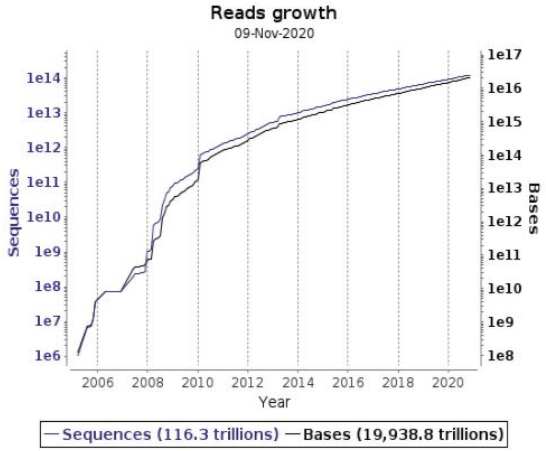
Data brokering

frama.link/ifb-ag20-data-brokering

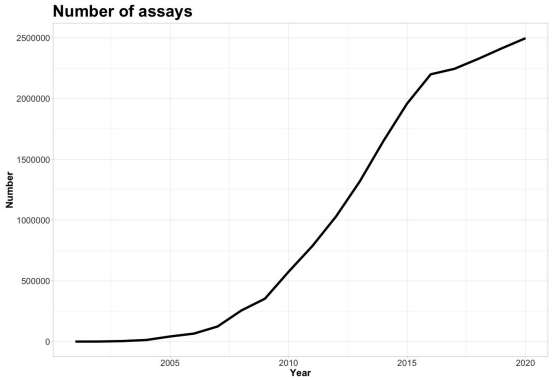
Thomas Denecker, Hélène Chiapello & Jacques van Helden

Contexte

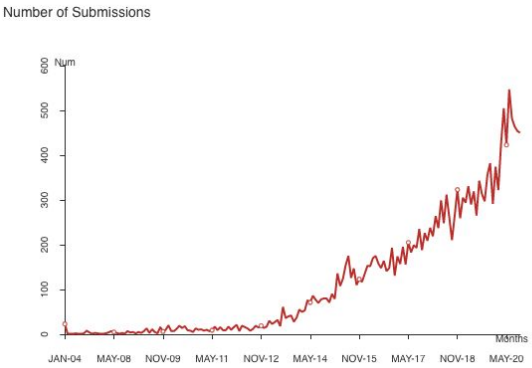
Toujours plus de données...



ENA
European Nucleotide Archive

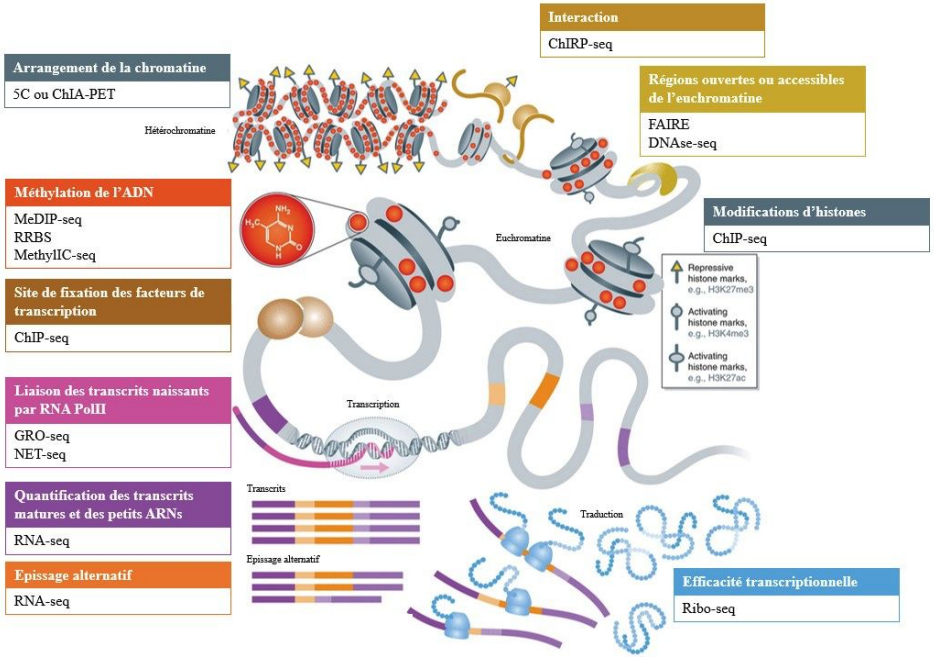


Array express



PRIDE
PRoteomics IDEntifications Database

… de plus en plus hétérogènes ...



(Soon et al. 2013)

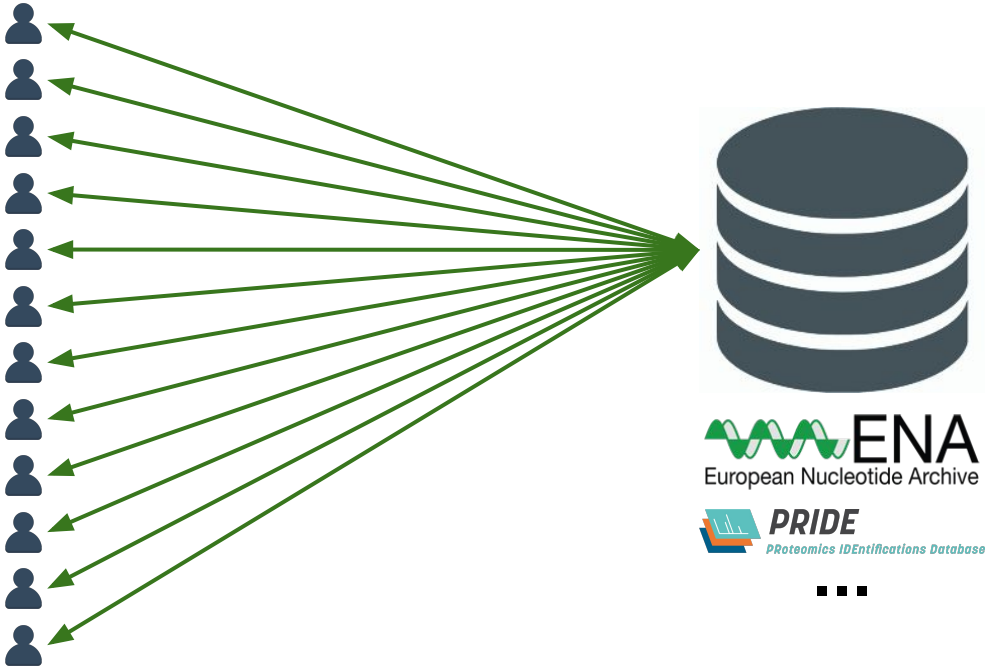
… stockées dans des bases de données de plus en plus nombreuses

