strain										
3	summary	Analysis of mechano-regulation of tenocyte metabolism at gene expression level. The hypothesis tested in the present study was that cyclic tensile strain influence the balance of anabolism/catabolism of tenocytes. Results provide important information of the response of tenocyte										
4	overall design	Total RNA obtained from isolated tendon fascicles subjected to 1 or 24 hours in vitro cyclic tensile strain compared to unstrained control fascicles.										
5	contributor	Jane,Doe										
6	contributor	John,A,Smith										
7												
8	SAMPLES											
9	# The corresponding example matrix table is included in the next worksheet.											
10	Sample name	title	source name	organism	idat file	characteristics: Strain	characteristics: age	characteristics: tiss	molecule	label	description	platform
11	Sample 1	Fascicle Strained 24h rep1	Rat tail tendon	Rattus norvegicus	4307579061_B_Grn.	Wistar	5 months	tail tendon	total RNA	biotin	replicate 1	GPL6101
12	Sample 2	Fascicle Unstrained 24h rep1	Rat tail tendon	Rattus norvegicus	4307579072_A_Grn.idat	Wistar	5 months	tail tendon	total RNA	biotin	replicate 1	GPL6101
13	Sample 3	Fascicle Strained 1h rep2	Rat tail tendon	Rattus norvegicus	4307579062_B_Grn.idat	Wistar	5 months	tail tendon	total RNA	biotin	replicate 2	GPL6101
14												
15	PROTOCOLS											
16	extract protocol	RNA was extracted with Trizol reagent, followed by clean-up and DNase I treatment with QIAGEN RNeasy mini kit in accordance with the prescribed protocol provided with the kit. Quality control was performed with Agilent Bioanalyser.										
17	label protocol	Biotinylated cRNA were prepared with the Ambion MessageAmp kit for Illumina arrays										
18	hyb protocol	Standard Illumina hybridization protocol										
19	scan protocol	Standard Illumina scanning protocol										
20	data processing	The data were normalised using quantile normalisation with IlluminaGUI in R										
21	value definition	quantile normalized										
22												
23												
24												
25												

Metadata Template | Matrix normalized | Matrix non-normalized | Metadata Example | Matrix normalized Example | Matrix non-normalized Example

Prêt

100 %

Les outils complémentaires : GeoToR

exemple : GSE25724

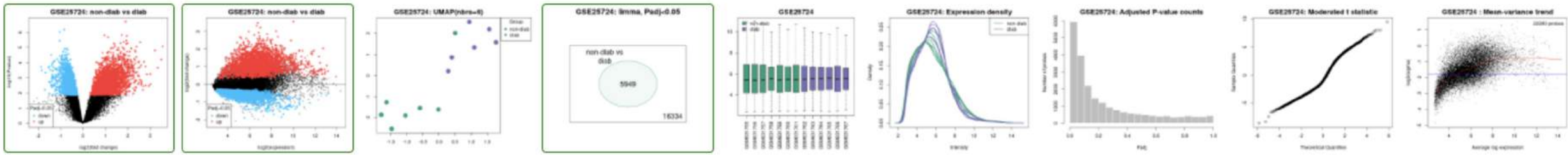
GEO accession GSE25724 Set — Expression data from type 2 diabetic and non-diabetic isolated human islets

Samples ▸ Define groups — Selected **13** out of **13** samples

Columns Set

Group	Accession	Title	Source name	Tissue	Disease state	Age	Gender	Characteristics
non-diab	GSM631755	Non-diabetic islets, rep1	human islets, non-diabetic	pancreatic islets	non-diabetic	47 yrs	male	bmi (kg/m2): 27.7
non-diab	GSM631756	Non-diabetic islets, rep2	human islets, non-diabetic	pancreatic islets	non-diabetic	33 yrs	male	bmi (kg/m2): 22.9
non-diab	GSM631757	Non-diabetic islets, rep3	human islets, non-diabetic	pancreatic islets	non-diabetic	47 yrs	male	bmi (kg/m2): 28.4
non-diab	GSM631758	Non-diabetic islets, rep4	human islets, non-diabetic	pancreatic islets	non-diabetic	54 yrs	male	bmi (kg/m2): 23.1
non-diab	GSM631759	Non-diabetic islets, rep5	human islets, non-diabetic	pancreatic islets	non-diabetic	76 yrs	female	bmi (kg/m2): 25.9
non-diab	GSM631760	Non-diabetic islets, rep6	human islets, non-diabetic	pancreatic islets	non-diabetic	77 yrs	female	bmi (kg/m2): 23.8
non-diab	GSM631761	Non-diabetic islets, rep7	human islets, non-diabetic	pancreatic islets	non-diabetic	73 yrs	female	bmi (kg/m2): 22
diab	GSM631762	Type 2 diabetic islets, rep1	human islets, diabetic	pancreatic islets	type 2 diabetes	79 yrs	male	bmi (kg/m2): 27.5
diab	GSM631763	Type 2 diabetic islets, rep2	human islets, diabetic	pancreatic islets	type 2 diabetes	76 yrs	male	bmi (kg/m2): 26
diab	GSM631764	Type 2 diabetic islets, rep3	human islets, diabetic	pancreatic islets	type 2 diabetes	73 yrs	female	bmi (kg/m2): 29
diab	GSM631765	Type 2 diabetic islets, rep4	human islets, diabetic	pancreatic islets	type 2 diabetes	75 yrs	female	bmi (kg/m2): 26.5
diab	GSM631766	Type 2 diabetic islets, rep5	human islets, diabetic	pancreatic islets	type 2 diabetes	54 yrs	female	bmi (kg/m2): 23.9
diab	GSM631767	Type 2 diabetic islets, rep6	human islets, diabetic	pancreatic islets	type 2 diabetes	66 yrs	male	bmi (kg/m2): 23.1

Visualization



https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html


GISAID

Qui a déjà soumis à GISAID ?

C’était facile ?

