

The BioSamples Database

Supporting multi-omics data integration with
FAIR sample records

Tony Burdett

Head of BioSamples & Joint Head of ENA

tburdett@ebi.ac.uk



A bit about me...



- I joined EBI in 2005
- I have a biological/medical background
- My career has been heavily focused on service engineering in bioinformatics and FAIR data management
- I've built, helped develop, or run the development teams for...

- ArrayExpress



- Expression Atlas



- BioSamples



- Ontology tooling

- GWAS Catalog

- Human Cell Atlas DCP



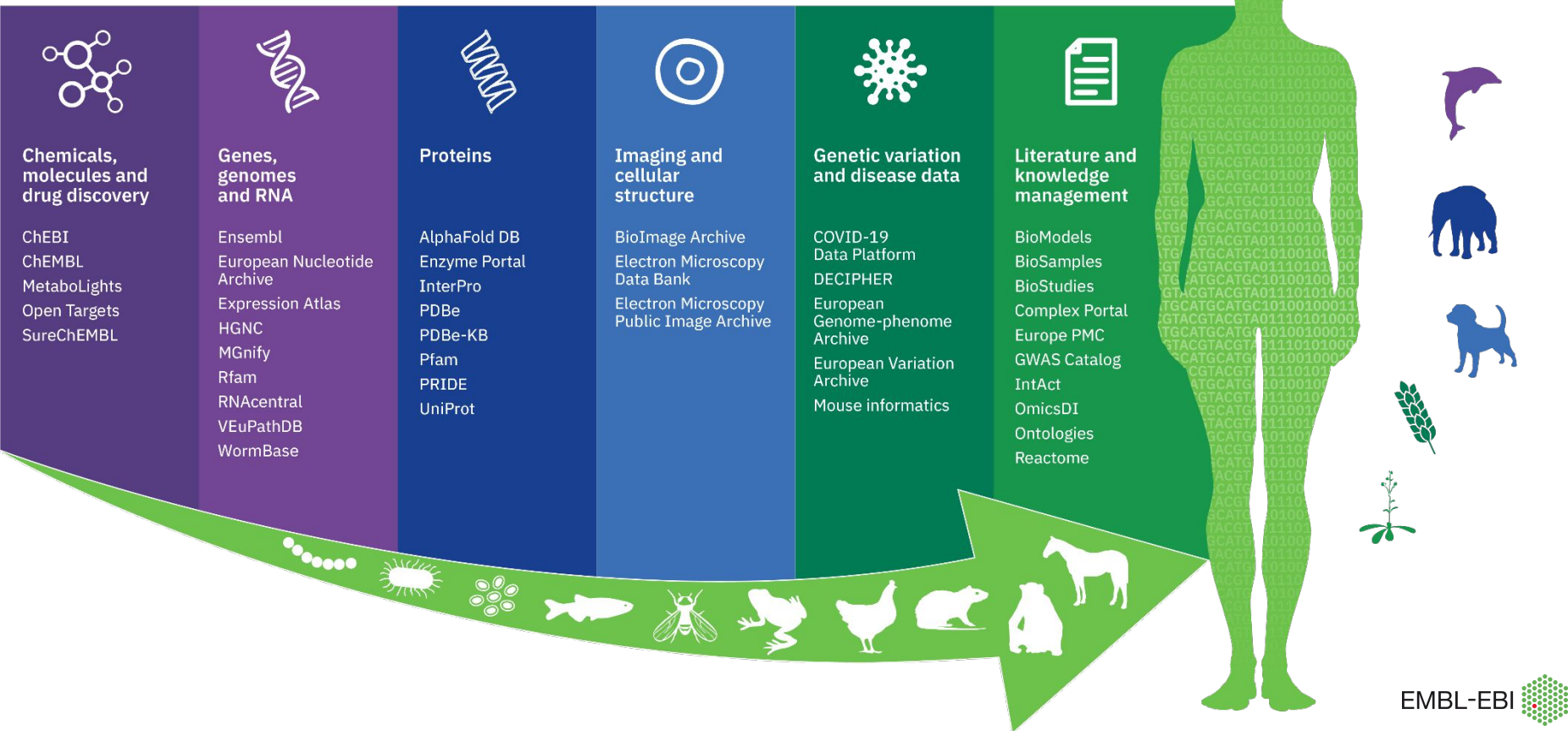
HUMAN
CELL
ATLAS

What is EMBL-EBI?

- World leading source of public biomolecular data
- Our vision is to benefit humankind by advancing scientific discovery and impact through bioinformatics.
- Part of the European Molecular Biology Laboratory (EMBL), Europe's flagship laboratory for the life sciences.



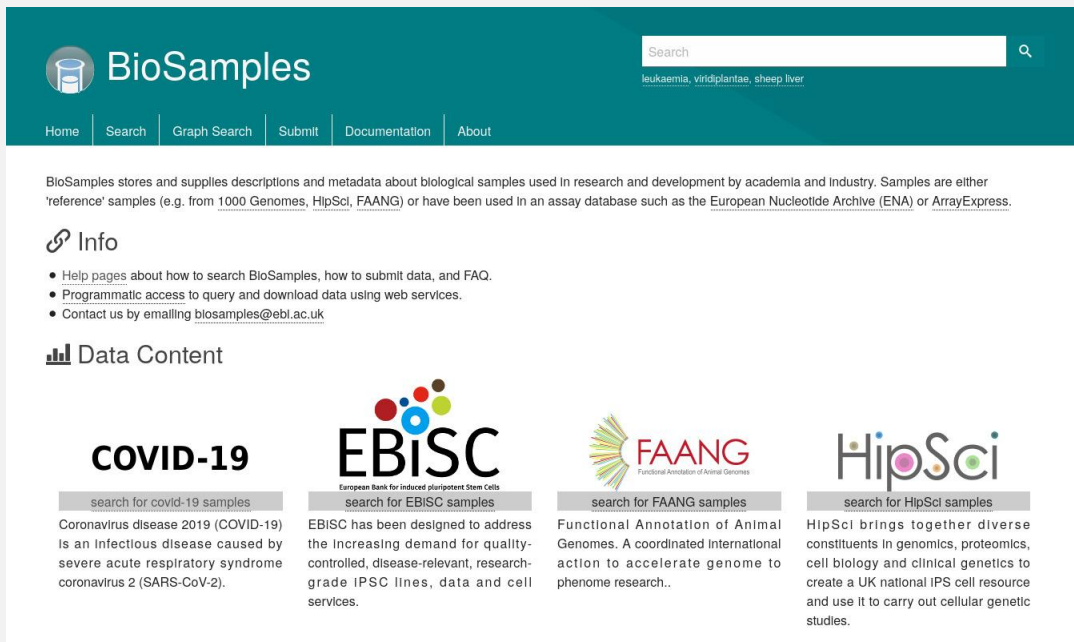
Data resources at EMBL-EBI



BioSamples

The BioSamples database, an ELIXIR deposition database for sample metadata storage and linkage to other EMBL-EBI resources.

BioSamples stores and supplies descriptions and metadata about biological samples used in research and development by academia and industry.



The screenshot shows the BioSamples website. The header is teal with the BioSamples logo and a search bar. Below the header is a navigation menu with links: Home, Search, Graph Search, Submit, Documentation, and About. The main content area has a teal background. It starts with a paragraph about the database's purpose. Below this is an 'Info' section with a list of links: 'Help pages about how to search BioSamples, how to submit data, and FAQ.', 'Programmatic access to query and download data using web services.', and 'Contact us by emailing biosamples@ebi.ac.uk'. The 'Data Content' section features four cards: 'COVID-19' (search for covid-19 samples), 'EBISC' (search for EBISC samples), 'FAANG' (search for FAANG samples), and 'HipSci' (search for HipSci samples). Each card provides a brief description of the dataset.

BioSamples

Search

leukaemia, viridiplantae, sheep liver

Home | Search | Graph Search | Submit | Documentation | About

BioSamples stores and supplies descriptions and metadata about biological samples used in research and development by academia and industry. Samples are either 'reference' samples (e.g. from [1000 Genomes](#), [HipSci](#), [FAANG](#)) or have been used in an assay database such as the [European Nucleotide Archive \(ENA\)](#) or [ArrayExpress](#).

Info

- [Help pages](#) about how to search BioSamples, how to submit data, and FAQ.
- [Programmatic access](#) to query and download data using web services.
- [Contact us](#) by emailing biosamples@ebi.ac.uk

Data Content

COVID-19
search for covid-19 samples
Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

EBISC
search for EBISC samples
European bank for induced pluripotent Stem Cells
EBISC has been designed to address the increasing demand for quality-controlled, disease-relevant, research-grade iPSC lines, data and cell services.

FAANG
search for FAANG samples
Functional Annotation of Animal Genomes. A coordinated international action to accelerate genome to phenotype research..

HipSci
search for HipSci samples
HipSci brings together diverse constituents in genomics, proteomics, cell biology and clinical genetics to create a UK national IPS cell resource and use it to carry out cellular genetic studies.

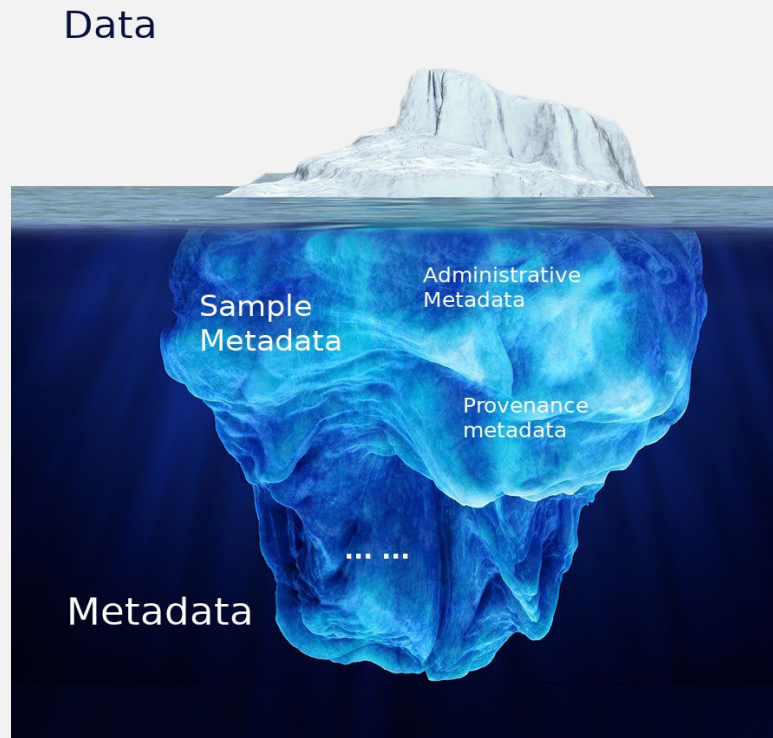
Metadata

Metadata, data about data

Structured information that describes, explains, locates or make it easier to retrieve, use, or manage an information resource.

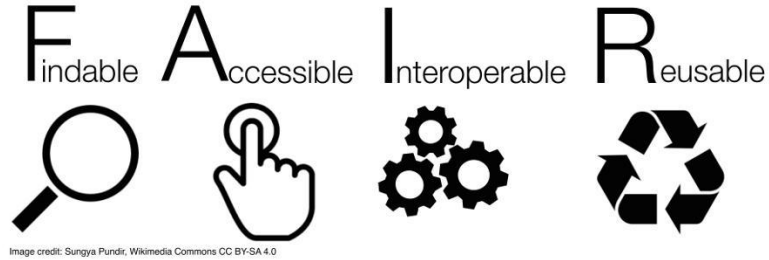
Sample metadata, for the better use of multi-omics data

- Description of biomaterial used in the experiment
- Support data interpretation.
- Improves data reusability and reproducibility

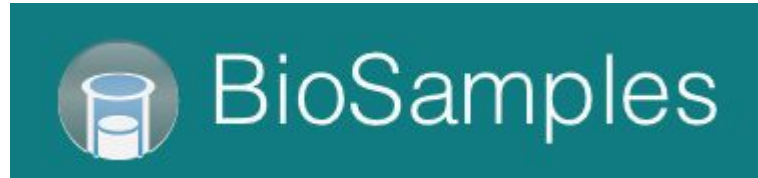


Good sample management needs...