≈ **1700** bases de données référencées dans NAR Database en 2018

(Rigden et al, 2018)

Soumission

Soumission

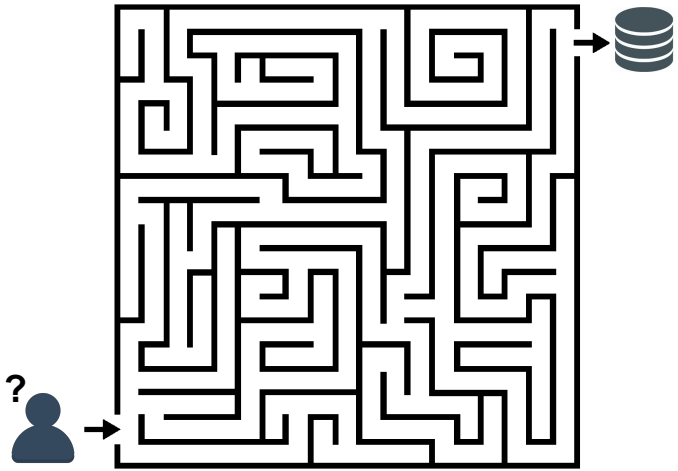


Aujourd'hui

Soumission par l'équipe expérimentale ou bioinformatique

Allers-retours entre la base de données et l'utilisateur

Soumission du point de vue de l'utilisateur



Le sentiment général ?

"*Pas toujours si simple*"

Des questions récurrentes

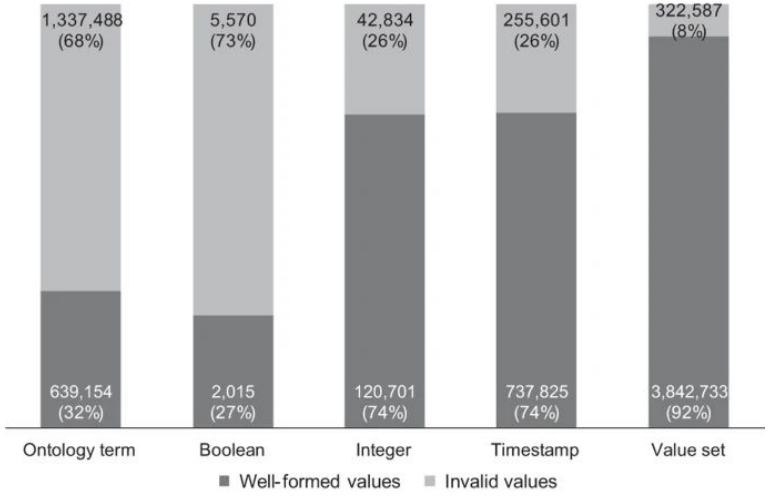
"*Dans quelle base de données soumettre ? Quelles données soumettre ? Les brutes, les traitées ? Accompagnées de quelles informations supplémentaires ? …*"

Bilan après soumission

Manque de consistance des soumissions

Métadonnées de mauvaise qualité

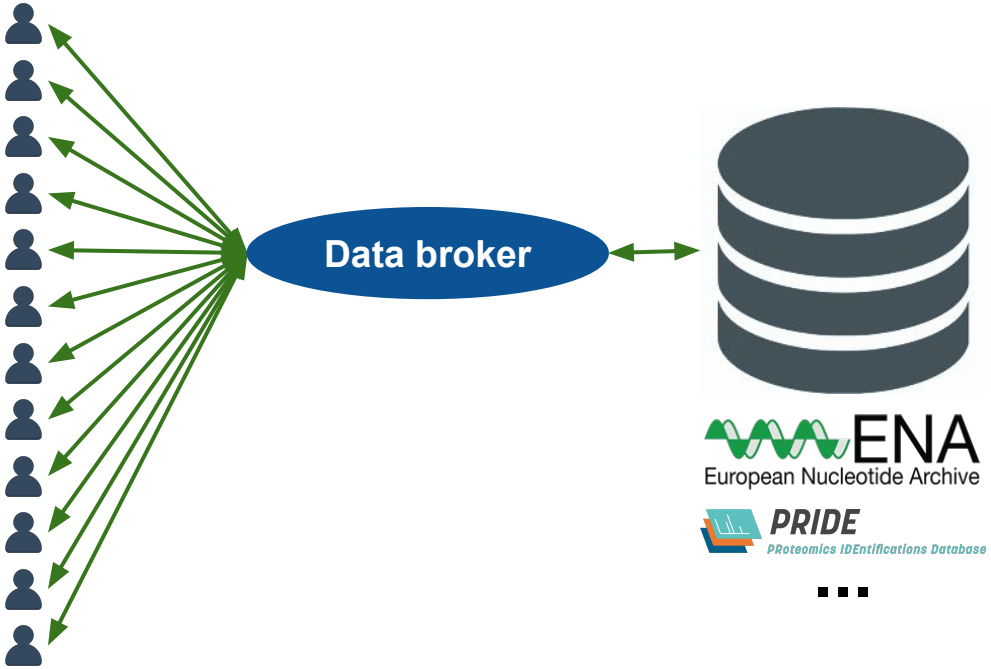
- Incohérentes
- Incomplètes
- Peu informatives
- Redondantes



(Gonçalves et al., 2019)