Présentation de la base

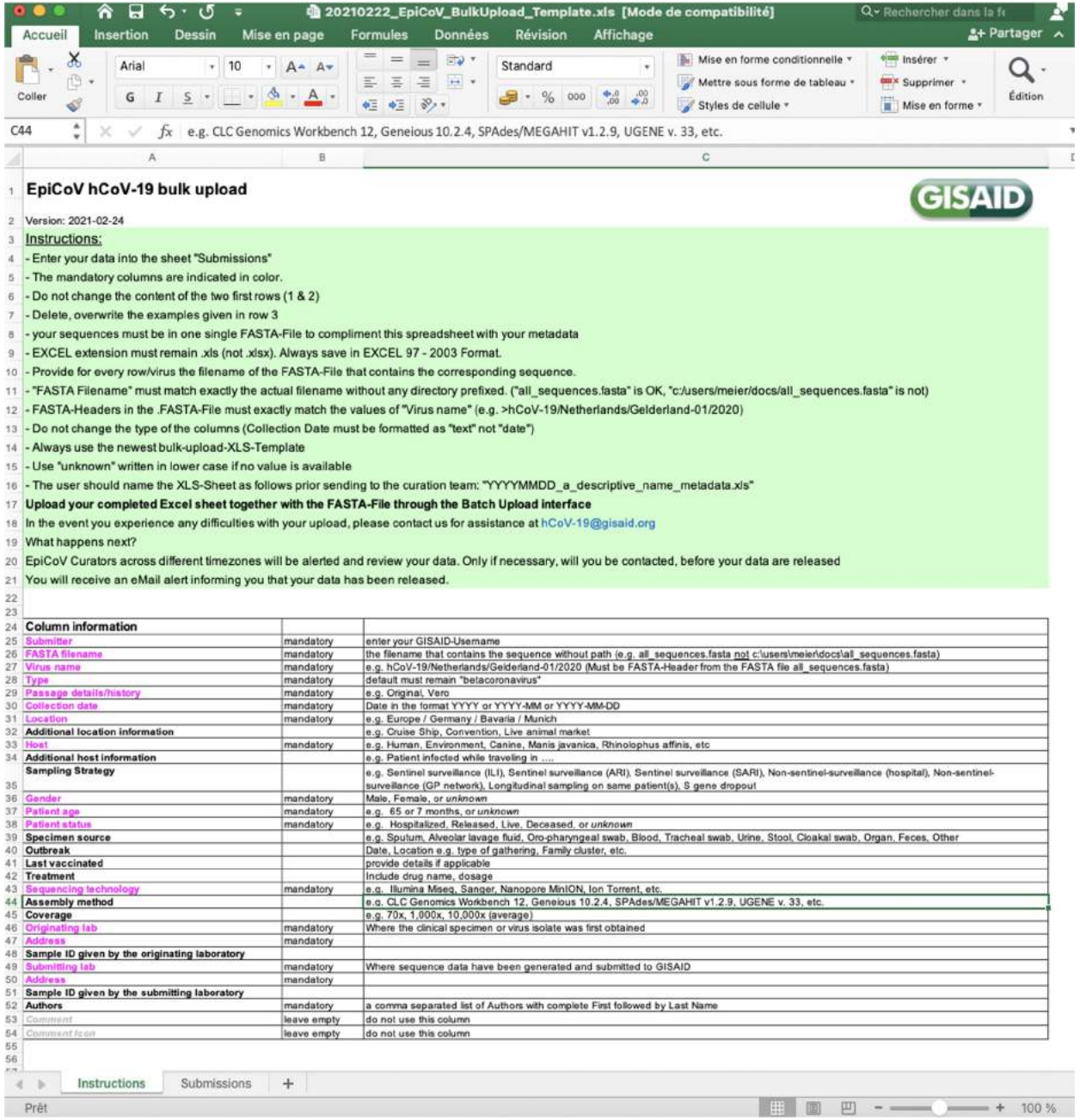
Données de tous les virus de la grippe et du **coronavirus à l'origine du COVID-19** : séquence génétique et les données cliniques et épidémiologiques associées aux virus humains, ainsi que les données géographiques et spécifiques aux espèces associées aux virus aviaires et autres virus animaux, pour aider les chercheurs à comprendre comment les virus évoluent et se propagent pendant les épidémies et les pandémies.

GISAID le fait en surmontant les obstacles et les restrictions dissuasifs, qui découragent ou empêchent le partage des données virologiques avant la publication officielle.

L'Initiative garantit que le libre accès aux données de GISAID est fourni gratuitement à toutes les personnes qui ont accepté de **s'identifier et de respecter le mécanisme de partage de GISAID régi par son accord d'accès à la base de données.**

Le fichier de métadonnées

Fichier excel



20210222_EpiCoV_BulkUpload_Template.xls [Mode de compatibilité]

Accueil | Insertion | Dessin | Mise en page | Formules | Données | Révision | Affichage | Partager

C44: e.g. CLC Genomics Workbench 12, Geneious 10.2.4, SPAdes/MEGAHIT v1.2.9, UGENE v. 33, etc.

EpiCoV hCoV-19 bulk upload

Version: 2021-02-24

Instructions:

- Enter your data into the sheet "Submissions"
- The mandatory columns are indicated in color.
- Do not change the content of the two first rows (1 & 2)
- Delete, overwrite the examples given in row 3
- your sequences must be in one single FASTA-File to compliment this spreadsheet with your metadata
- EXCEL extension must remain .xls (not .xlsx). Always save in EXCEL 97 - 2003 Format.
- Provide for every row/virus the filename of the FASTA-File that contains the corresponding sequence.
- "FASTA Filename" must match exactly the actual filename without any directory prefixed. ("all_sequences.fasta" is OK, "c:/users/meier/docs/all_sequences.fasta" is not)
- FASTA-Headers in the .FASTA-File must exactly match the values of "Virus name" (e.g. >hCoV-19/Netherlands/Gelderland-01/2020)
- Do not change the type of the columns (Collection Date must be formatted as "text" not "date")
- Always use the newest bulk-upload-XLS-Template
- Use "unknown" written in lower case if no value is available
- The user should name the XLS-Sheet as follows prior sending to the curation team: "YYYYMMDD_a_descriptive_name_metadata.xls"

Upload your completed Excel sheet together with the FASTA-File through the Batch Upload interface

In the event you experience any difficulties with your upload, please contact us for assistance at hCoV-19@gisaid.org

What happens next?

EpiCoV Curators across different timezones will be alerted and review your data. Only if necessary, will you be contacted, before your data are released

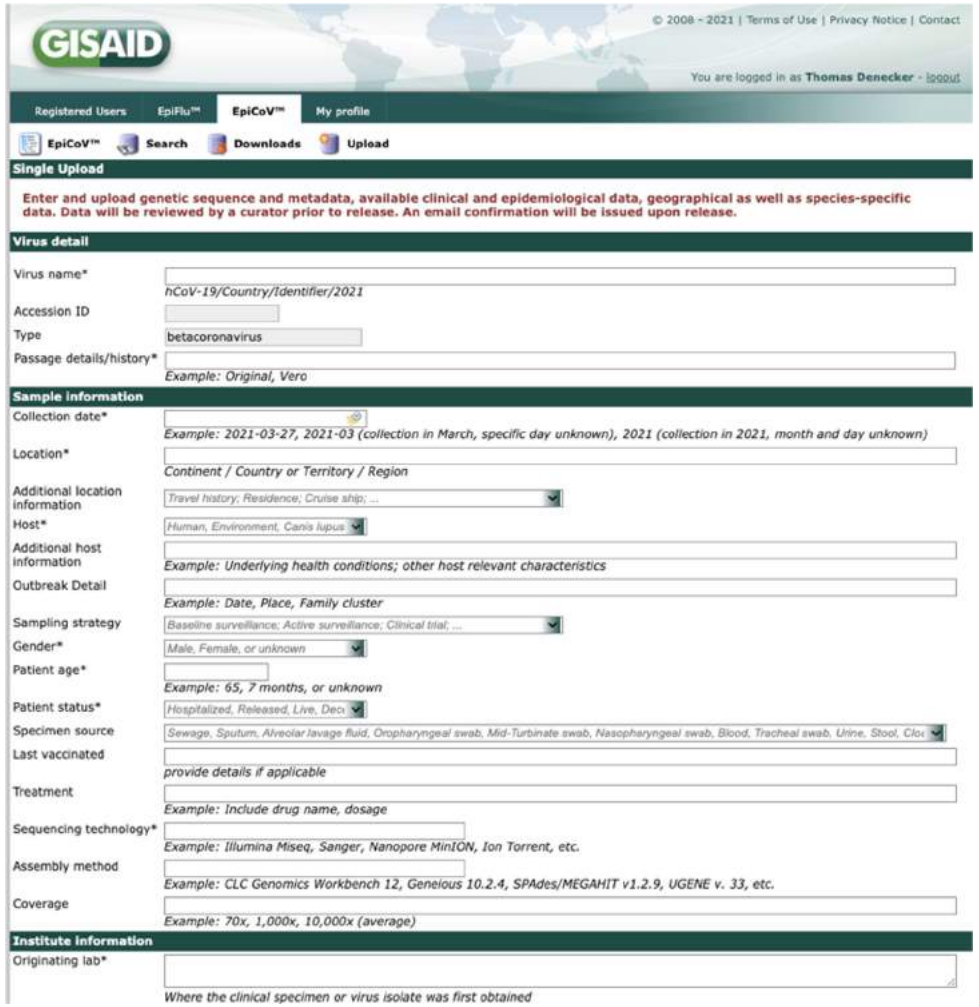
You will receive an eMail alert informing you that your data has been released.