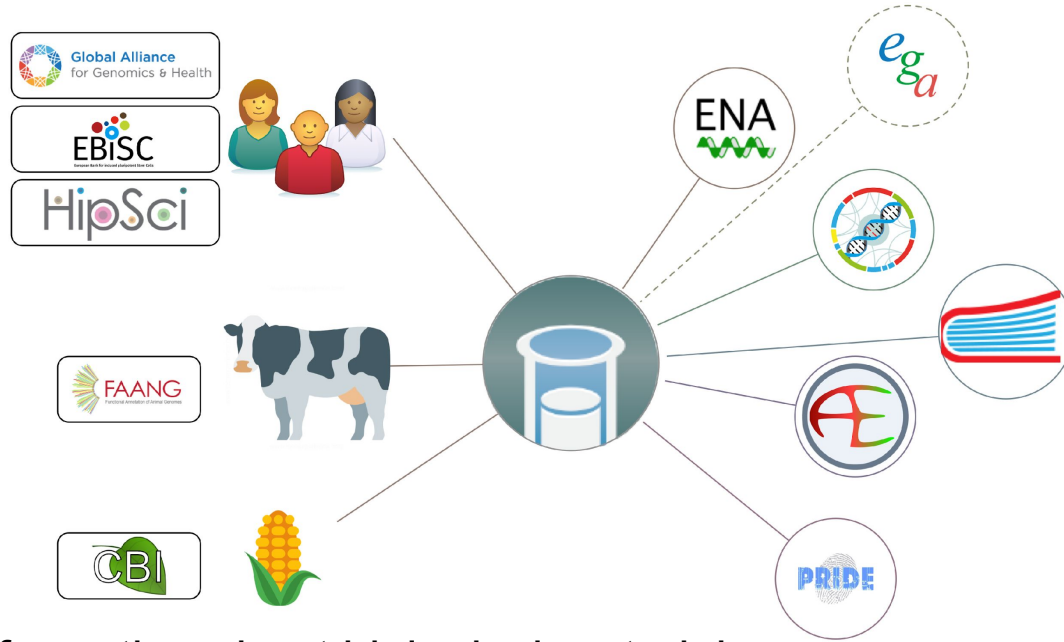
- FAIR data



- Infrastructure



BioSamples as a metadata hub



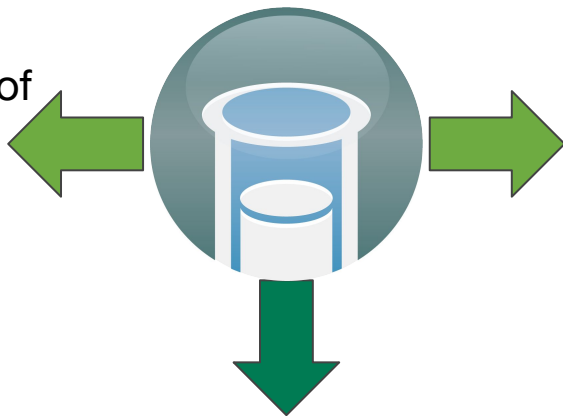
- Archive of information about biological materials
- From internal and external sources, and directly from submitters.
- Enables technology independent linking between assay data and sample metadata

BioSamples: Our mission to add value



Data owners:

- Where to publish details of our samples?
- What are the essential information?
- How to register samples and submit data?



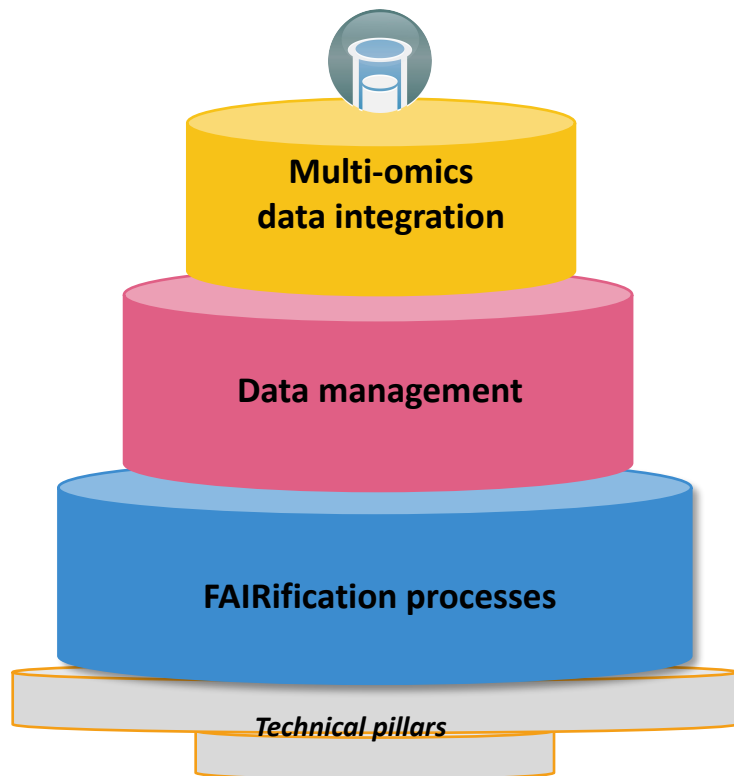
Data consumers:

- Is there (meta)data available?
- Where to find the (meta)data?
- How to search for the correct (meta)data?
- Is the (meta)data downloadable and interpretable?

FAIR (meta)data indicators:

- Metadata includes the identifier for the data
- Metadata is accessed through a standardized protocol
- Metadata uses standard vocabularies
- Metadata complies with a community standard

Biosamples “Layer Cake”



Interconnectivity across archives

- Single hosting place for sample metadata
 - Linking sample, assay and publication

Data deposition for collaborative projects

- Easier submission process
- Adjust to community requirements

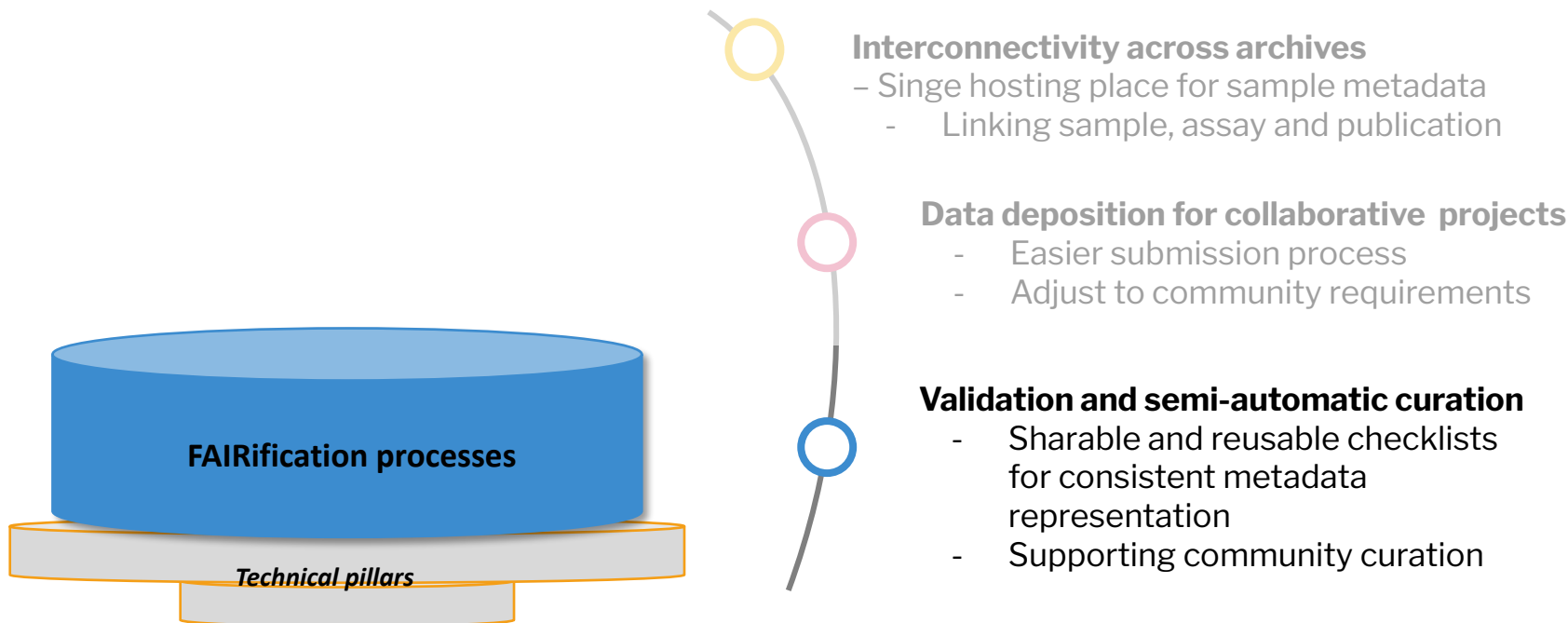
Validation and semi-automatic curation

- Sharable and reusable checklists for consistent metadata representation
- Supporting community curation



FAIRification processes

Biosamples “Layer Cake”



Use Cases: FAIRification processes



- Validation against sample checklists - e.g. MixS, MIAPPE - at the point of submission
- Retrospective certification against sample checklists (including entirely new checklists or simply newer versions)
- Integration with ELIXIR BioValidator
- (Semi-)Automated ontology annotation
- Ontology-enabled search

FAIR, specifically, in BioSamples context...

Findable

- Globally unique, resolvable, and persistent **identifiers**
- Machine-readable **descriptions** to support structured search and filtering

Accessible

- **Metadata** are accessible beyond the lifetime of the digital resource
- Clearly defined **access and security protocols (FAIR != Open)**

Interoperable

- Extensible **machine interpretable** formats for data + metadata
- Use **vocabularies** and **link** to other resources

Reusable

- Provide **licensing, provenance** and meet community **standards**

Credit: Michael Dunningham, Newcastle University

BioSample records

- A FAIR digital record
- Persistent identifier
- Open file format
- External links to data

SAMEA6807931

COG-UK/CAMB-785F5

Download as:

[XML](#)
[JSON](#)
[Bioschemas](#)
[Phenopacket](#)

Attributes

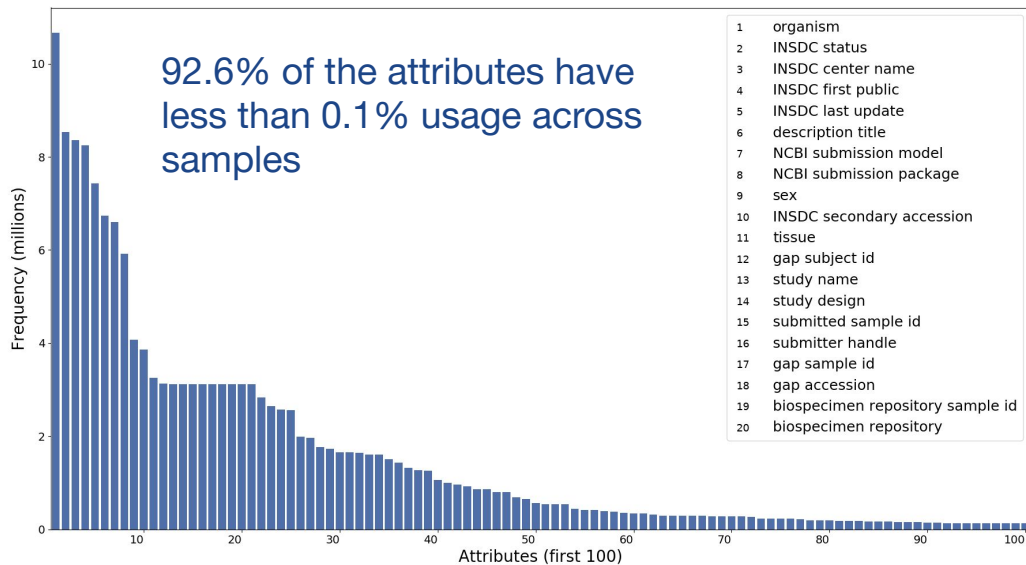
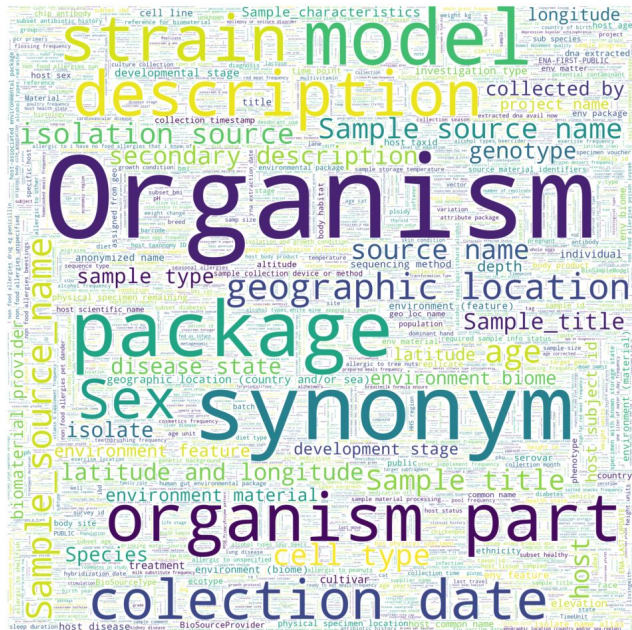
Type	Value
ENA first public	2020-04-29
ENA last update	2020-04-29
ENA-CHECKLIST	ERC000011
External Id	SAMEA6807931
INSDC center alias	University of Cambridge
INSDC center name	University of Cambridge
INSDC first public	2020-04-29T16:22:10Z
INSDC last update	2020-04-29T16:22:10Z
INSDC status	public
SRA accession	ERS4535582
Submitter Id	COG-UK/CAMB-785F5
broker name	COVID-19 Genomics UK Consortium
collection date	2020-04-06
geographic location (country and/or sea)	United Kingdom
geographic location (region and locality)	England
organism	Severe acute respiratory syndrome coronavirus 2
title	COG-UK/CAMB-785F5

External Links

ENA

Metadata curation

Redundancy and inconsistency in real life data





How to find all COVID-19 samples?