Data brokering

Un *data broker*



Intermédiaire entre le producteur de données et la ressource internationale de stockage/archivage

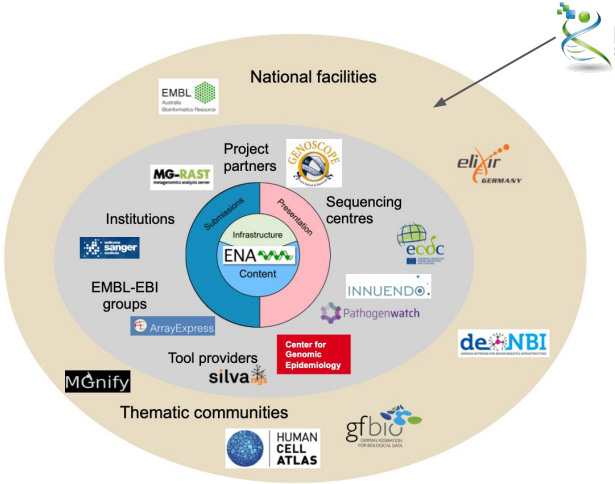
Rationalisation des échanges

Data broker & Soumission

Missions

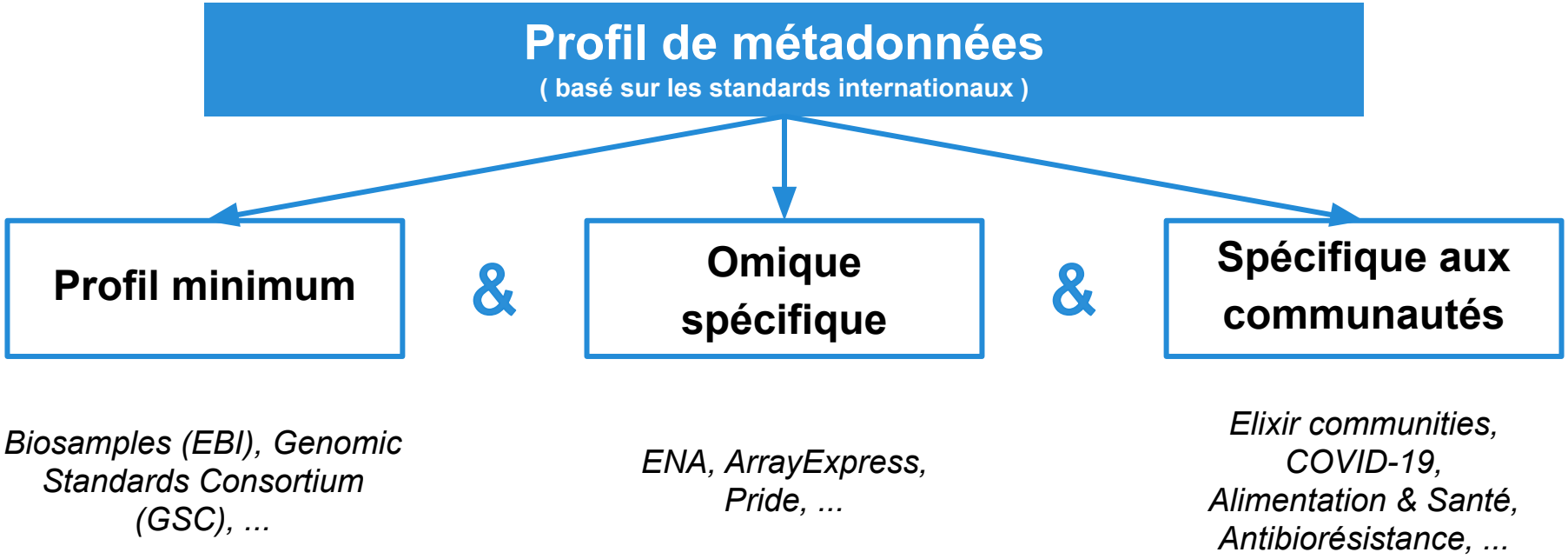
1. Identifier et collecter les informations obligatoires pour décrire et soumettre les données
2. Centraliser ces informations pour les soumettre dans une base de données adéquate
3. Implémenter des outils automatisés permettant la fluidification du flux de données

Qui ?



(Cochrane, 2018)

1. Identifier et collecter les informations obligatoires pour décrire et soumettre les données



Les métadonnées

Définition

"Metadata are the in-depth, **controlled description** of the sample that your sequence was taken from. Essentially, the '**what, where, how, and when**' of your study from collection to sequence generation, plus contextual data such as environmental conditions (latitude, longitude, temperature) or clinical observations."

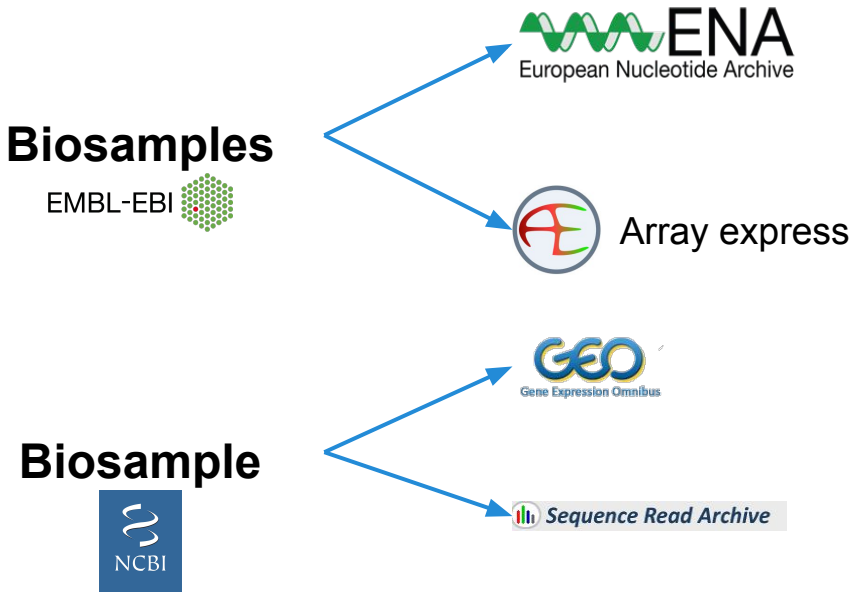
EMBL-EBI

genomic STANDARDS consortium

GGBN

...