Column information		
Submitter	mandatory	enter your GISAID-Username
FASTA filename	mandatory	the filename that contains the sequence without path (e.g. all_sequences.fasta not c:\users\meier\docs\all_sequences.fasta)
Virus name	mandatory	e.g. hCoV-19/Netherlands/Gelderland-01/2020 (Must be FASTA-Header from the FASTA file all_sequences.fasta)
Type	mandatory	default must remain "betacoronavirus"
Passage details/history	mandatory	e.g. Original, Vero
Collection date	mandatory	Date in the format YYYY or YYYY-MM or YYYY-MM-DD
Location	mandatory	e.g. Europe / Germany / Bavaria / Munich
Additional location information		e.g. Cruise Ship, Convention, Live animal market
Host	mandatory	e.g. Human, Environment, Canine, Manis javanica, Rhinolophus affinis, etc
Additional host information		e.g. Patient infected while traveling in ....
Sampling Strategy		e.g. Sentinel surveillance (ILI), Sentinel surveillance (ARI), Sentinel surveillance (SARI), Non-sentinel-surveillance (hospital), Non-sentinel-surveillance (GP network), Longitudinal sampling on same patient(s), S gene dropout
Gender	mandatory	Male, Female, or *unknown*
Patient age	mandatory	e.g. 65 or 7 months, or *unknown*
Patient status	mandatory	e.g. Hospitalized, Released, Live, Deceased, or *unknown*
Specimen source		e.g. Sputum, Alveolar lavage fluid, Oro-pharyngeal swab, Blood, Tracheal swab, Urine, Stool, Cloakal swab, Organ, Feces, Other
Outbreak		Date, Location e.g. type of gathering, Family cluster, etc.
Last vaccinated		provide details if applicable
Treatment		Include drug name, dosage
Sequencing technology	mandatory	e.g. Illumina Miseq, Sanger, Nanopore MinION, Ion Torrent, etc.
Assembly method		e.g. CLC Genomics Workbench 12, Geneious 10.2.4, SPAdes/MEGAHIT v1.2.9, UGENE v. 33, etc.
Coverage		e.g. 70x, 1,000x, 10,000x (average)
Originating lab	mandatory	Where the clinical specimen or virus isolate was first obtained
Address	mandatory	
Sample ID given by the originating laboratory		
Submitting lab	mandatory	Where sequence data have been generated and submitted to GISAID
Address	mandatory	
Sample ID given by the submitting laboratory		
Authors	mandatory	a comma separated list of Authors with complete First followed by Last Name
Comment	leave empty	do not use this column
Comment Icon	leave empty	do not use this column

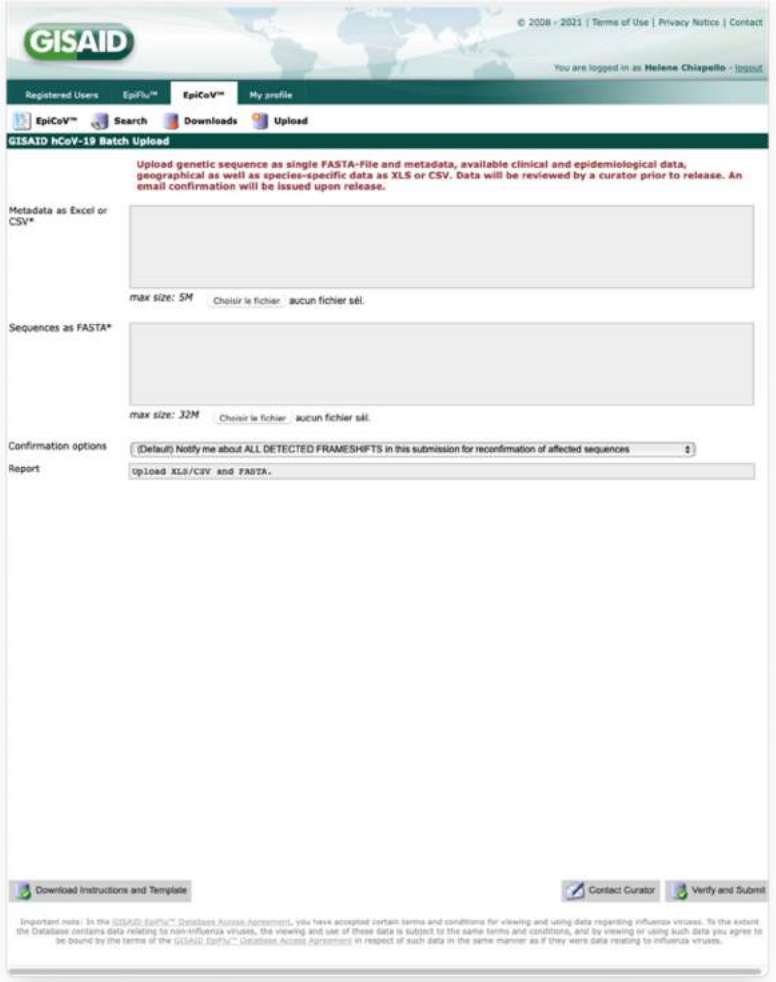
Instructions | Submissions | +

Prêt | 100 %

WEB - Single


GISAID
© 2008 - 2021 | Terms of Use | Privacy Notice | Contact
You are logged in as Thomas Denecker - logout
Registered Users
EpiFlu™
EpiCoV™
My profile
EpiCoV™
Search
Downloads
Upload
Single Upload
Enter and upload genetic sequence and metadata, available clinical and epidemiological data, geographical as well as species-specific data. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.
Virus detail
Virus name*
hCoV-19/Country/Identifier/2021
Accession ID
Type
betacoronavirus
Passage details/history*
Example: Original, Vero
Sample information
Collection date*
Example: 2021-03-27, 2021-03 (collection in March, specific day unknown), 2021 (collection in 2021, month and day unknown)
Location*
Continent / Country or Territory / Region
Additional location information
Travel history; Residence; Cruise ship; ...
Host*
Human, Environment, Canis lupus
Additional host information
Example: Underlying health conditions; other host relevant characteristics
Outbreak Detail
Example: Date, Place, Family cluster
Sampling strategy
Baseline surveillance; Active surveillance; Clinical trial; ...
Gender*
Male, Female, or unknown
Patient age*
Example: 65, 7 months, or unknown
Patient status*
Hospitalized, Released, Live, Dec
Specimen source
Sewage, Sputum, Alveolar lavage fluid, Oropharyngeal swab, Mid-Turbinate swab, Nasopharyngeal swab, Blood, Tracheal swab, Urine, Stool, Cloa
Last vaccinated
provide details if applicable
Treatment
Example: Include drug name, dosage
Sequencing technology*
Example: Illumina Miseq, Sanger, Nanopore MinION, Ion Torrent, etc.
Assembly method
Example: CLC Genomics Workbench 12, Geneious 10.2.4, SPAdes/MEGAHIT v1.2.9, UGENE v. 33, etc.
Coverage
Example: 70x, 1,000x, 10,000x (average)
Institute information
Originating lab*
Where the clinical specimen or virus isolate was first obtained