Covid 19-related attributes in BioSamples:

- *severe acute respiratory syndrome*
- *COVID19*
- *novel coronavirus pneumonia*
- *nCoV pneumonia*
- *COVID-19*
- *Coronavirus infected disease-19 (COVID-19)*

Metadata curation

- Text curation
- Semantic annotation

Common challenges in sample metadata

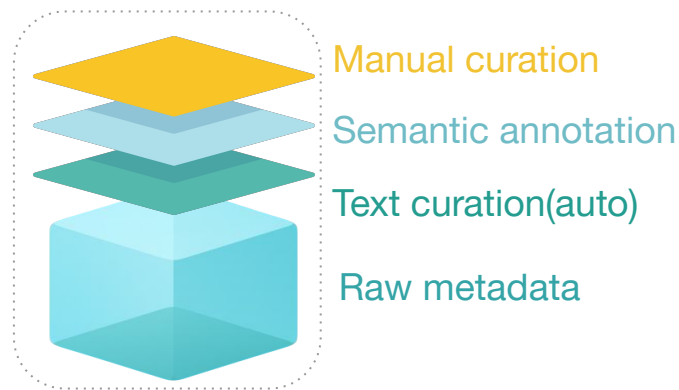
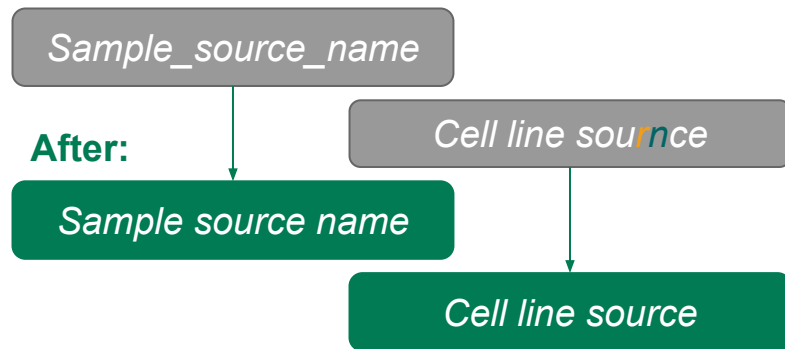
- Special characters (*COVID19* vs *COVID-19*)
- Acronyms (*T2D* for *diabetes*)
- Typo
- Synonyms



Text curation

- Automatic curation by pipelines
 - Remove empty values
 - Handling special characters
 - Normalize attribute format
- Manual curation by experts
- Curation tool based on manual curation and machine learning

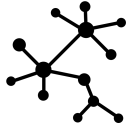
Before:



Semantic annotation

Ontology-based semantic annotation

Ontology:



A formal representation and definition of concepts and categories in a subject area or domain that shows their properties and the relations between them.

Example ontologies: Gene Ontology, EFO



Automated ontology annotation

ZOOMA

An annotation service for finding possible ontology terms for free text terms.



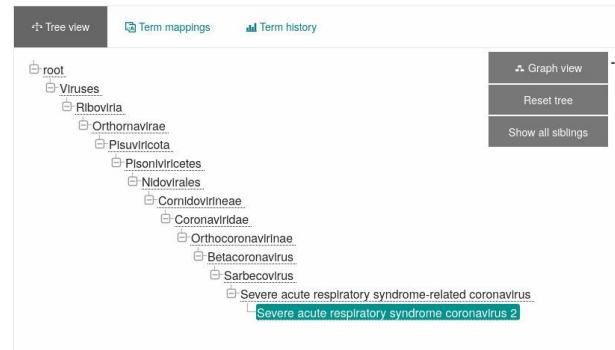
SARS-CoV-2 related values

- SARS-CoV2
- Wuhan coronavirus
- Human coronavirus 2019
- SARS-CoV-2
- 2019-nCoV
- COVID-19 virus

OLS / NCBI organismal classification [NCBITAXON](#) / [NCBITaxon:2697049](#) [Copy](#)

Severe acute respiratory syndrome coronavirus 2

http://purl.obolibrary.org/obo/NCBITaxon_2697049 [Copy](#)



Access and visualize ontologies

Ontology Lookup Service (OLS)

A repository for biomedical ontologies that aims to provide a single point of access to the latest ontology versions.

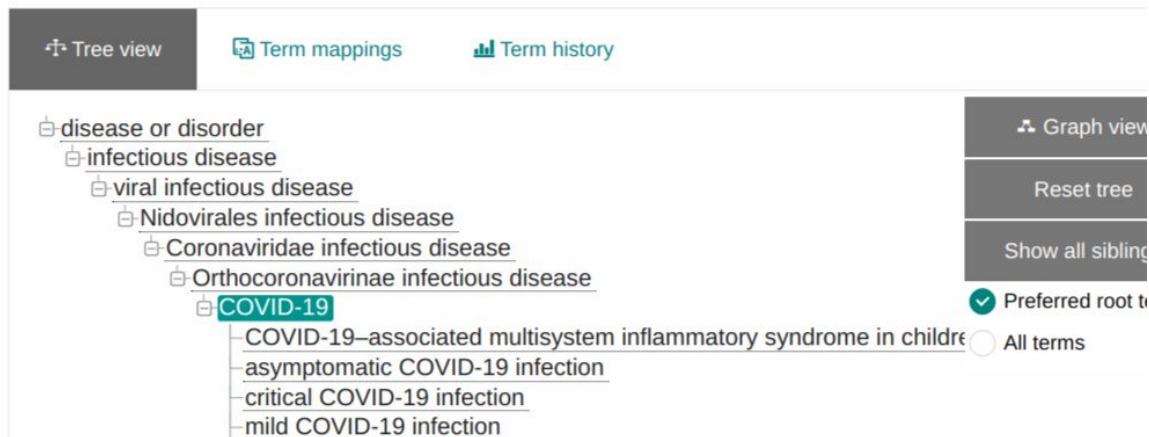


COVID-19

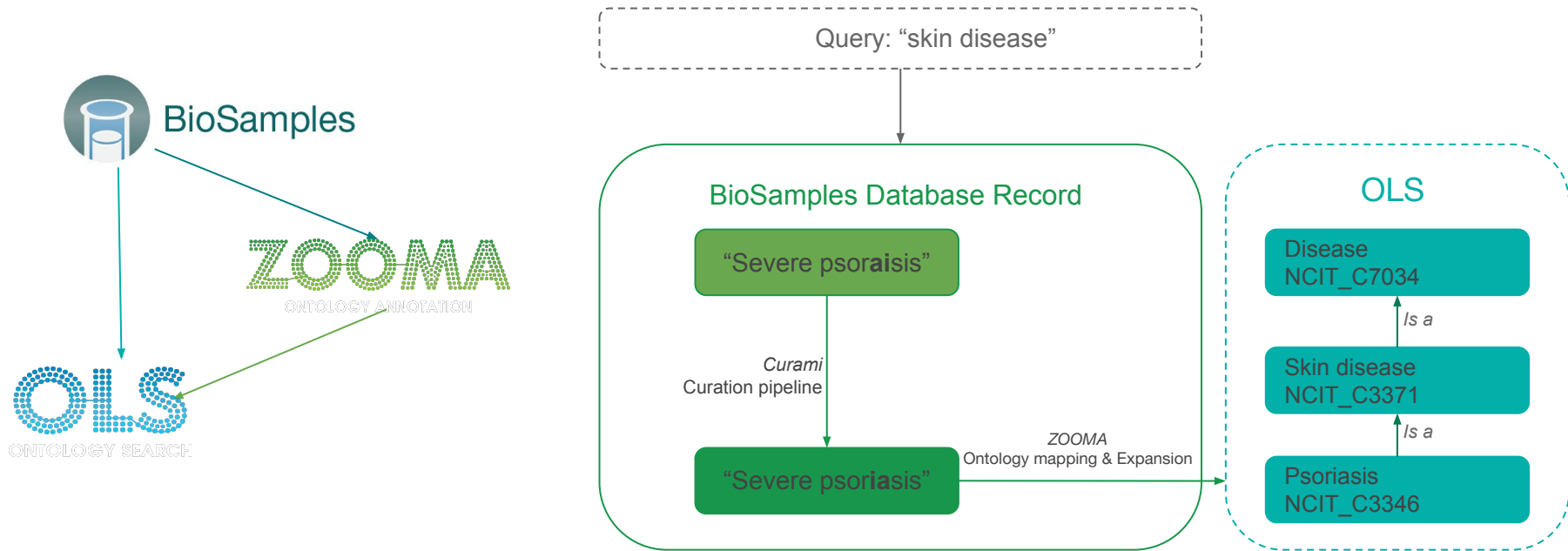
http://purl.obolibrary.org/obo/MONDO_0100096 [Copy](#)

A disease caused by infection with severe acute respiratory syndrome coronavirus 2. [[https://www.cdc.gov/c](https://www.cdc.gov/coronavirus)]

Synonyms: [2019 novel coronavirus infection](#) [2019-nCoV infection](#) [coronavirus disease 2019](#)



Curation improves findability



Ontology powered search expansion

 BioSamples

cancer

leukaemia, viridiplantae, sheep liver

[Home](#) | [Search](#) | [Submit](#) | [Documentation](#) | [About](#)

Search results

Apply filters

Clear filters

Organism	362105
Homo sapiens	315490
Mus musculus	17270
human gut metagenome	9079
Saccharomyces cerevisiae	6347
Drosophila melanogaster	2306
Rattus norvegicus	1820
Caenorhabditis elegans	856
human skin metagenome	745

Showing 479991 to 480000 of 521477 results

« Previous 1 ... 47999 **48000** 48001 ... 52148 Next »

[source GSM509314 1](#)

SAMEA107475

Updated on: 04-04-2018 16:02

Organism **Homo sapiens** Sample_source_name **HL-60 cells treated with DMSO for 3 h**

cell type **human promyelocytic leukemia cells**

description **HL-60 cells were treated with 0.99 mM ethylbenzene for 3 h, and the RNA was subjected to microarr...**

strain **HL-60** treatment **DMSO** treatment time **3 h** has member (reverse) **SAMEG10295**

Ontology powered search expansion

 BioSamples

leukaemia, viridiplantae, sheep liver

[Home](#) | [Search](#) | [Submit](#) | [Documentation](#) | [About](#)

Search results

[Apply filters](#) [Clear filters](#)

Organism	362105
Homo sapiens	315490
Mus musculus	17270
human gut metagenome	9079
Saccharomyces cerevisiae	6347
Drosophila melanogaster	2306
Rattus norvegicus	1820
Caenorhabditis elegans	856
human skin metagenome	745

Showing 479991 to 480000 of 521477 results

« Previous 1 ... 47999 **48000** 48001 ... 521477 Next

[source GSM509314 1](#)

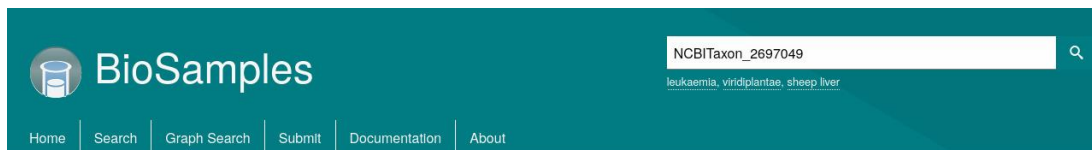
SAMEA107475

Updated on: 04-04-2018 16:02

Organism **Homo sapiens** Sample_source_name **HL-60 cells treated with DMSO for 3 h**
cell type **human promyelocytic leukemia cells**
description **HL-60 cells were treated with 0.99 μ M ethylbenzene for 3 h, and the RNA was subjected to microarr...**
strain **HL-60** treatment **DMSO** treatment time **3 h** has member (reverse) **SAMEG10295**

Search “cancer”
expands to
“leukemia”

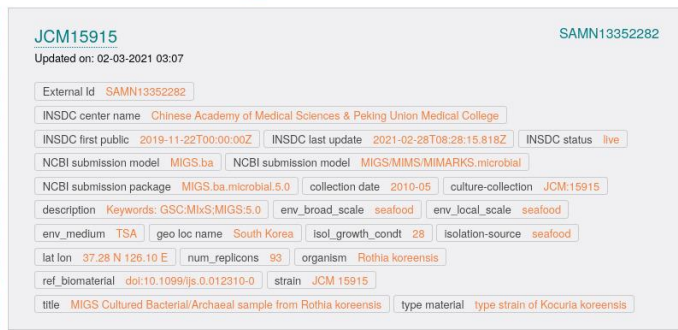
Curation improves findability



The BioSamples search interface features a teal header with the BioSamples logo on the left and a search bar on the right. The search bar contains the text 'NCBITaxon_2697049' and a magnifying glass icon. Below the search bar, the text 'leukaemia, viridiplantae, sheep liver' is displayed. The header also includes a navigation menu with links for Home, Search, Graph Search, Submit, Documentation, and About.