Des bases de données de métadonnées



2- Centraliser ces informations pour les soumettre dans une base de données adéquate

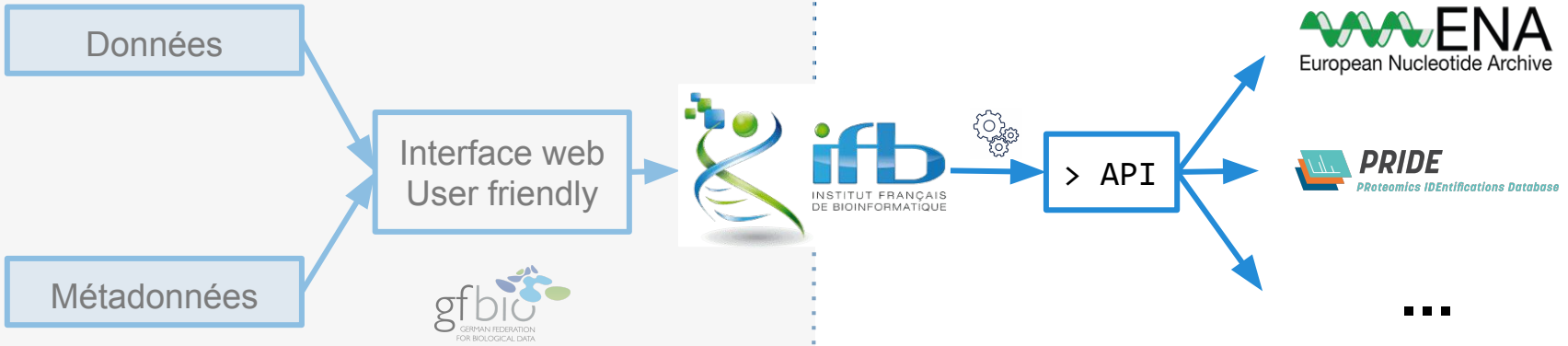
3- Implémenter des outils automatisés permettant la fluidification du flux de données

2- Centraliser ces informations pour les soumettre dans une base de données adéquate

3- Implémenter des outils automatisés permettant la fluidification du flux de données

2- Centraliser ces informations pour les soumettre dans une base de données adéquate

3- Implémenter des outils automatisés permettant la fluidification du flux de données



Exemples de brokers

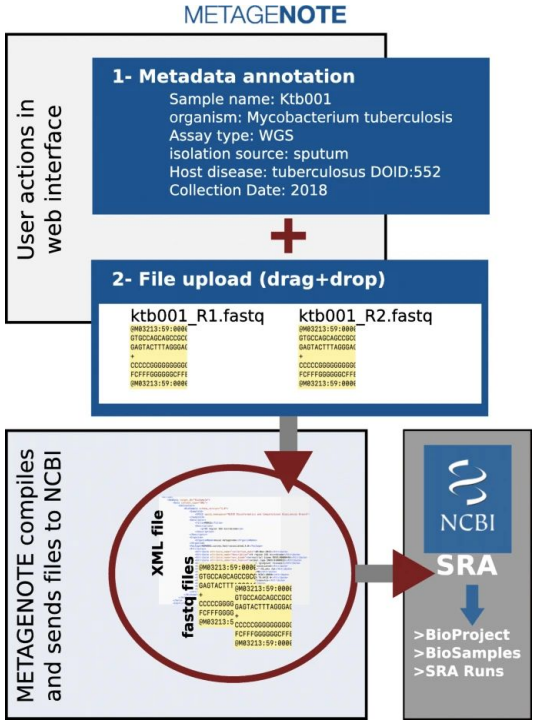
METAGENOTE

Basé sur les *guidelines* du Genomic Standards Consortium (GSC) et des ontologies (ENVO, FMA, …)

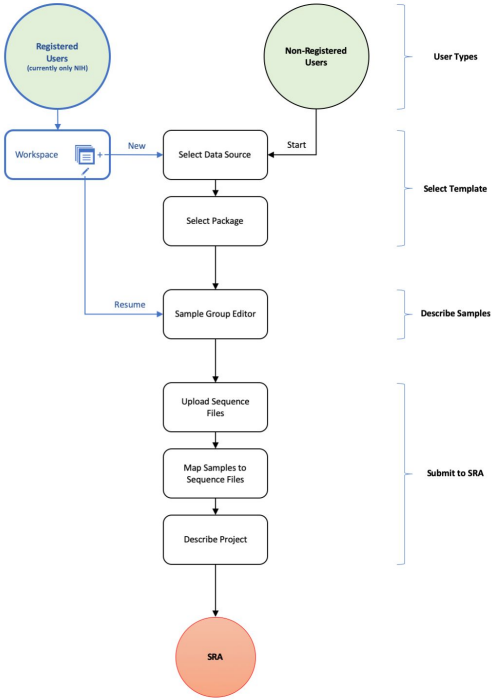
Interface web + MariaDb

Soumission automatique sur SRA via API

(Quiñones et al. - 2020 - https://doi.org/10.1186/s12859-020-03694-0)

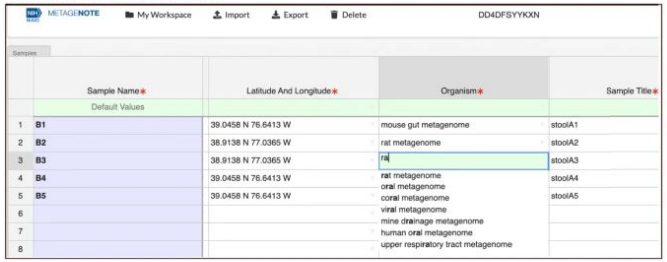


METAGENOTE

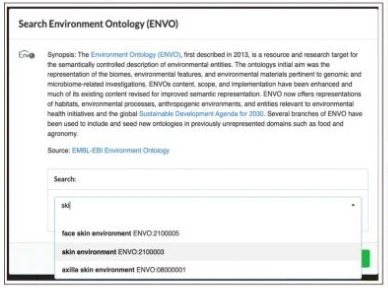


Via une interface web

2a: Sample group annotation table

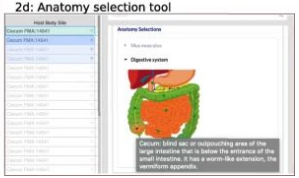


2b: Ontology search box (right+click menu option)