Web - Batch upload

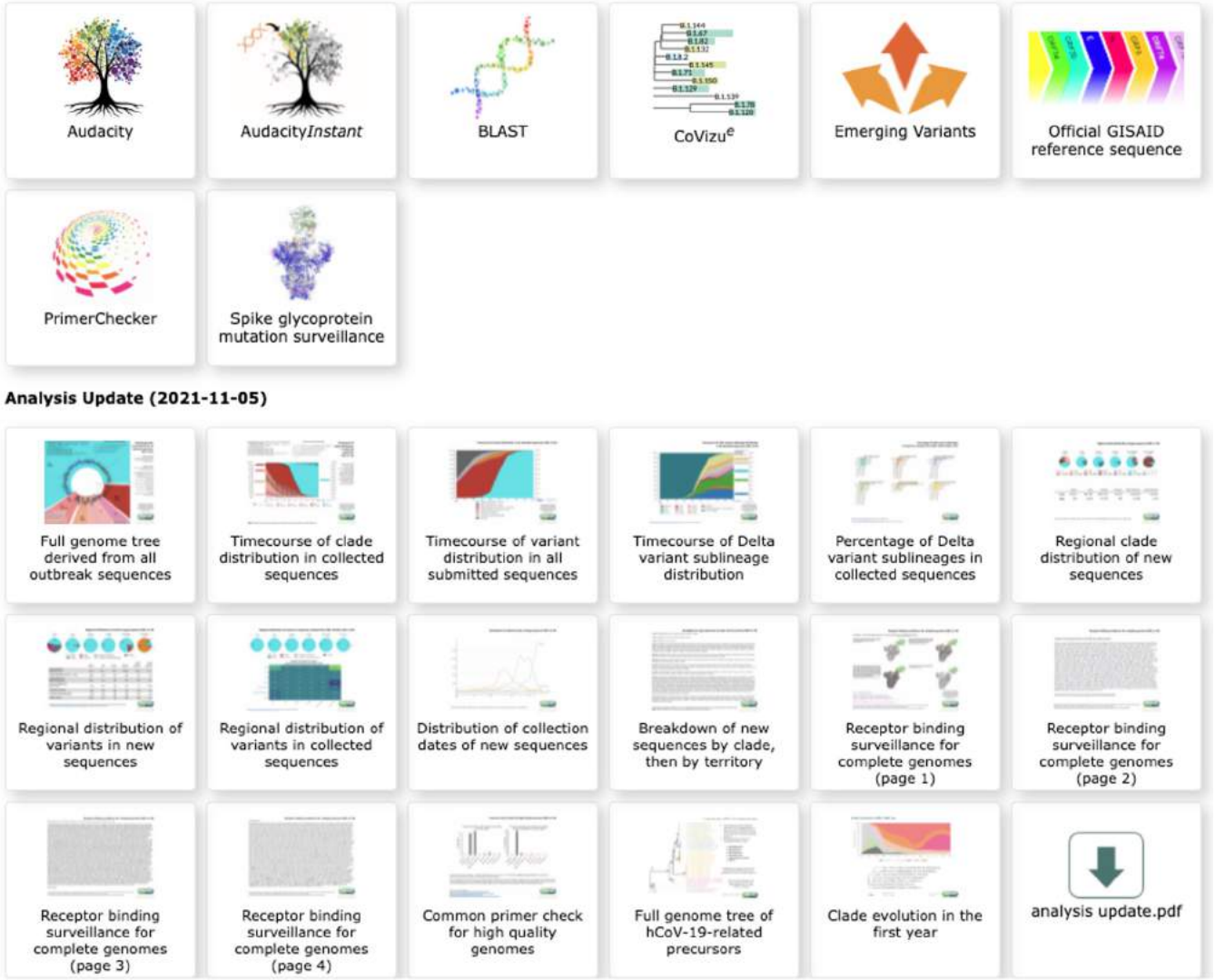


GISAID CLI2

Version 2 Command Line Interface (CLI) for batch uploading

```
usage: cli2 upload [-h] [--database {EpiCoV,EpiFlu,EpiRSV}] [--token TOKEN] --metadata METADATA --fasta FASTA
                   [--frameshift {catch_all,catch_novel,catch_none}] [--failed FAILED] [--proxy PROXY] [--debug] [--log LOG]

Perform upload of sequences and metadata to GISAID's curation zone.

optional arguments:
-h, --help            show this help message and exit
--database {EpiCoV,EpiFlu,EpiRSV}
                      Target GISAID database. (default: EpiCoV)
--token TOKEN         Authentication token. (default: ./gisaid.authtoken)
--metadata METADATA   The csv-formatted metadata file. (default: None)
--fasta FASTA         The fasta-formatted nucleotide sequences file. (default: None)
--frameshift {catch_all,catch_novel,catch_none}
                        'catch_none': catch none of the frameshifts and release immediately; 'catch_all': catch all frameshifts and require email
                        confirmation; 'catch_novel': catch novel frameshifts and require email confirmation. (default: catch_all)
--failed FAILED       Name of CSV output to contain failed records. (default: ./failed.out)
--proxy PROXY         Proxy-configuration for HTTPS-Request in the form: http(s)://username:password@proxy:port. (default: None)
--debug               Switch off debugging information (dev purposes only). (default: True)
--log LOG             All output logged here. (default: ./upload.log)
```

Les outils complémentaires



Data brokering à l’IFB

Pourquoi le développer à l'IFB

Constat

- Les soumissions sont souvent complexes et difficiles à réaliser par les équipes expérimentales.
- Les métadonnées sont souvent mal comprises, ce qui entraîne des soumissions incomplètes, redondantes et incohérentes.