Search results for NCBITaxon_2697049 showing 1 to 10 of 362604 samples

Apply filters	Clear filters
organism	362601
Severe acute respiratory syndrome coronavi rus 2	249
Homo sapiens	249
human nasopharyngeal metagenome	45
Arachis hypogaea	42
Macaca mulatta	40
food metagenome	37
Severe acute respiratory syndrome-related coron avirus	22
Chlorocebus aethiops	17
Mustela putorius furo	12
Enterobacter hormaechei subsp. xiangfangensis	9
external reference	353542



The sample details for JCM15915 (SAMN13352282) are displayed. The sample was updated on 02-03-2021 03:07. The external ID is SAMN13352282. The INSDC center name is Chinese Academy of Medical Sciences & Peking Union Medical College. The INSDC first public date is 2019-11-22T00:00:00Z, and the INSDC last update is 2021-02-28T08:28:15.818Z. The INSDC status is live. The NCBI submission model is MIGS.ba, and the NCBI submission model is MIGS/MIMS/MIMARKS.microbial. The NCBI submission package is MIGS.ba.microbial.5.0, the collection date is 2010-05, and the culture-collection is JCM:15915. The description is Keywords: GSC.MixS:MIGS.5.0, env_broad_scale seafood, env_local_scale seafood. The env_medium is TSA, the geo loc name is South Korea, the isol_growth_condt is 28, and the isolation-source is seafood. The lat lon is 37.28 N 126.10 E, the num_replicons is 93, and the organism is Rothia koreensis. The ref_biomaterial is doi:10.1099/ijs.0.012310-0, and the strain is JCM 15915. The title is MIGS Cultured Bacterial/Archaeal sample from Rothia koreensis, and the type material is type strain of Kocuria koreensis.

Over 8.6M COVID samples recorded to date

One of the most comprehensive search systems for Covid-19 data in the world.

https://www.ebi.ac.uk/biosamples/samples?text=NCBITaxon_2697049

Curation improves findability



Search results

Apply filters

organism

Severe acute respiratory syndrome
Homo sapiens
human nasopharyngeal metagenome
Arachis hypogaea
Macaca mulatta
food metagenome
Severe acute respiratory syndrome
Chlorocebus aethiops
Mustela putorius furo
Enterobacter hormaechei subsp.

external reference

https://www.ebi.



COVID-19 Data Portal

About ▾

Tools ▾

FAQ

Related Resources

Bulk Downloads

Submit Data

[Viral Sequences](#)

[Host Sequences](#)

[Expression](#)

[Proteins](#)

[Networks](#)

[Cohorts](#)

[More ▾](#)

Samples

Biomaterials relating to SARS-CoV-2 and its research

Search

Examples: [ACE2](#) , [Severe acute respiratory syndrome 2](#)...

[Advanced search](#)

Showing 6 of 7,425,785 in All > Samples

Data types

All

Samples (7,425,785)

Assayed samples
(7,425,659)

Cell lines (126)

Assayed samples 7,425,659 results

COG-UK/PHWC-2AFB8

Source: [biosamples-covid19](#)

Oro-Nasopharyngeal swab

Source: [biosamples-covid19](#)

amples

n systems
ne world.

Data export standards



Which format for interoperable phenotype data?

BioSamples exports 35 million samples using shared community standard formats, and supports downloading as PFX files.



Data consumers

Download as:

XML

JSON

Bioschemas

Phenopacket

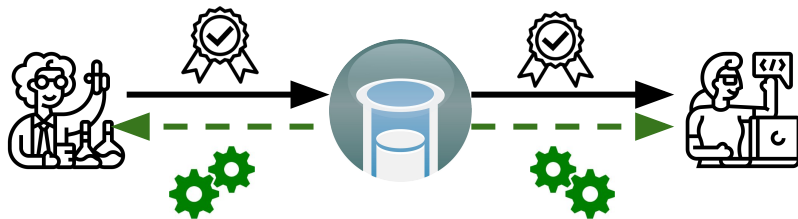


Phenopackets

- An open standard for sharing disease and phenotype information.
- Represented as PFX (Phenotype Exchange Format) files, which may be encoded in JSON or YAML

Tools for data standards

BioSamples also provides tools for checking data compliance through the data flow.



elixir-europe / json-schema-validator
forked from HumanCellAtlas/ingest-validator-js

Watch 4 Star 2 Fork 2

Code Pull requests 0 Projects 0 Wiki Insights

JavaScript validator for HCA metadata

247 commits 11 branches 6 releases 5 contributors Apache-2.0

Branch: master New pull request Create new file Upload files Find File Clone or download

This branch is 18 commits ahead of HumanCellAtlas:master. Pull request Compare

simonjupp release 1.3.8 with updated event stream to fix flatmap-stream vulnera... Latest commit 6e61ee2 on 19 Dec 2018

examples	3.1.3 release supports load schema by supplying a base file path	4 months ago
src	allow config OLS uri for expanding curies	3 months ago
test	pass keyword options to validator	3 months ago
.dockerignore	added ability to pass custom loadRefs function and OLS base url options	4 months ago
.gitignore	Restructure, better logging, docker and new custom keyword (#13)	9 months ago
.travis.yml	Restructure, better logging, docker and new custom keyword (#13)	9 months ago
Dockerfile	removed HCA specific code	4 months ago
LICENSE.md	Initial commit.	a year ago
README.md	update the docs	4 months ago
index.js	added index.js	4 months ago
package-lock.json	release 1.3.8 with updated event stream to fix flatmap-stream vulnera...	3 months ago
package.json	release 1.3.8 with updated event stream to fix flatmap-stream vulnera...	3 months ago

README.md

JSON Schema Validator

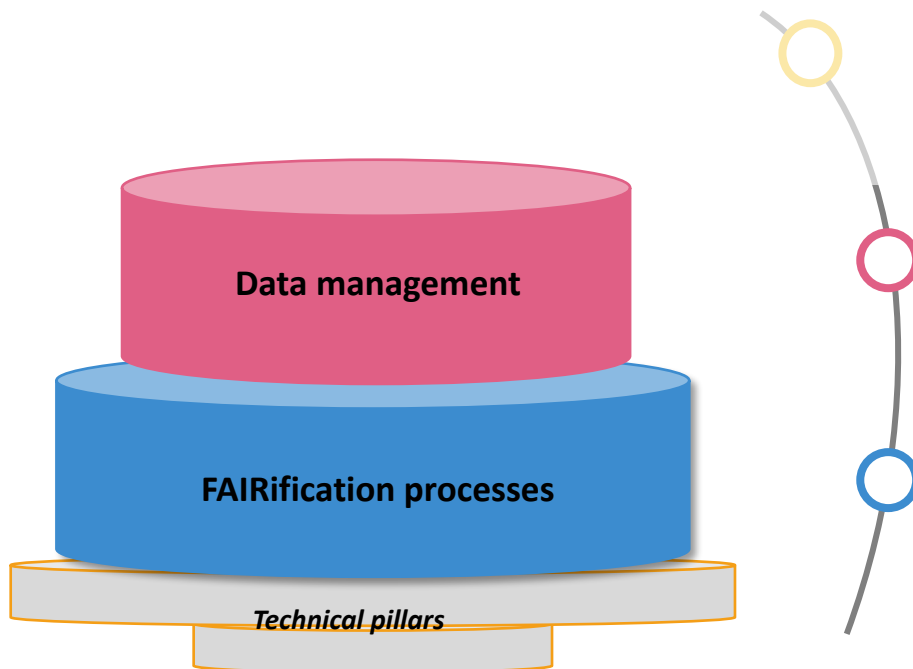
build passing tested with jest

This repository contains a [JSON Schema](#) that includes custom extensions of life science data. This package can be used directly or set to run as a node server that receives validation requests and gives back results. The validation is done using the [AJV](#) library version ^6.0.0 that fully supports the JSON Schema draft-07.



Data Management

Biosamples “Layer Cake”



Interconnectivity across archives

- Single hosting place for sample metadata
 - Linking sample, assay and publication

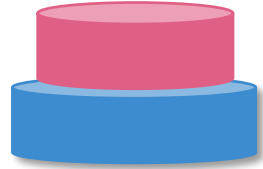
Data deposition for collaborative projects

- Easier submission process
- Adjust to community requirements

Validation and semi-automatic curation

- Sharable and reusable checklists for consistent metadata representation
- Supporting community curation

Use Cases: Data Management



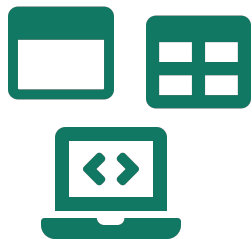
- Out-of-the-box sample management
- Integration with EBI search
- Integration with standards supported by ELIXIR (BioSchemas, AAI) and GA4GH (Phenopackets, DUO, GA4GH passports)
- Standardising sample:sample relationships (“derived from”, “child of”, “same as”)

Data Generation Lifecycle



[Meta]data
planning

- [Meta]data standard specification
- Consensus amongst key stakeholders



[Meta]data
collection

- Collection in a format that everyone understands
- Collection that meets the standard



[Meta]data
upload

- Upload service from data generator to storage platform (possibly staging)



[Meta]data
validation

- [Meta]data validated against the standard
- Experimental designs validated
- File formats validated