2c: Right pane description search box

2d: Anatomy selection tool



Preuve de concept

COVID-19

Constat

SCIENCES · MEDECINE

Partage

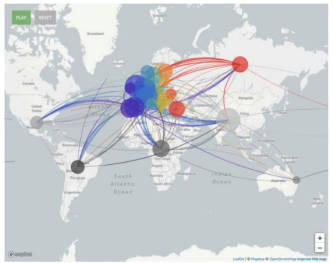
Covid-19 : les chercheurs français peu partageurs des séquences génétiques

La mise en commun massive permet une étude plus précise du virus et de son évolution, mais les scientifiques français y sont réticents.

Par David Larousserie · Publié le 31 août 2020 à 06h30 - Mis à jour le 01 septembre 2020 à 12h47

Lecture 7 min.

Article réservé aux abonnés



Représentation des origines des diverses importations du coronavirus en Europe entre le 7 avril et le 1er juillet 2020, tirée du séquençage de leurs génomes. Image extraite du site Nextstrain

Il n'y a de pire aveugle que celui qui ne veut pas voir. En matière de Covid-19, le dicton s'appliquerait-il à la France ? Notre pays semble en effet peu enclin à utiliser un outil de pointe qui permettrait de répondre à des questions importantes sur l'épidémie, comme

(Larousserie, Le Monde, 31 août 2020)

De nouveaux outils pour faciliter la soumission

New tool simplifies the submission of SARS-CoV-2 data to open databases

ELIXIR Belgium and ELIXIR Germany (de.NBI) help researchers share FAIR COVID-19 data

ELIXIR Belgium, in collaboration with ELIXIR Germany and the European COVID-19 Data Platform, have developed a tool to simplify the submission of viral sequencing data to the European Nucleotide Archive (ENA), an ELIXIR Core Data Resource providing open access to nucleotide sequences. The new submission tool offers an easy-to-use interface, guides researchers through the submission process and verifies the data format and description.

Why submit data to ENA?

(ELIXIR, 17 November 2020)

Conclusion & Perspectives

Les actions à venir

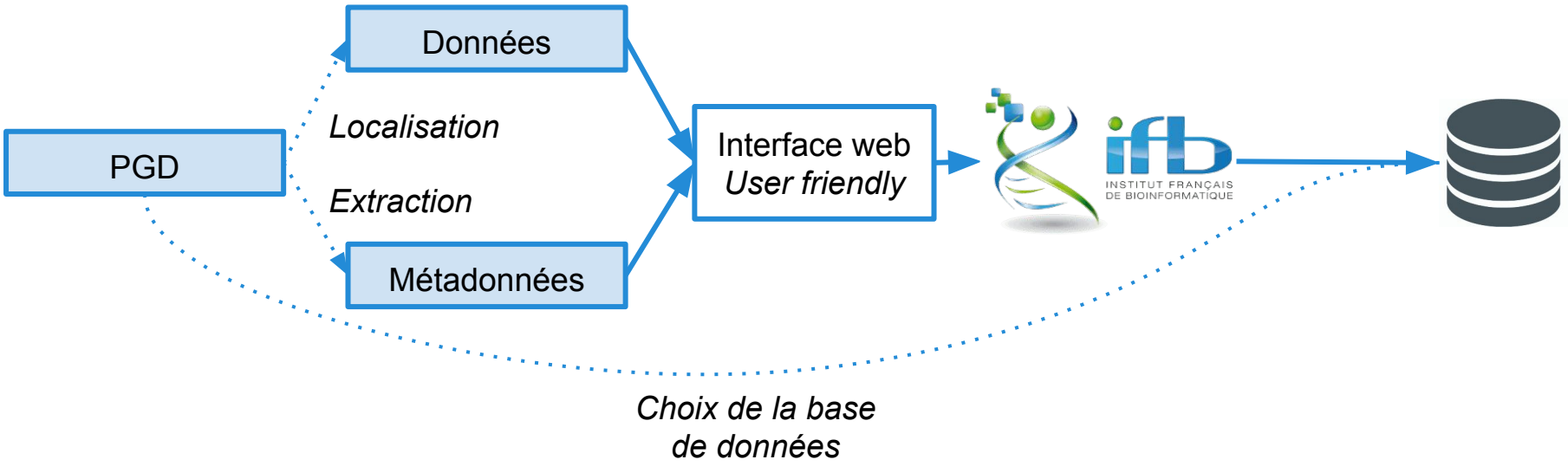
Entre l'utilisateur et le *data broker*

1) Établir et collecter un ensemble d'informations nécessaires lors d'une soumission et proposer un **profil de métadonnées**
2) Mettre en place un **système *User friendly* de collecte** des métadonnées et des données

Entre le *data broker* et la base de données

3) Mettre en place des ***workflows* automatiques** de soumission adaptés au type de données

Une connexion avec le plan de gestion de données



Mise à jour de la soumission lors de la mise à jour du PGD (intégration continue)

Annexes

Un point sur les métadonnées

Quelques définitions

"Metadata refers to **descriptive information** about the overall study, individual samples, all protocols, and references to processed and raw data file names."

GEO

"Metadata characterize biological resources by **core information** including a name, a description of its input and its output (parameters or format), its address, and various additional properties."

Encyclopedia of Database Systems, 2009

"Metadata are the in-depth, **controlled description** of the sample that your sequence was taken from. Essentially, the 'what, where, how, and when' of your study from collection to sequence generation, plus contextual data such as environmental conditions (latitude, longitude, temperature) or clinical observations."