L'ENA a demandé à l'IFB de devenir le data broker français

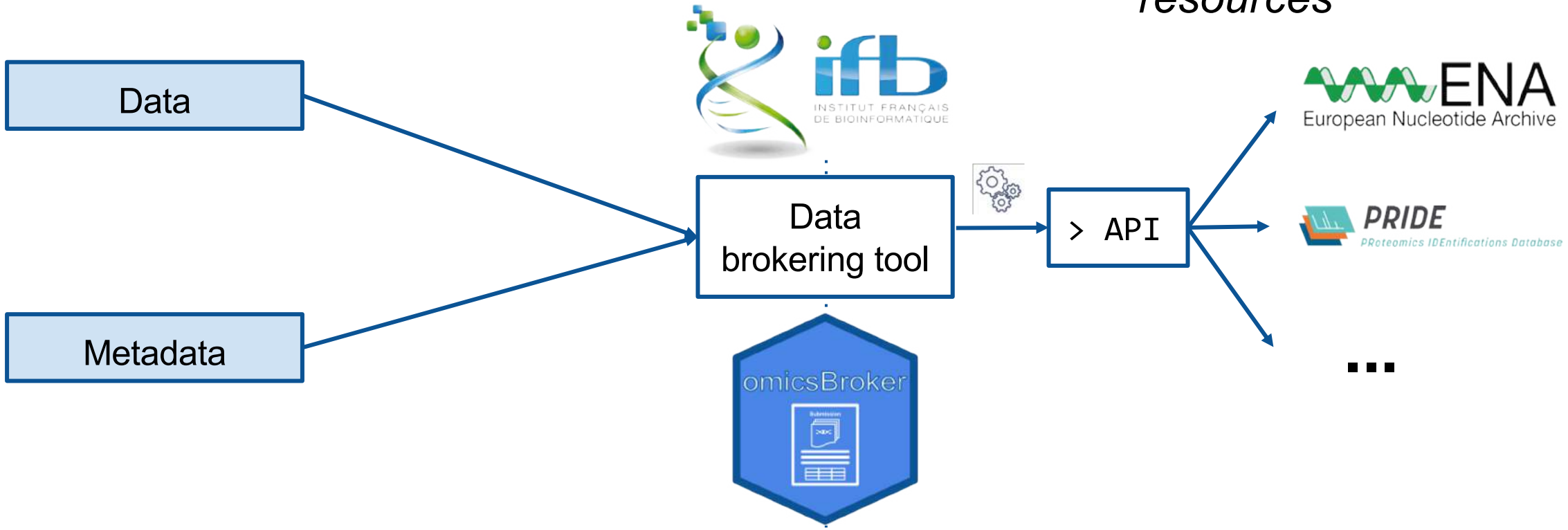
Idée principale : offrir un service national de data brokering à IFB pour **simplifier** et **rationaliser** les échanges de données entre les ressources internationales et le nœud Elixir français IFB.

3 types d'activités : le développement d'outils, la formation et le support aux utilisateurs.

Data Brokering service developed by IFB

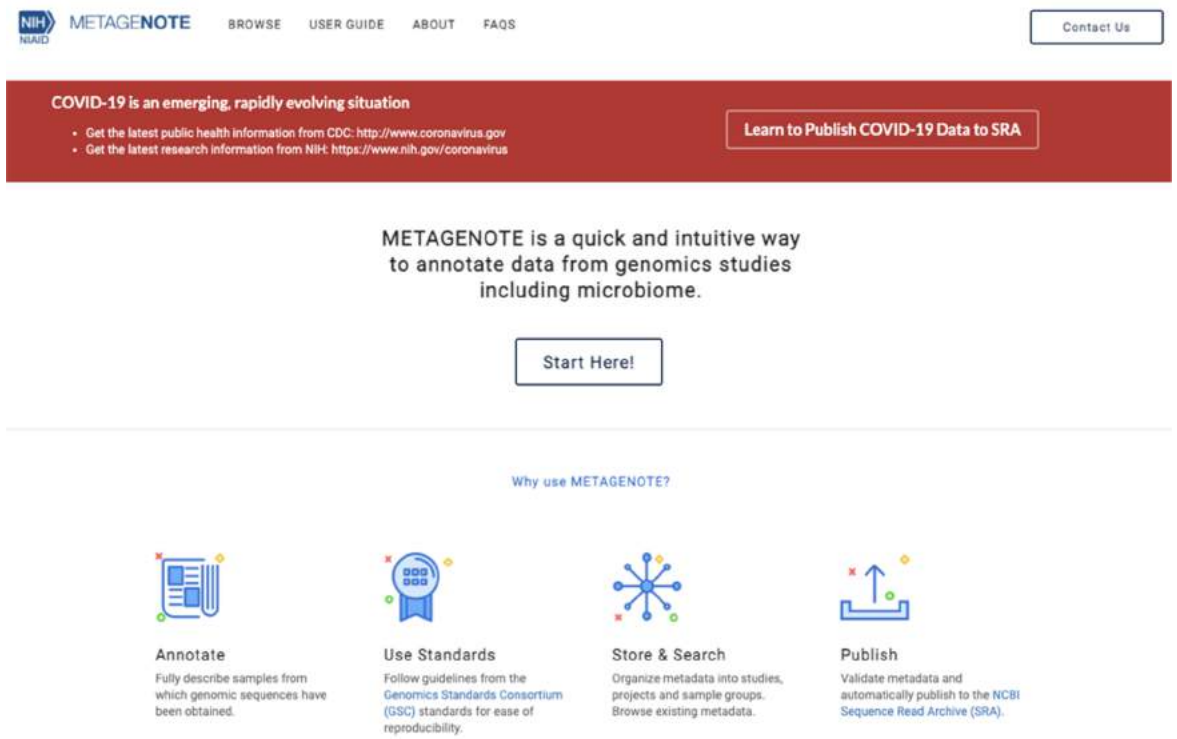
IFB services to manage and centralize data and metadata of a project

IFB services to submit data and metadata of a project to international resources



Des outils de data brokering déjà disponibles

https://www.gfbio.org/



https://metagenote.niaid.nih.gov/

Publication d’articles/ OpenLink

Les entrepôts de données

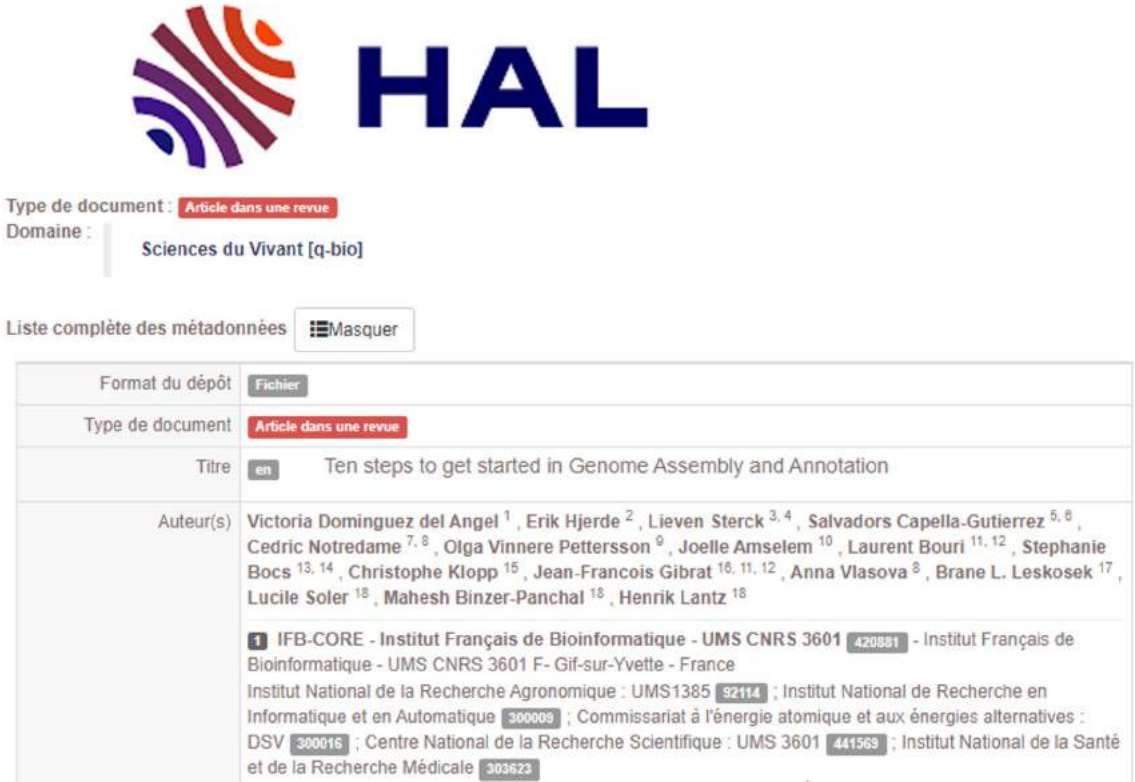
open access

- Applications Web open source pour partager, conserver, citer, explorer et analyser les données de recherche.
- Facilite la mise à disposition des données aux autres et vous permet de reproduire leurs travaux.