[Meta]data
brokering

- Submission of metadata and data files to final location

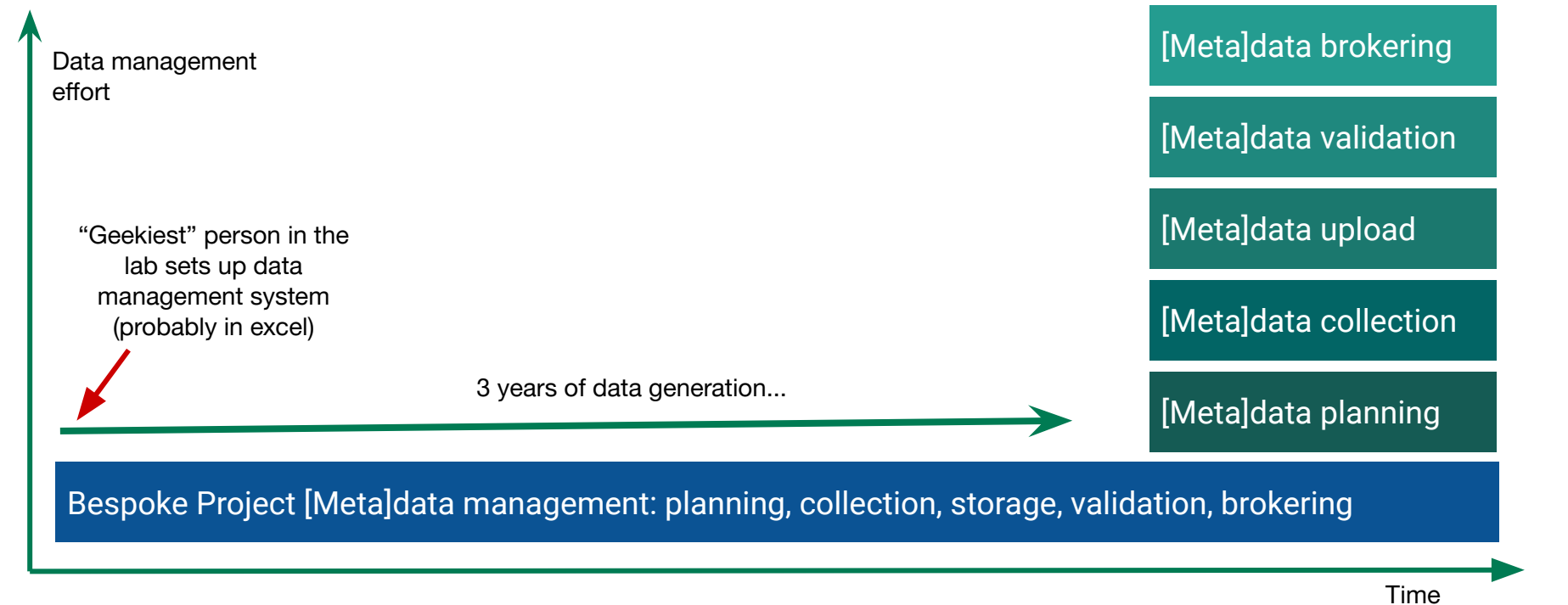
Current data sharing process



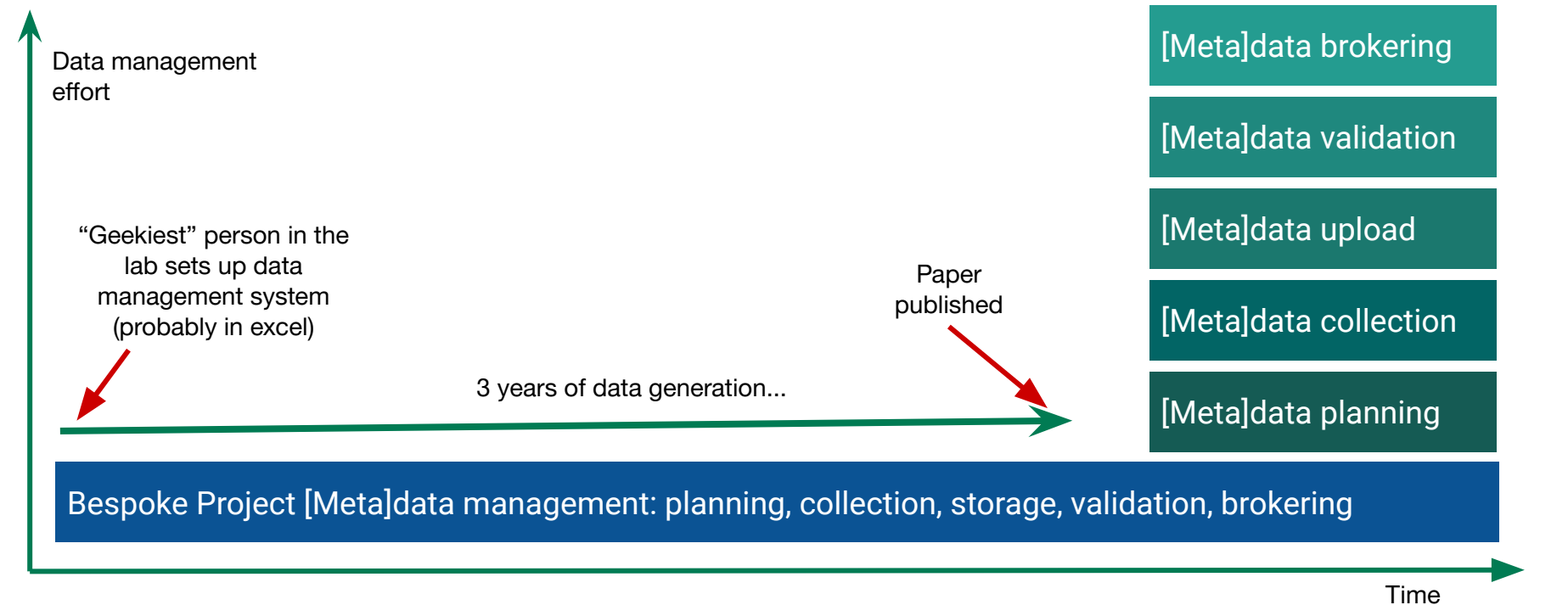
Current data sharing process



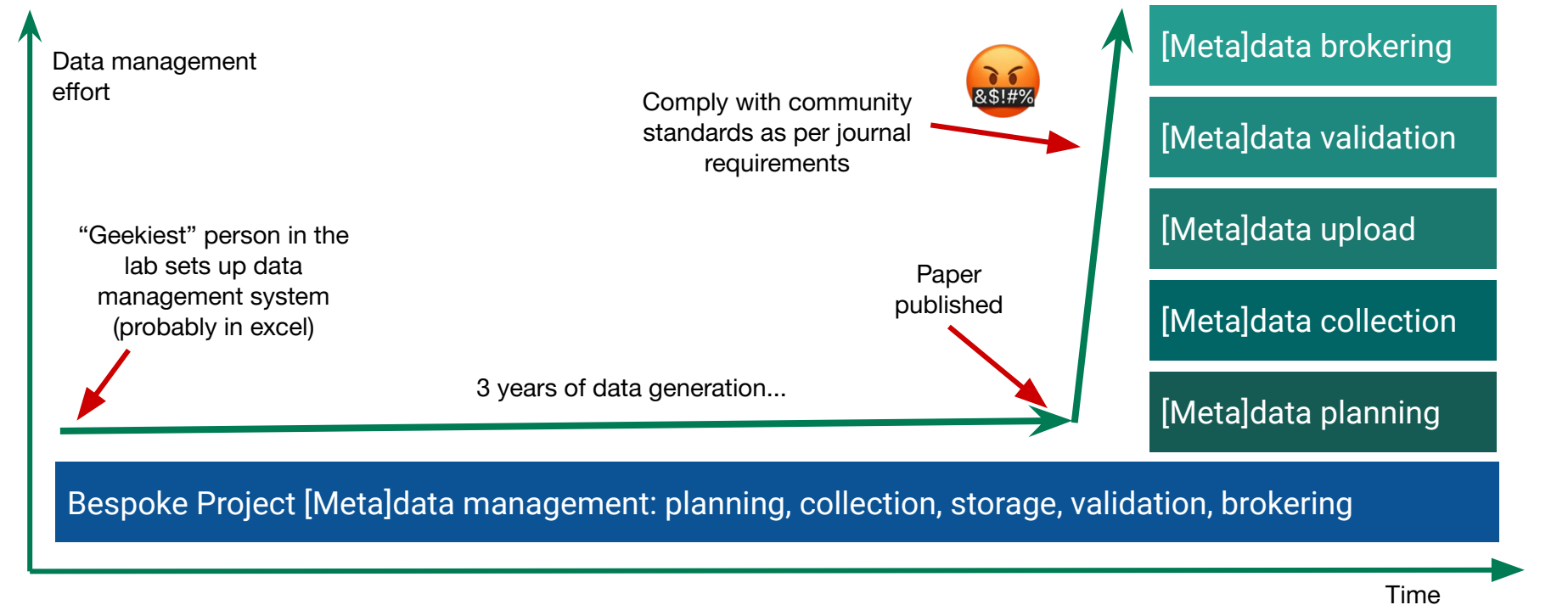
Current data sharing process



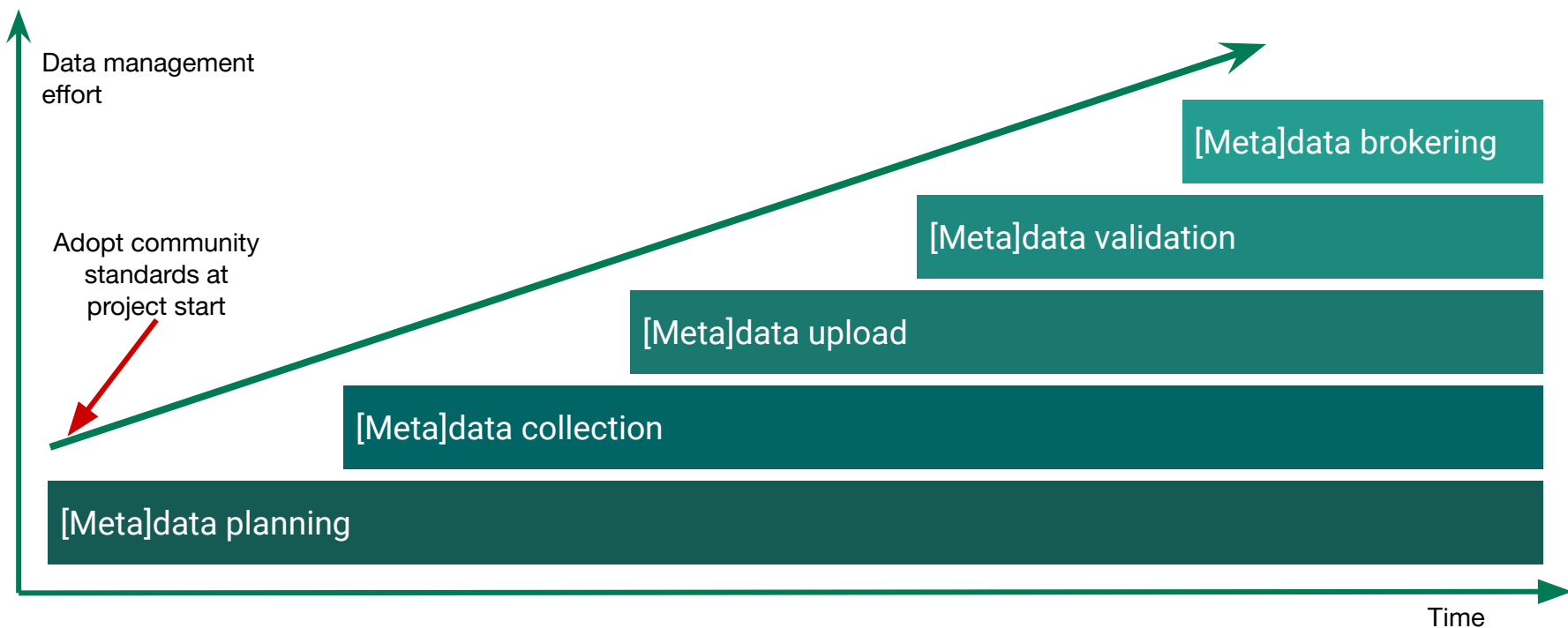
Current data sharing process



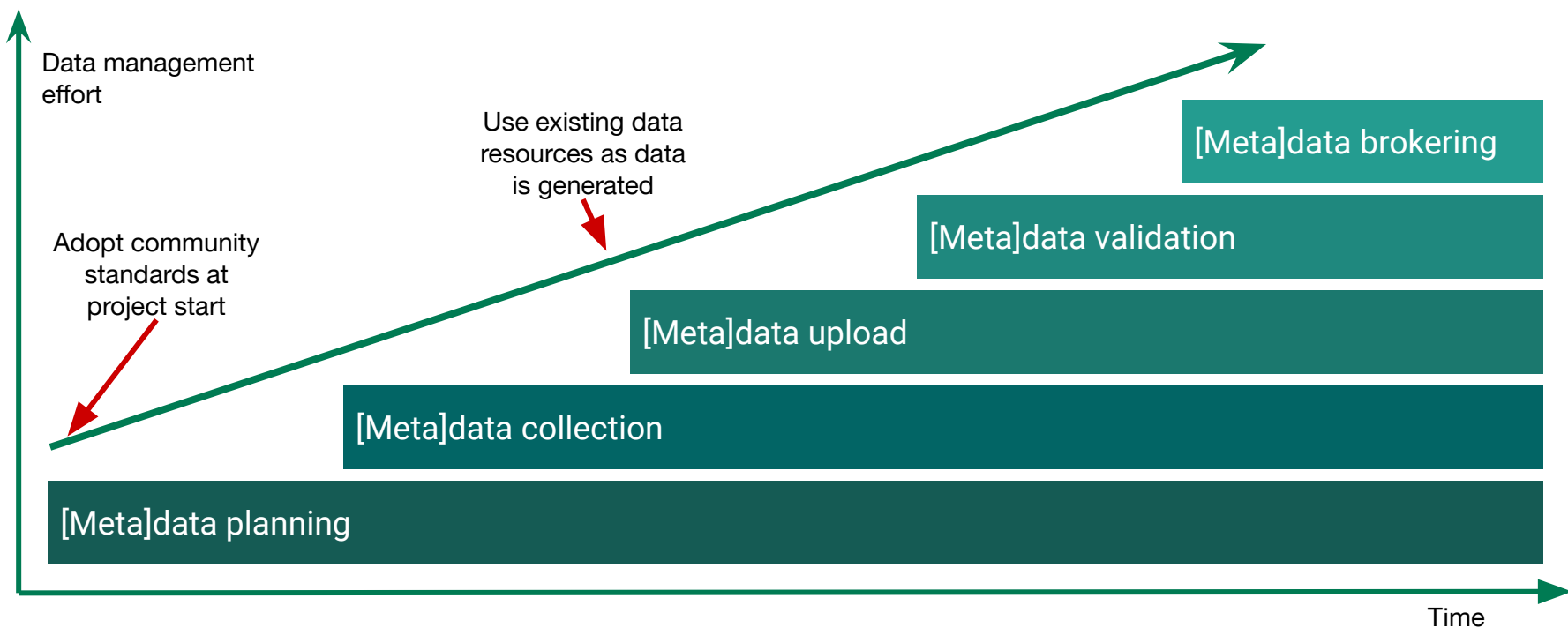
Current data sharing process



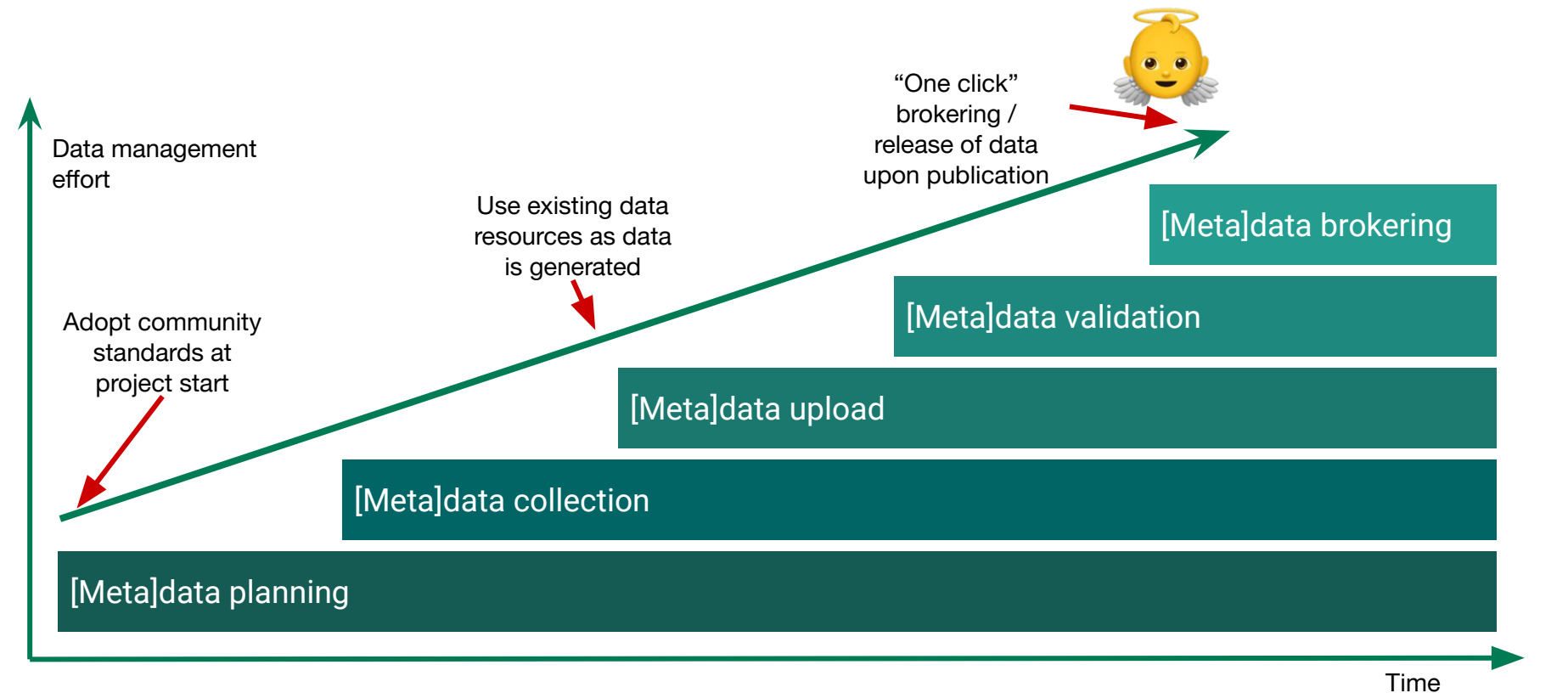
Better data sharing process



Better data sharing process



Better data sharing process



Data standards



ENA sample checklists

Requirements on the minimum metadata expected to describe biological samples to meet the needs of different research communities.

Checklist: ERC000033

ENA virus pathogen reporting standard checklist

Minimum information about a virus pathogen. A checklist for reporting metadata of virus pathogen samples associated with genomic data. This minimum metadata standard was developed by the COMPARE platform for submission of virus surveillance and outbreak data (such as Ebola) as well as virus isolate information.

Checklist Fields

Filter fields...	Q				
Filter by type:					
Human surveillance data					
Collection event information					
sample collection					
host disorder					
host description					
Virus isolate information					
hospitalisation	?	text choice	options	optional	
illness duration	?	free text		optional	
illness symptoms	?	free text		optional	
collection date	?	restricted text	regular expression ?	recommended	
geographic location (country and/or sea)	?	text choice	options	mandatory	
geographic location (latitude)	?	restricted text	regular expression ?	recommended	DD
geographic location (longitude)	?	restricted text	regular expression ?	recommended	DD

[Meta]data
planning

[Meta]data
collection

[Meta]data
upload

[Meta]data
validation

[Meta]data
brokering



Drag and drop one step sample submission

- Non-programmatic submitters
- Both instant response and asynchronous submissions (for very large sets of samples)
- Select a checklist for validation during submission
- Currently supports ISA-TAB but more could be added

[Submit Samples](#) [View Submissions](#) [Logout](#)

self.helpdesk

ENA-GSC_MlxS_host_associated(ERC000013)

0.5 MB

package-loc...

Submit

Metadata validation

BioSamples reuses the ENA sample checklists, validates samples against such checklists, and assigns certificates to valid sample records.

Certificates

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



Data
owners



[Meta]data
planning

[Meta]data
collection

[Meta]data
upload

[Meta]data
validation

[Meta]data
brokering

Checklist validation for data consistency