EMBL-EBI

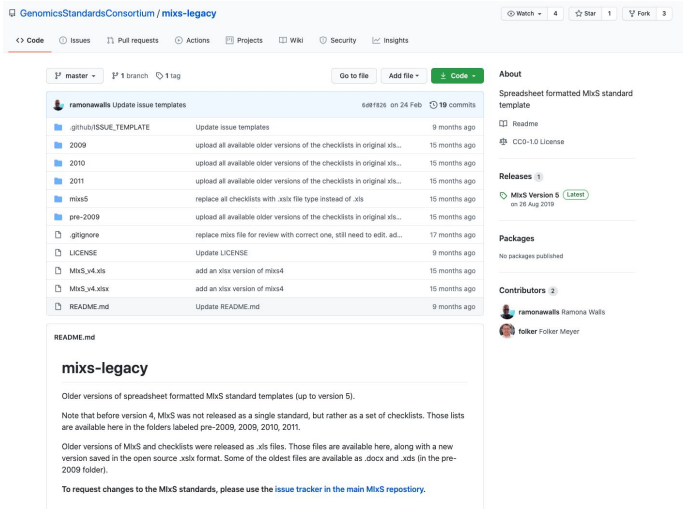
Genomic Standards Consortium (GSC)

Checklist
(obligatoire)

&

Environmental packages
(informations complémentaires)

Fichier excel téléchargeable sur GitHub
(dernière MAJ 26 Aug 2019)



Genomic Standards Consortium (GSC)

Checklist *- Informations minimales pour les séquences (MIxS)*

94 items (± obligatoires) décrits dans 22 colonnes :

The MIxS checklist
Column 1 - structured comment name: name of a checklist item as it will appear in GenBank structured comments
Column 2 - item: full name of item as it appears in the publication
Column 3 - definition: a description of the item, including links to ontologies and other resources that can be used to fill in values for the item
Column 4 - expected value: short description and/or expected value of an item
Column 5 - value syntax: the proper syntax for writing the value for a given item
Column 6 - example: examples of values for an item
Column 7 - section: the section of an item
Columns 8 through 18-migs_eu,migs_ba,migs_pl,migs_vi,migs_org,me,mimarks_s,mimarks_c,misag,mimag,miuvig: information about whether an item is mandatory (M), conditional mandatory (C), optional (X), environment-dependent (E) or not applicable (-) for a given checklist type
Column 19 - preferred units: a unit suggestion if a measurement value is given
Column 20 - occurrence: indicates whether a given item can be used only once (1), multiple times (m), or none (0)
Column 21 - position: position of item as it appears in the publication
Column 22 - MIXS ID: a unique of an item

Genomic Standards Consortium (GSC)

genomic STANDARDS consortium

Checklist *(obligatoire)*

Informations minimales pour les séquences (MIxS) - 94 items (± obligatoires)

Exemple : la méthode de séquençage (obligatoire)

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>seq_meth</td><td>sequencing method</td><td>Sequencing method used; e.g. Sanger, pyrosequencing, ABI-solid</td><td>enumeration</td><td>[MinION|GridION|PromethION|454 GS|454 GS 20|454 GS FLX|454 GS FLX+|454 GS FLX Titanium|454 GS Junior|Illumina Genome Analyzer|Illumina Genome Analyzer II|Illumina Genome Analyzer IIx|Illumina HiSeq 4000|Illumina HiSeq 3000|Illumina HiSeq 2500|Illumina HiSeq 2000|Illumina HiSeq 1500|Illumina HiSeq 1000|Illumina HiScanSQ|Illumina MiSeq|Illumina HiSeq X Five|Illumina HiSeq X Ten|Illumina NextSeq 500|Illumina NextSeq 550|AB SOLiD System|AB SOLiD System 2.0|AB SOLiD System 3.0|AB SOLiD 3 Plus System|AB SOLiD 4 System|AB SOLiD 4hq System|AB SOLiD PI System|AB 5500 Genetic Analyzer|AB 5500xl Genetic Analyzer|AB 5500xl-W Genetic Analysis System|Ion Torrent PGM|Ion Torrent Proton|Ion Torrent S5|Ion Torrent S5 XL|PacBio RS|PacBio RS II|Sequel|AB 3730xL Genetic Analyzer|AB 3730 Genetic Analyzer|AB 3500xL Genetic Analyzer|AB 3500 Genetic Analyzer|AB 3130xL Genetic Analyzer|AB 3130 Genetic Analyzer|AB 310 Genetic Analyzer|BGISEQ-500]</td><td>Illumina HiSeq 1500</td><td>sequencing</td><td>M</td><td>M</td><td>M</td><td>M</td><td>M</td><td>M</td><td>M</td><td>M</td><td>M</td><td>M</td><td>M</td><td></td><td>1</td><td>50</td><td>MIXS:0000050</td></tr>
</table>

Genomic Standards Consortium (GSC)