The Dataverse® Project

zenodo

HAL


Type de document : Article dans une revue
Domaine : Sciences du Vivant [q-bio]

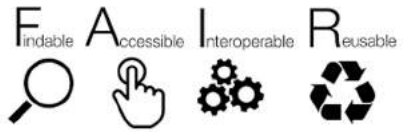
Liste complète des métadonnées Masquer

Format du dépôt	Fichier
Type de document	Article dans une revue
Titre	en Ten steps to get started in Genome Assembly and Annotation
Auteur(s)	Victoria Dominguez del Angel [1], Erik Hjerde [2], Lieven Sterck [3, 4], Salvadors Capella-Gutierrez [5, 6], Cedric Notredame [7, 8], Olga Vinnere Pettersson [9], Joelle Amselem [10], Laurent Bouri [11, 12], Stephanie Bocs [13, 14], Christophe Klopp [15], Jean-Francois Gibrat [16, 11, 12], Anna Vlasova [8], Brane L. Leskosek [17], Lucile Soler [18], Mahesh Binzer-Panchal [18], Henrik Lantz [18] 1 IFB-CORE - Institut Français de Bioinformatique - UMS CNRS 3601 420881 - Institut Français de Bioinformatique - UMS CNRS 3601 F- Gif-sur-Yvette - France Institut National de la Recherche Agronomique : UMS1385 92114 ; Institut National de Recherche en Informatique et en Automatique 300009 ; Commissariat à l'énergie atomique et aux énergies alternatives : DSV 300016 ; Centre National de la Recherche Scientifique : UMS 3601 441569 ; Institut National de la Santé et de la Recherche Médicale 303623

Organisation des données avec OpenLink

Buts

Une **vision claire des données** associées à chaque projet de recherche



Réduire les obstacles à l'adoption des principes FAIR

Limiter l'impact de FAIR sur le temps de gestion des données

Assister les chercheurs dans la **publication de leurs données**

Organisation des données avec OpenLink

Solution

Source code: https://gitlab.com/igbmc/openlink
Documentation: https://openlink.readthedocs.io/en/latest

Une **application web open-source** basée sur le framework Django (langage Python)

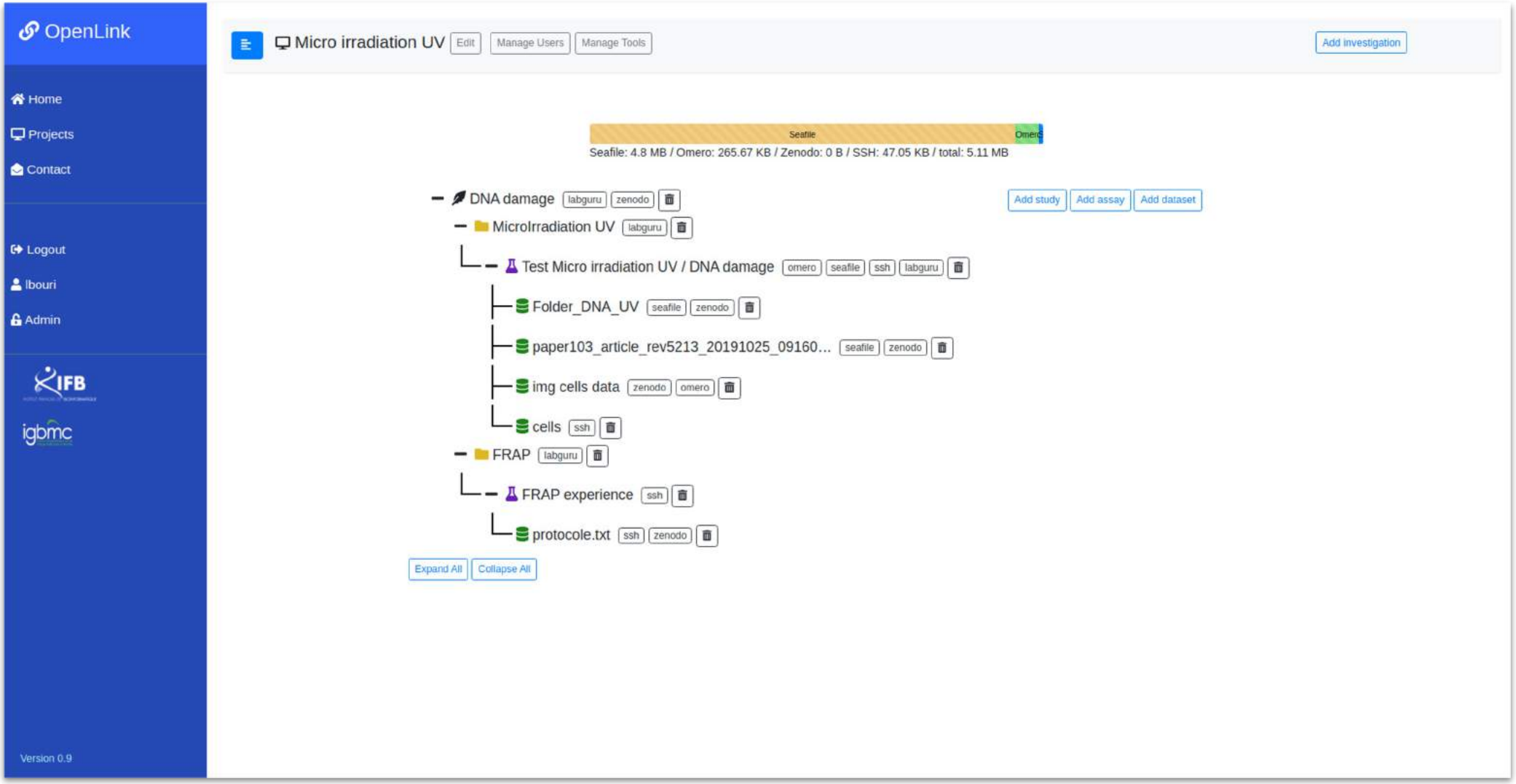
Une base de données pour créer des **liens** entre la structure d'un projet de recherche (modèle ISA) et de multiples sources de données

Une **collection évolutive de connecteurs** aux outils couramment utilisés par les chercheurs: LabGuru, Omero, Seafile, mass storage, etc.