- Mandatory validation against the BioSamples minimum checklist. (All samples MUST have organism)
- Finding samples by checklists
<https://www.ebi.ac.uk/biosamples/samples?filter=attr:checklist:BSDC00001>

genus	69522
Hordeum	67174
Triticum	2346
Aegilops	2
checklist	69858
✓ BSDC00001	69858

Shared checklists registry

- Development of JSON schema store
- Integration with BioSamples
- MIAPPE and SARS-CoV-2 checklists

SCHEMA STORE HOME SEARCH EDITOR CHECKLIST ABOUT

Search Schema

Items per page: 5 1 - 5 of 43 < >

biosamples-minimal v1.0.0 BSDC00001
<https://www.ebi.ac.uk/biosamples/schemas/biosamples-minimal/1.0.0>
BioSamples minimal checklist
Minimum required fields to submit a sample into EBI BioSamples database
List of Versions Click to See more

plant-miappe v1.2.0 BSDC00002
<https://www.ebi.ac.uk/biosamples/schemas/plant-miappe/1.2.0>
Plant MIAPPE checklist for omics and phenomics interoperability
This is a checklist developed by MIAPPE group for Plant omics and phenomics interoperability
List of Versions Click to See more

[Meta]data
planning

[Meta]data
collection

[Meta]data
upload

[Meta]data
validation

[Meta]data
brokering

Collaboration with other communities

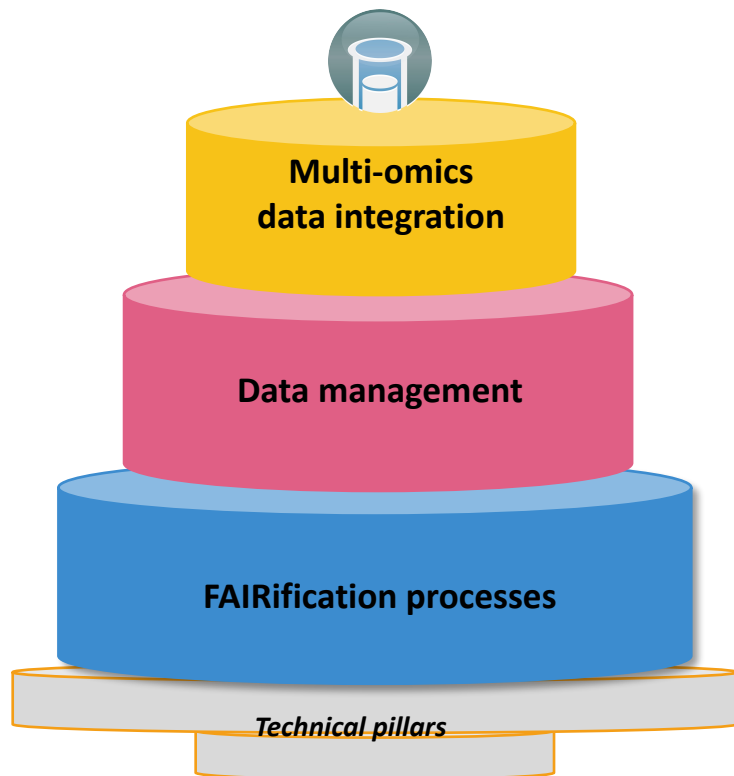
BioSamples works with multiple communities to provide services, develop standards, build tools, and support their use cases.





Multomics Data Integration

Biosamples “Layer Cake”



Interconnectivity across archives

- Single hosting place for sample metadata
 - Linking sample, assay and publication

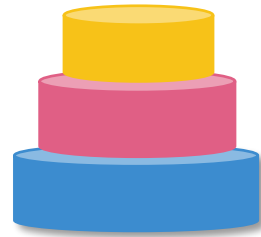
Data deposition for collaborative projects

- Easier submission process
- Adjust to community requirements

Validation and semi-automatic curation

- Sharable and reusable checklists for consistent metadata representation
- Supporting community curation

Use Cases: Multiomics Data Integration

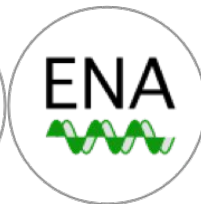
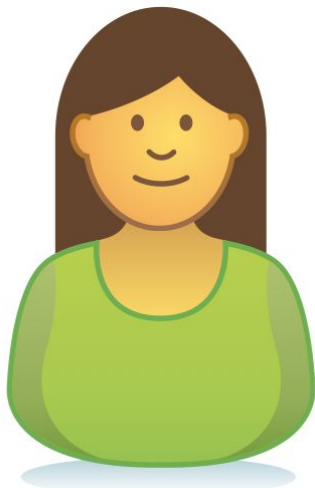


- Generally: links between different data modalities, taken from the same samples, supports combinatorial analysis, including...
 - Host/pathogen interactions (EGA to ENA in e.g. covid)
 - Functional variation effects (ENA/EVA to ArrayExpress)
 - Single cell spatial analysis (ENA/ArrayExpress to e.g. BioImage archive)
 - Any other sequence analysis (ENA to BioStudies)
 - ...and many more

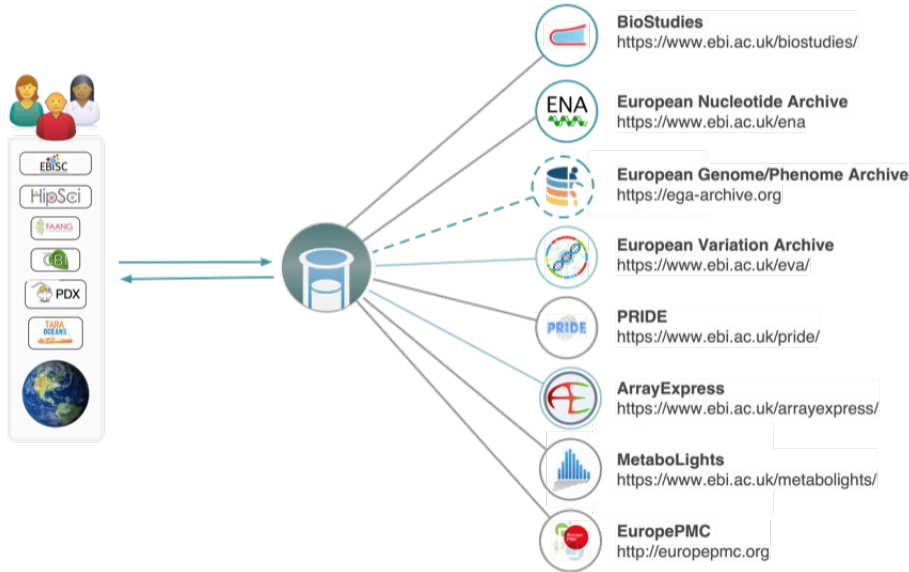
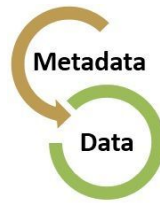
Example driver use case

Transcriptomics study

Project page in BioStudies, samples in BioSamples, sequence in ENA, experimental metadata to ArrayExpress. Spatial transcriptomics images in BioImage Archive. Project specific metadata schema in addition to archives checklists.