Environmental packages *- 17 différents*

Environmental package	Structured comment name	Package item	Definition	Expected value	Value syntax	Example	Requirement	Preferred unit	Occurrence	Position	MIXS ID
air	alt	altitude	Altitude is a term used to identify heights of objects such as airplanes, space shuttles, rockets, atmospheric balloons and heights of places such as atmospheric layers and clouds. It is used to measure the height of an object which is above the earth's surface. In this context, the altitude measurement is the vertical distance between the earth's surface above sea level and the sampled position in the air	measurement value	{float} {unit}	100 meter	M	meter	1	0	MIXS.000009 4
air	elev	elevation	Elevation of the sampling site is its height above a fixed reference point, most commonly the mean sea level. Elevation is mainly used when referring to points on the earth's surface, while altitude is used for points above the surface, such as an aircraft in flight or a spacecraft in orbit	measurement value	{float} {unit}	100 meter	C	meter	1	0	MIXS.000009 5
air	barometric_press	barometric pressure	Force per unit area exerted against a surface by the weight of air above that surface	measurement value	{float} {unit}	5 millibar	X	millibar	1	1	MIXS.000009 6
air	carb_dioxide	carbon dioxide	Carbon dioxide (gas) amount or concentration at the time of sampling	measurement value	{float} {unit}	410 parts per million	X	micromole per liter, parts per million	1	1	MIXS.000009 7
air	carb_monoxide	carbon monoxide	Carbon monoxide (gas) amount or concentration at the time of sampling	measurement value	{float} {unit}	0.1 parts per million	X	micromole per liter, parts per million	1	1	MIXS.000009 8
air	chem_administration	chemical administration	List of chemical compounds administered to the host or site where sampling occurred, and when (e.g. Antibiotics, n fertilizer, air filter); can include multiple compounds. For chemical entities of biological interest ontology (chebi) (v 163), http://purl.bioontology.org/ontology/chebi	CHEBI;timestamp	{termLabel} {[termID]};{timestamp}	agar [CHEBI:2509];2018-05-11T20:00Z	X		m	1	MIXS.000075 1
air	humidity	humidity	Amount of water vapour in the air, at the time of sampling	measurement value	{float} {unit}	25 gram per cubic meter	X	gram per cubic meter	1	1	MIXS.000010 0
air	methane	methane	Methane (gas) amount or concentration at the time of sampling	measurement value	{float} {unit}	1800 parts per billion	X	micromole per liter, parts per billion, parts per million	1	1	MIXS.000010 1
air	misc_param	miscellaneous parameter	Any other measurement performed or parameter collected, that is not listed here	parameter name;measurement value	{text};{float} {unit}	Bicarbonate ion concentration;2075 micromole per kilogram	X		m	1	MIXS.000075 2
air	organism_count	organism count	Total cell count of any organism (or group of organisms) per gram, volume or area of sample, should include name of organism followed by count. The method that was used for the enumeration (e.g. qPCR, atp, mpn, etc.) Should also be provided. (example: total prokaryotes; 3.5e7 cells per ml; qpcr)	organism name;measurement value;enumeration	{text};{float} {unit};[qPCR|ATP|MPN|other]	total prokaryotes;3.5e7 cells per milliliter;qPCR	X	number of cells per cubic meter, number of cells per milliliter, number of cells per cubic centimeter	m	1	MIXS.000010 3
air	oxygen	oxygen	Oxygen (gas) amount or concentration at the time of sampling	measurement value	{float} {unit}	600 parts per million	X	milligram per liter, parts per million	1	1	MIXS.000010 4
air	oxy_stat_samp	oxygenation status of sample	Oxygenation status of sample	enumeration	[aerobic|anaerobic|other]	aerobic	X		1	1	MIXS.000075 3
air	perturbation	perturbation	Type of perturbation, e.g. chemical administration, physical disturbance, etc., coupled with perturbation regimen including how many times the perturbation was repeated, how long each perturbation lasted, and the start and end time of the entire perturbation period; can include multiple perturbation types	perturbation type name;perturbation interval and duration	{text};{R#/start_time/end_time/duration}	antibiotic addition;R2/2018-05-11T14:30Z/2018-05-11T19:30Z/P1H30M	X		m	1	MIXS.000075 4
air	pollutants	pollutants	Pollutant types and, amount or concentrations measured at the time of sampling; can report multiple pollutants by entering numeric values preceded by name of pollutant	pollutant name;measurement value	{text};{float} {unit}	lead;0.15 microgram per cubic meter	X	gram, mole per liter, milligram per liter, microgram per cubic meter	m	1	MIXS.000010 7
air	resp_part_matter	respirable particulate matter	Concentration of substances that remain suspended in the air, and comprises mixtures of organic and inorganic substances (PM10 and PM2.5); can report multiple PM's by entering numeric values preceded by name of PM	particulate matter name;measurement value	{text};{float} {unit}	PM2.5;10 microgram per cubic meter	X	microgram per cubic meter	m	1	MIXS.000010 8
air	samp_salinity	sample salinity	Salinity is the total concentration of all dissolved salts in a liquid or solid (in the form of an extract obtained by centrifugation) sample. While salinity can be measured by a complete chemical analysis, this method is difficult and time consuming. More often, it is instead derived from the conductivity measurement. This is known as practical salinity. These derivations compare the specific conductance of the sample to a salinity standard such as seawater	measurement value	{float} {unit}	1 milligram per liter	X	milligram per liter, practical salinity unit, percentage	1	1	MIXS.000010 9
air	samp_store_dur	sample storage duration	Duration for which the sample was stored	duration	{duration}	P1Y6M	X		1	1	MIXS.000011 6
air	samp_store_loc	sample storage location	Location at which sample was stored, usually name of a specific freezer/room	location name	{text}	Freezer no:5	X		1	1	MIXS.000075 5
air	samp_store_temp	sample storage temperature	Temperature at which sample was stored, e.g. -80 degree Celsius	measurement value	{float} {unit}	-80 degree Celsius	X	degree Celsius	1	1	MIXS.000011 0
air	samp_vol_we_dna_ext	sample volume or weight for DNA extraction	Volume (ml), weight (g) of processed sample, or surface area swabbed from sample for DNA extraction	measurement value	{float} {unit}	1500 milliliter	X	milliliter, gram, milligram, square centimeter	1	1	MIXS.000011 1
air	solar_irradiance	solar irradiance	The amount of solar energy that arrives at a specific area of a surface during a specific time interval	measurement value	{float} {unit}	1.36 kilowatts per square meter per day	X	kilowatts per square meter per day, ergs per square centimeter per second	1	1	MIXS.000011 2
air	temp	temperature	Temperature of the sample at the time of sampling	measurement value	{float} {unit}	25 degree Celsius	X	degree Celsius	1	1	MIXS.000011 3
air	ventilation_rate	ventilation rate	Ventilation rate of the system in the sampled premises	measurement value	{float} {unit}	750 cubic meter per minute	X	cubic meter per minute, liters per second	1	1	MIXS.000011 4
air	ventilation_type	ventilation type	Ventilation system used in the sampled premises	ventilation type name	{text}	Operable windows	X		1	1	MIXS.000075 6
air	volatile_org_comp	volatile organic compounds	Concentration of carbon-based chemicals that easily evaporate at room temperature; can report multiple volatile organic compounds by entering numeric values preceded by name of compound	volatile organic compound name;measurement value	{text};{float} {unit}	formaldehyde;500 nanogram per liter	X	microgram per cubic meter, parts per million, nanogram per liter	m	1	MIXS.000011 5
air	wind_direction	wind direction	Wind direction is the direction from which a wind originates	wind direction name	{text}	Northwest	X		1	1	MIXS.000075 7
air	wind_speed	wind speed	Speed of wind measured at the time of sampling	measurement value	{float} {unit}	21 kilometer per hour	X	meter per second, kilometer per hour	1	1	MIXS.000011 8

Métadonnées

BioSamples

BioSamples stocke et fournit des descriptions et des métadonnées sur les échantillons biologiques utilisés dans la recherche et le développement par les universités et l'industrie.