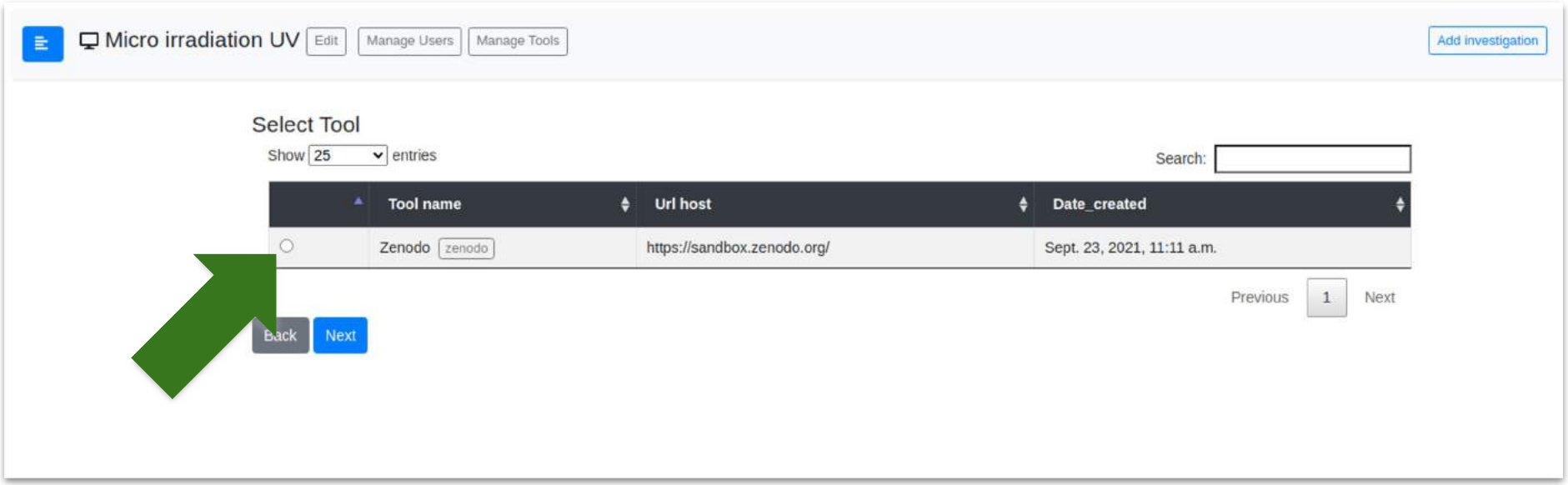
Des **outils intégrés** pour faciliter la manipulation des données

Openlink: De la recherche à la publication de données

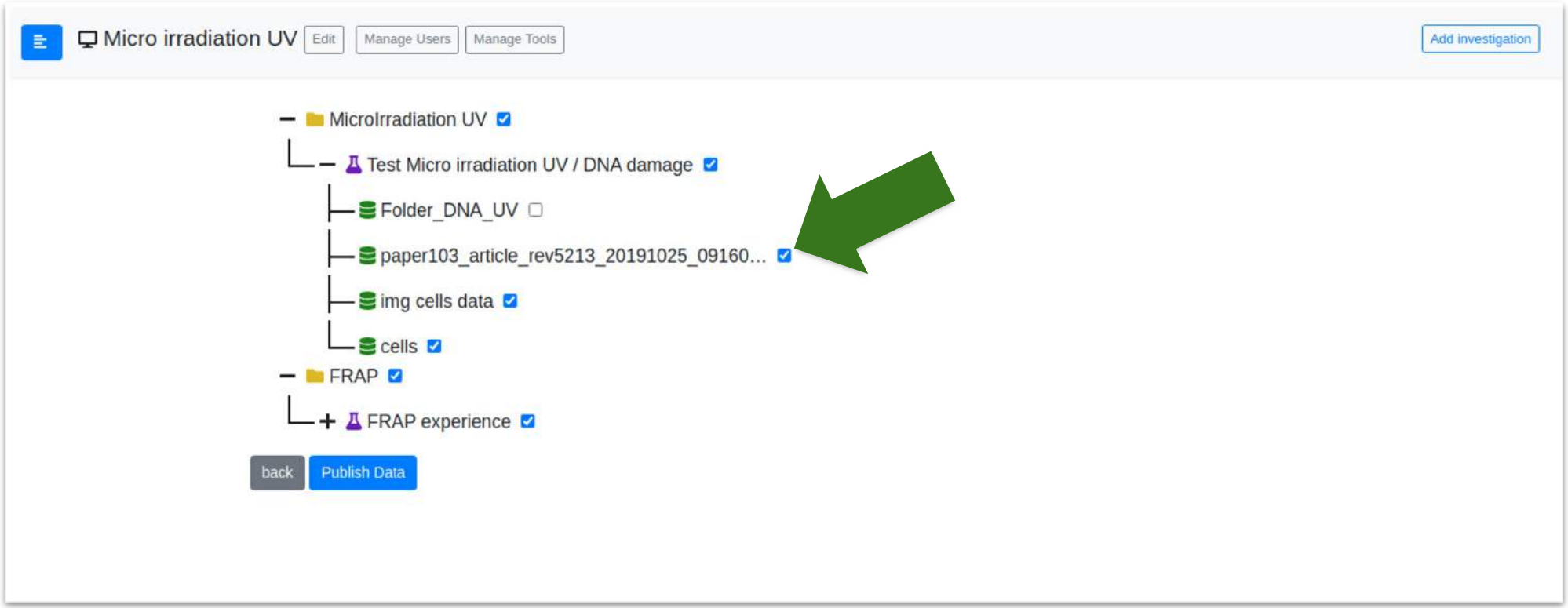
Projet complet avec plus de données liées