Linking metadata



SAMEA100007

RestFacetIntegration_testEnaRestFacet

Released on 2015 / 03 / 22 08:30:23 UTC
Created on 2019 / 11 / 27 10:29:24 UTC
Updated on 2019 / 11 / 27 10:29:24 UTC

Download as:

XML
JSON
Bioschemas
Phenopacket

Attributes

Type	Value
Description	Test description
MultiCategoryCodeField	heart and lung
Organism Part	Lung
test_Type	test_value

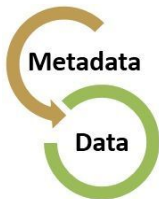
External Links

From: EGA Dataset [DUO:0000001] [DUO:0000005] [DUO:0000007]
ArrayExpress
ENA

External Links

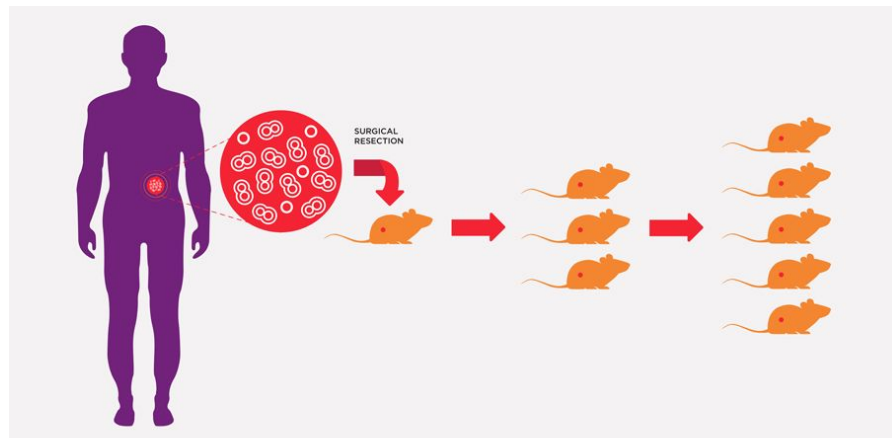
EGA Dataset [DUO:0000001] [DUO:0000005] [DUO:0000007]
ArrayExpress
ENA

Linking samples



Sample relationships

Relationship types	Reverse relationships	Description
<code>derived from</code>	<code>derived from (reverse)</code>	<i>Sample A is derived from Sample B.</i> <i>E.g.</i> <ul style="list-style-type: none">- Tissue samples derived from donor samples- Cell line samples derived from tissue samples- Viral samples separated from saliva samples- Organoid samples cultured from tissue samples
<code>same as</code>	<code>same as</code>	<i>Sample A is the same as Sample B. This can be used to link duplicated samples</i>
<code>has member</code>	<code>has member (reverse)</code>	<i>Sample A is a member of Sample group G. BioSamples create a sample group for each sampleTab submission*. It's also possible to put patient samples as a sample group.</i>
<code>child of</code>	<code>child of (reverse)</code>	<i>Sample A is the child of Sample B.</i> <i>E.g.</i> <ul style="list-style-type: none">- Patient A is the child of Patient B



<https://www.ebi.ac.uk/biosamples/docs/guides/relationships>

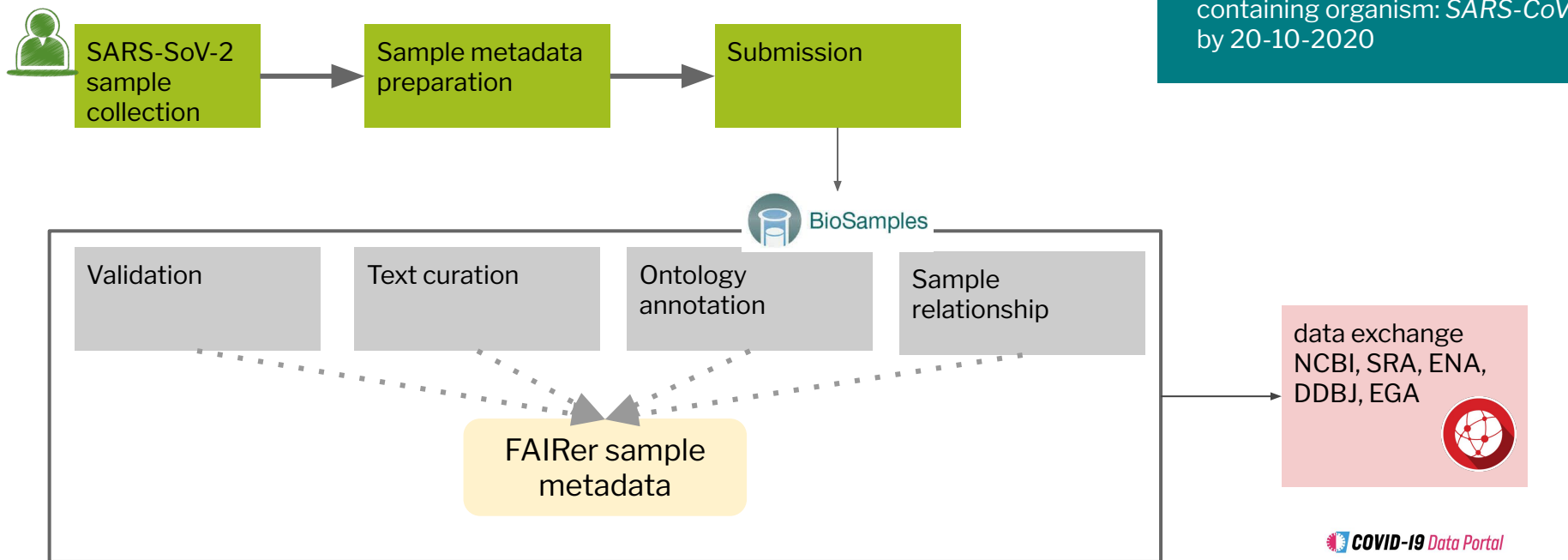
Related patient-derived xenografts(PDX) samples

The journey of a SARS-Cov-2 sample

BioSamples COVID-19

Sample collection:

89068 samples
containing organism: SARS-CoV-2
by 20-10-2020



COVID19 viral sample collection:

https://www.ebi.ac.uk/biosamples/samples?text=NCBITaxon_2697049

 **COVID-19 Data Portal**

[Viral Sequences](#) [Host Sequences](#) [Expression](#) [PI](#)

Viral sequence

A FAIRer SARS-Cov-2 sample

SAMN16272508

CS2217

Attributes

Type	Value
External Id	SAMN16272508
INSDC center name	CSIR-Institute of Genomics ar
INSDC first public	2020-09-26T00:00:00Z
host	Home system
host disease	1 COVID19
isolation source	Respiratory infection swab
lat lon	11.1271 N 78.6569 E
organism	Severe acute respiratory syndrome coronavirus 2
replicate	Biological Replicate 106
strain	SARS-CoV-2
title	Sample106

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

External Links

SRA	ENA
---------------------	---------------------

Download as:

XML
JSON
Bioschemas
Phenopacket

Added values:

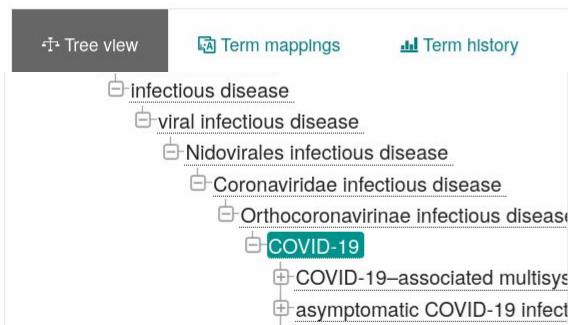
1. Linked to ontology terms and supporting ontology expansion search e.g COVID-19

OLS / Mondo Disease Ontology **MONDO** / **MONDO:0100096**

COVID-19

http://purl.obolibrary.org/obo/MONDO_0100096 [Copy](#)

Synonyms: [2019 novel coronavirus infection](#) [SARS-CoV-2](#) [2019 novel coronavirus](#) [severe acute respiratory syndrome coronavirus 2](#) [coronavirus](#) [SARS-coronavirus 2](#)

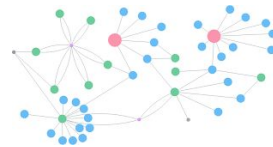


<https://www.ebi.ac.uk/biosamples/samples/SAMN16272508>

EMBL-EBI



A FAIRer SARS-Cov-2 sample



SAMN16272508

CS2217

Download as:

XML
JSON
Bioschemas
Phenopacket

Attributes

Type	Value
External Id	SAMN16272508
INSDC center name	CSIR-Institute of Genomics ar
INSDC first public	2020-09-26T00:00:00Z
host	Human sapiens
host disease	1 COVID19
isolation source	nasopharyngeal swab
lat lon	11.1271 N 78.6569 E
organism	Severe acute respiratory syndrome coronavirus 2
replicate	Biological Replicate 106
strain	SARS-CoV-2
title	Sample106

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

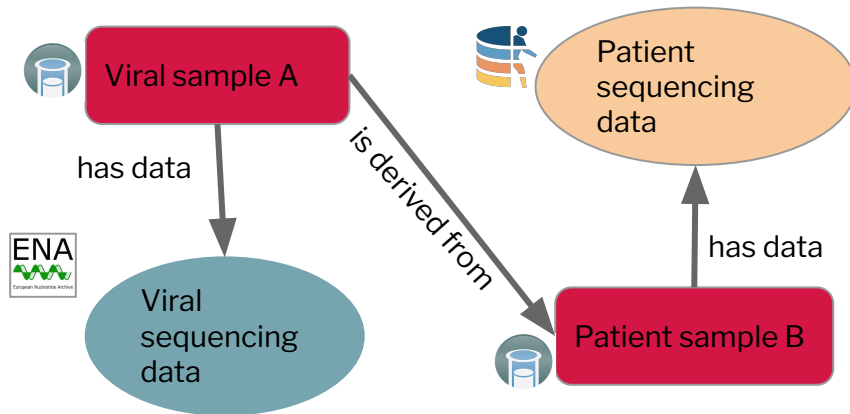
External Links

2

[SRA](#) [ENA](#)

Added values:

1. Linked to ontology terms and supporting ontology expansion search
2. Graph search supporting the explore relationships/interactions between samples and external entities



<https://www.ebi.ac.uk/biosamples/samples/SAMN16272508>

A FAIRer SARS-Cov-2 sample

SAMN16272508

CS2217

Attributes

Type	Value
External Id	SAMN16272508
INSDC center name	CSIR-Institute of Genomics ar
INSDC first public	2020-09-26T00:00:00Z
host	Home system
host disease	1 COVID19
isolation source	Hospitals/Ingenioweb
lat lon	11.1271 N 78.6569 E
organism	Severe acute respiratory syndrome coronavirus 2
replicate	Biological Replicate 106
strain	SARS-CoV-2
title	Sample106

Certificates

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

External Links

SRA	ENA
---------------------	---------------------

Download as:

4

[XML](#)
[JSON](#)
[Bioschemas](#)
[Phenopacket](#)

Added values:

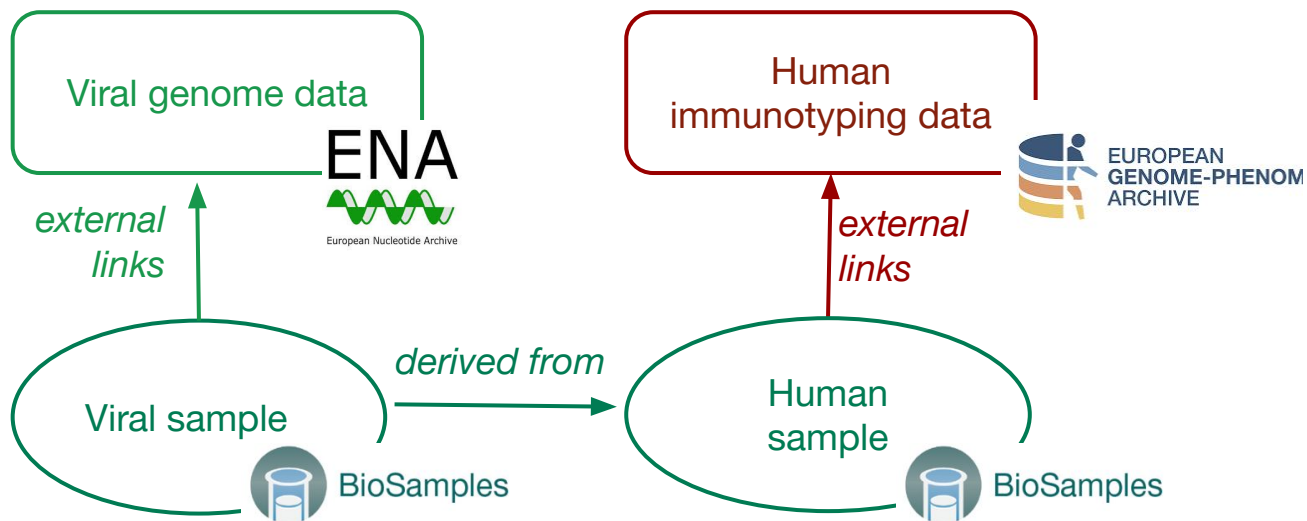
1. Linked to ontology terms and supporting ontology expansion search e.g COVID-19
2. Graph search supporting the explore relationships/interactions between samples and external entities
3. Validated against minimum information requirements
4. Exporting data in different formats

<https://www.ebi.ac.uk/biosamples/samples/SAMN16272508>

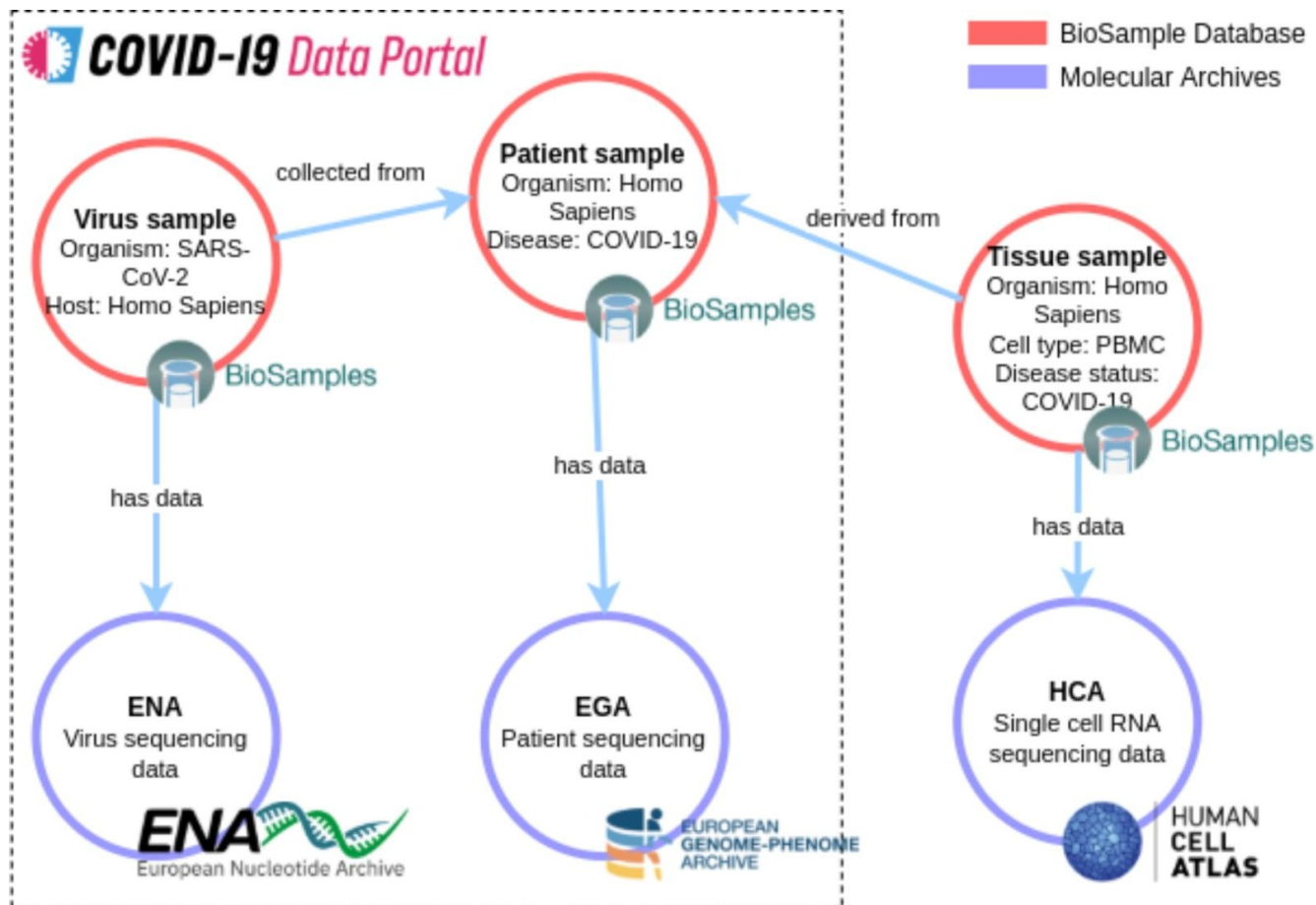
Search across archives



As a researcher, I want to find the immunotyping data of all lung samples from **COVID19 patients** and corresponding genome sequencing data of the **viral isolate**, to study how the immune systems response to viral infection.



Covid-19 Sample Linking



Final thoughts



“Historical” Data Workflows



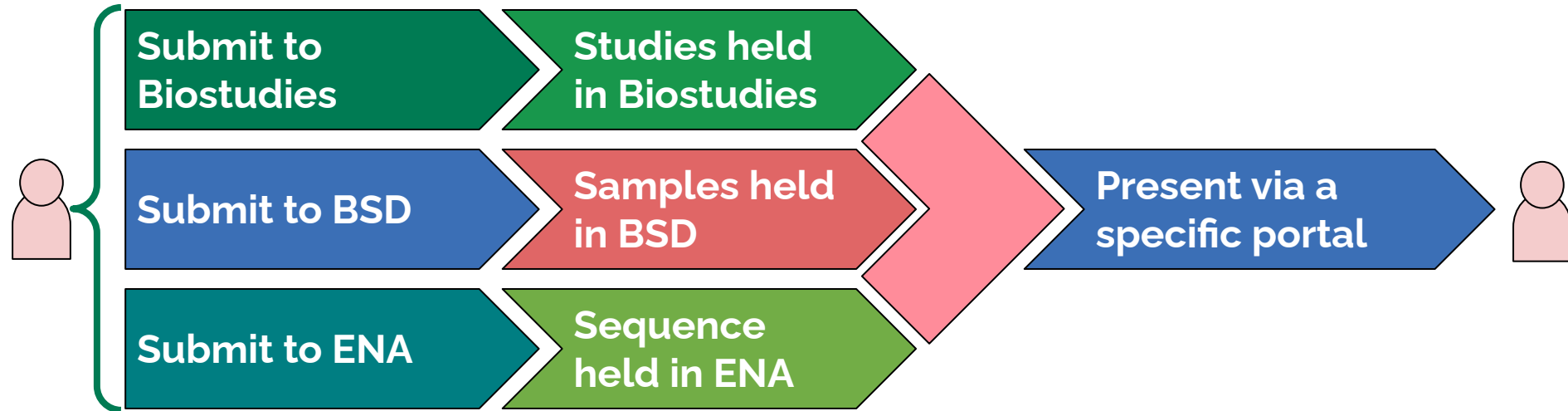
- Submitters interact with “target” archive (ENA)
- All data held in a single archive (e.g. ENA)
- Submissions systems, standards, storage, user interfaces all siloed
- Consumers interact with “branded” archive (ENA)

“Brokering” Data Workflows



- Submitters interact with “target” archive (ENA); samples are brokered
- Data held in two independent archives (ENA and BioSamples)
- Standards are aligned; storage systems are specialised; submissions systems and user interfaces remain siloed
- Consumers interact with “branded” archive (ENA); sample indexing is federated

“Coordinated” Data Workflows into the Future



- Submitters - and their initiatives - can establish more direct control over standards
- Submitters *are* required to do some specialist data coordination
- Data and metadata held in specialised systems and interlinked (Biostudies, BioSamples, ENA)
- Standards are aligned; submission, storage systems and consumer interfaces are all specialised

BioSamples Looking Forward



BioSamples as a data broker

- Single repository for samples metadata
- Connectivity between archives
 - Embedding sample ID generation into submission pipelines
 - Linking at submission time
 - Importing inter-archive links

Refined at annual “Samples Day” Events



BioSamples supporting the communities

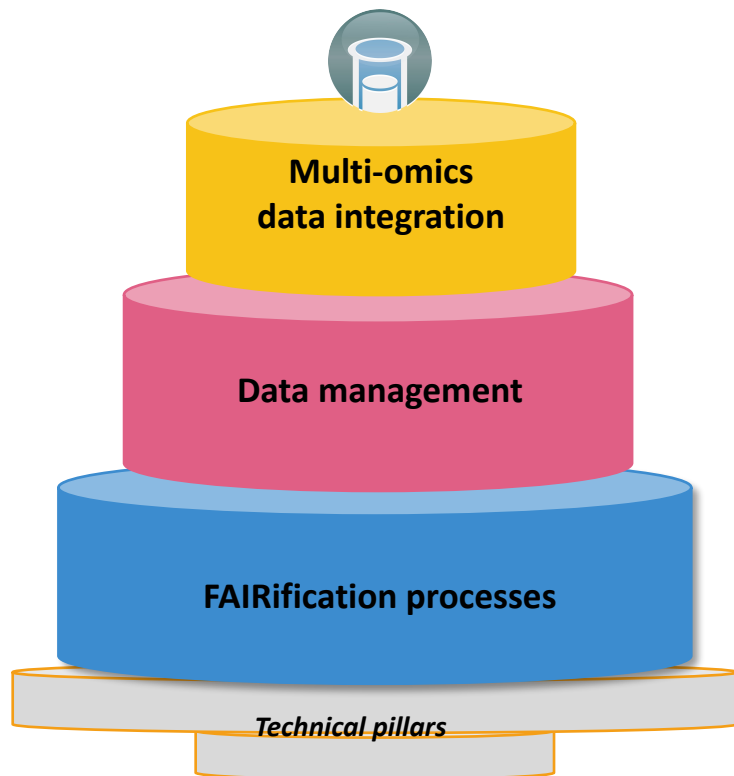
- Search: Graph search, case insensitive search /filtering, ontology based search, advanced search/filtering UI
- Curation: Curation provenance model
- Checklists: Checklist versioning



Technical infrastructure for better findability, availability and reliability

- Solr upgrade
- Service partitioning
- Defining SLO and error budgets for critical APIs
- Improve API monitoring

Biosamples “Layer Cake”



Interconnectivity across archives

- Single hosting place for sample metadata
 - Linking sample, assay and publication

Data deposition for collaborative projects

- Easier submission process
- Adjust to community requirements

Validation and semi-automatic curation

- Sharable and reusable checklists for consistent metadata representation
- Supporting community curation

Acknowledgements

BioSamples team, past and present



Fuqi Xu



Dipayan Gupta



Isuru Liyanage



Cincia Thion



Melanie Courtot

Partners



Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

Archive Infrastructure Technology
team
FAIRplus collaborators
GA4GH collaborators

Funding



EMBL-EBI



innovative
medicines
initiative



Questions?