ENA
European Nucleotide Archive

Array express

Sample

Sample content reference:

Field	Description	Type	Cardinality =======
name	The short name of the sample.	`String`	Required
release	The date at which the sample was first made public.	`Date ISO 8601`	Required
update	The date at which the sample was last updated.	`Date ISO 8601`	System Generated
domain	The AAP domain the sample belongs to.	`String`	Required
accession	The sample unique identifier in the BioSamples database. If not provided, one will be automatically assigned.	`String`	Required for `PUT` requests
characteristics	The key-value pairs representing the attributes of the sample.	`Object`	Optional
externalReferences	A list of links towards external references, such as datasets in other archives.	`Array`	Optional
relationships	A list of relationships this sample has to other, existing, samples.	`Array`	Optional
data	A more structured data format to allow submission of tables (eg. antibiogram) in addition to key-value pairs.	`Array`	Optional, required only for structured data submission using our POST, PUT or PATCH endpoints. Note - You must provide an AAP domain for your data. It can be same as the sample domain if you are the submitter of both the sample metadata and structured data.

https://www.ebi.ac.uk/biosamples/docs/references/api/submit#_submission_minimal_fields

Biosamples - Exemples

COVID-19

SAMN15743942

CS0753_1

Download as:

XML

JSON

Bioschemas

Phenopacket

Attributes

Type	Value
External Id	SAMN15743942
INSDC center name	CSIR-Institute of Genomics and Integrative Biology
INSDC first public	2020-08-10T00:00:00Z
INSDC last update	2020-08-10T13:03:44.987Z
INSDC secondary accession	SRS7175978
INSDC status	live
NCBI submission model	Pathogen.cl
NCBI submission package	Pathogen.cl.1.0
SRA accession	SRS7175978
organism	Severe acute respiratory syndrome coronavirus 2
replicate	Biological Replicate 753
strain	SARS-CoV-2
title	Negative Control 1

Certificates

Name	Version	File Name
biosamples-minimal	0.0.1	schemas/certification/biosamples-minimal.json
ncbi-candidate-schema	0.0.1	schemas/certification/ncbi-candidate-schema.json

Released on	2020 / 08 / 10 00:00:00 UTC
Created on	2020 / 08 / 06 02:48:18 UTC
Updated on	2020 / 10 / 15 02:49:50 UTC

EBiSC

European Bank for induced pluripotent Stem Cells

SAMEA4675912

CHDI-90002200

Attributes

Certificates

Relationships

External Links

FAANG

Functional Annotation of Animal Genomes

SAMEA104728707

ECA_UCD_S186

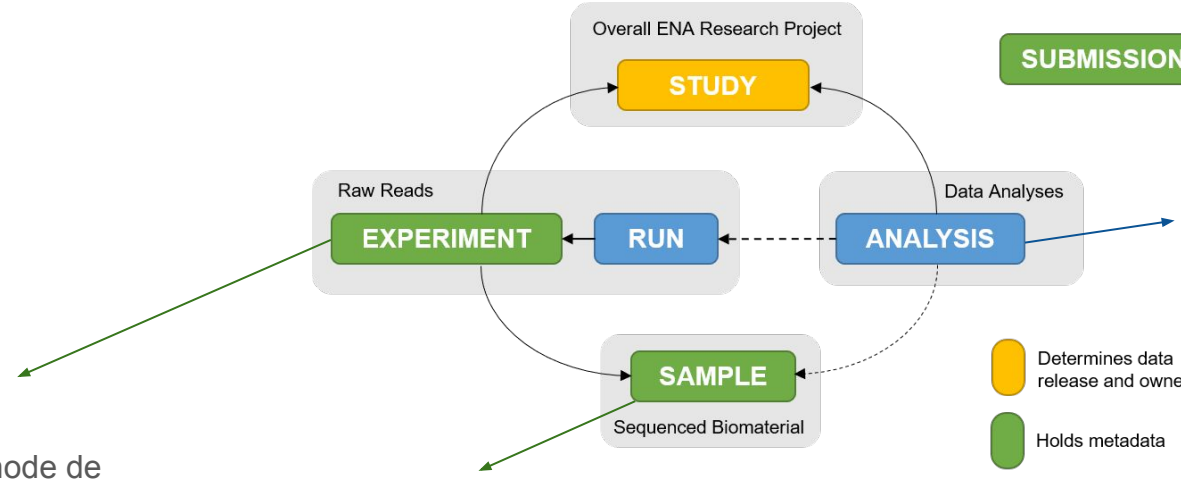
Attributes

Contacts

Organizations

Relationships

Modèle des métadonnées à l'ENA



Accessible le 13 novembre 2020
https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html

La qualité actuelle des métadonnées

The variable quality of metadata about biological samples used in biomedical experiments
Rafael S. Gonçalves & Mark A. Musen
Scientific Data volume 6, Article number: 190021 (2019)

11.4 millions de métadonnées testés issues de BioSample (NCBI) et BioSamples (EBI)

Bilan

- Les noms et les valeurs ne sont pas contrôlés ni standardisés
- Les valeurs ne sont pas toujours du bon type (ex : binaire)
- Manque d'outils de validation qui éviteraient des aberrances