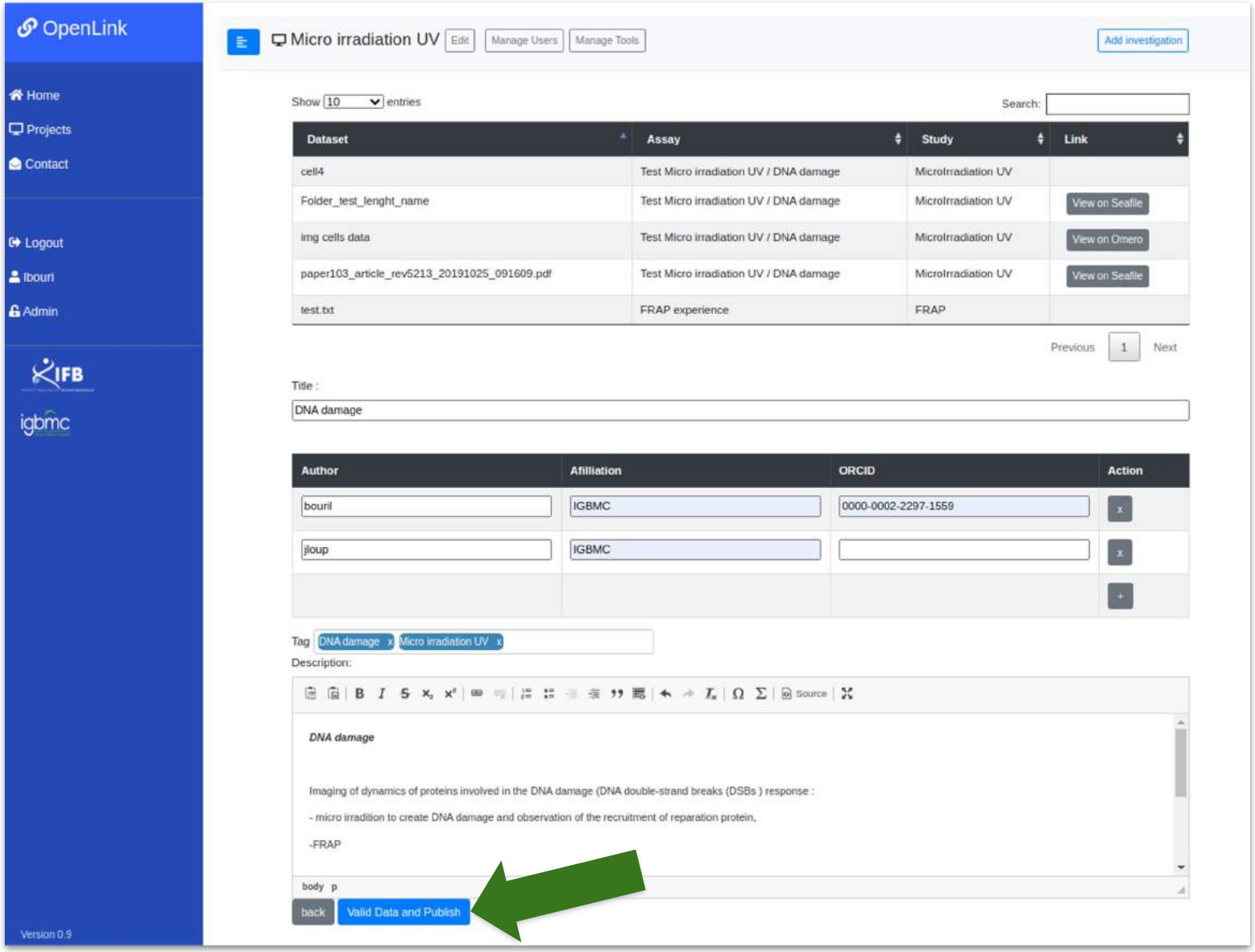
Publication sur Zenodo



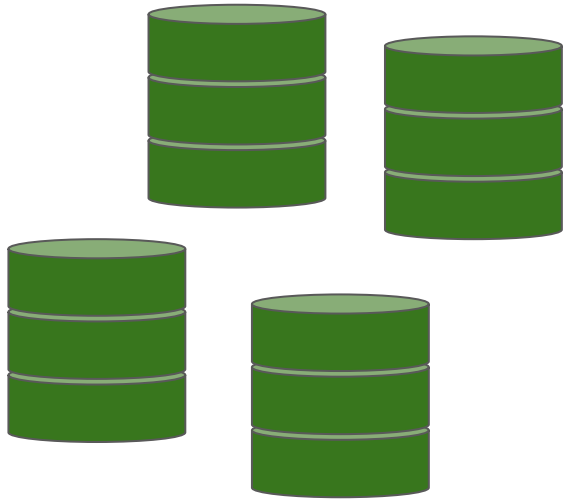
Publication sur Zenodo



Publication sur Zenodo



Publication sur Zenodo

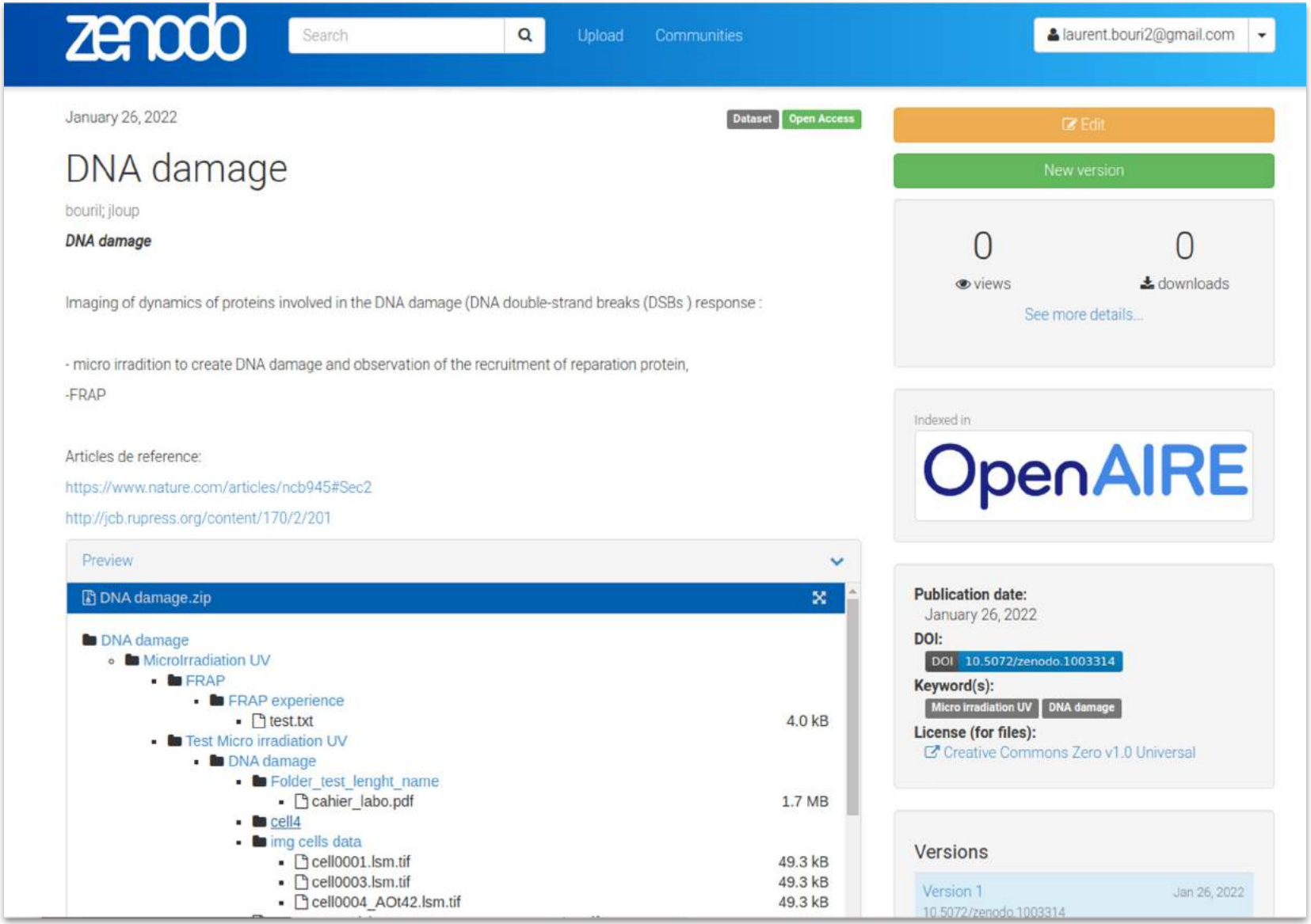


Download all datasets

Create archive

Upload to Zenodo
with metadata

Publication sur Zenodo



Supplementary data

The omicsBroker tool

omicsBroker is a tool to easily annotate and submit **omics data** to **international repositories**

Prototype disponible (soumission dans la zone de test de l'ENA)

- Développé en Django
- Disponible en Docker

Futurs développements

- Gestionnaire de soumission,
- API,
- **...**

Exemple du prototype

Metadata table

Excel

	Experience name	Organism	Platform	Instrument	Library layout	Insert size
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

Descriptions

Search

Platform

Definition

Platform name. Permitted values : https://ena-docs.readthedocs.io/en/latest/submit/reads/webin-cli.html#permitted-values-for-platform

Value

LS454 ; ILLUMINA ; PACBIO_SMRT ; ION_TORRENT ; CAPILLARY ; OXFORD_NANOPORE ; DNBSEQ

Harmonized Name

PLATFORM

* Mandatory