

Data clearinghouse, validation and curation of BioSamples, ENA, Breeding API endpoints, MAR databases

Proponent: **Luca Cherubin, Cyril Pommier**

On the behalf of the Elixir Validation Implementation study group

Project links

- [GitHub project](#)

Background

- Regular practice in life-sciences is for domain experts to manually curate and improve the quality of metadata associated with biological sample(s)
- Unfortunately this curated, high quality metadata often can't be embedded back into the original assay and results, reducing data FAIRness.

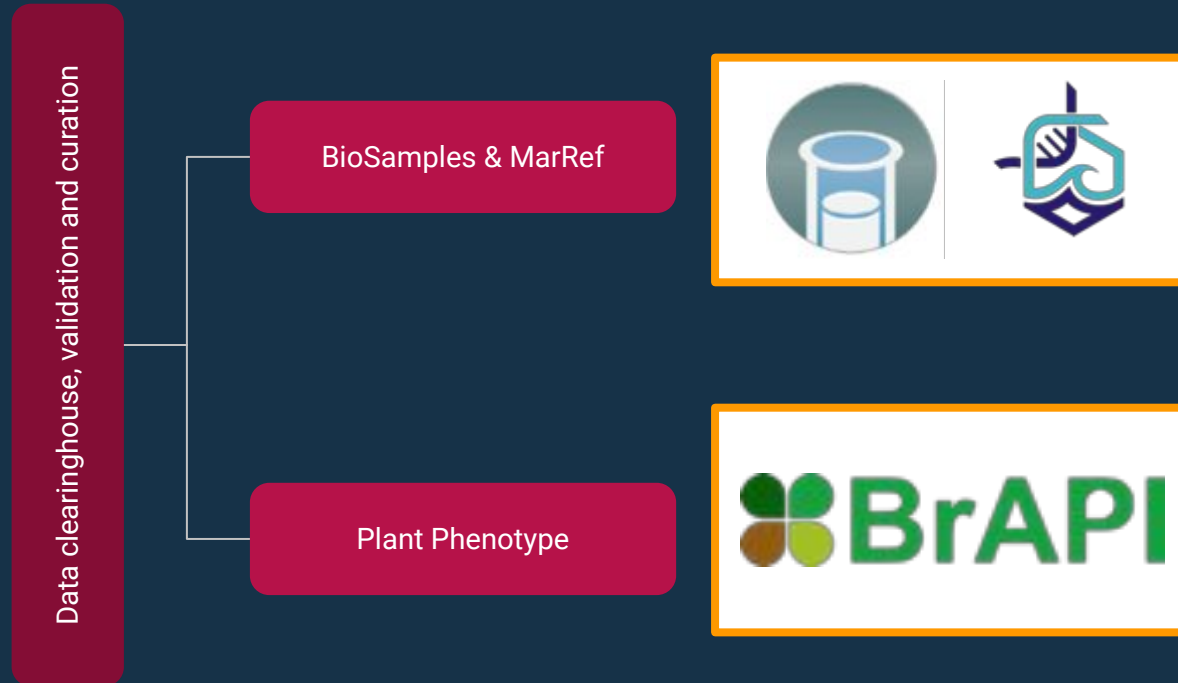


Goal and expected outcome

- **Expose metadata for programmatic access**
- **Programmatically validate the metadata against predefined schemas**
- **Store metadata in a central repository easily accessible from any resource**

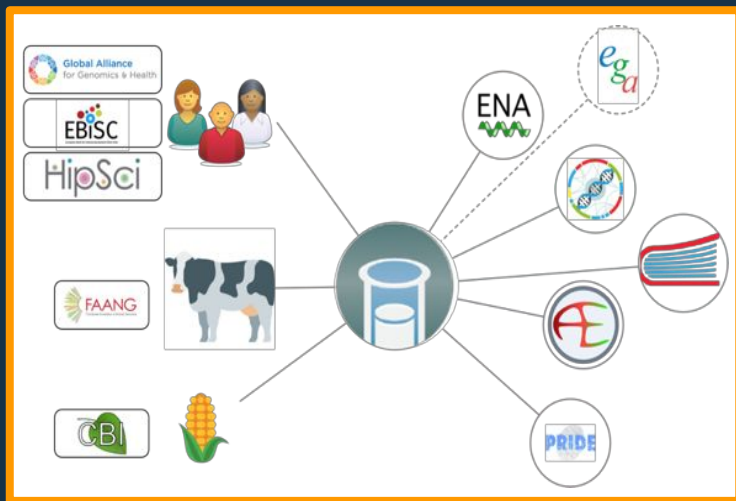


Use cases



BioSample, MarRef Use Case

- Biosamples: EMBL-EBI hub for sample metadata



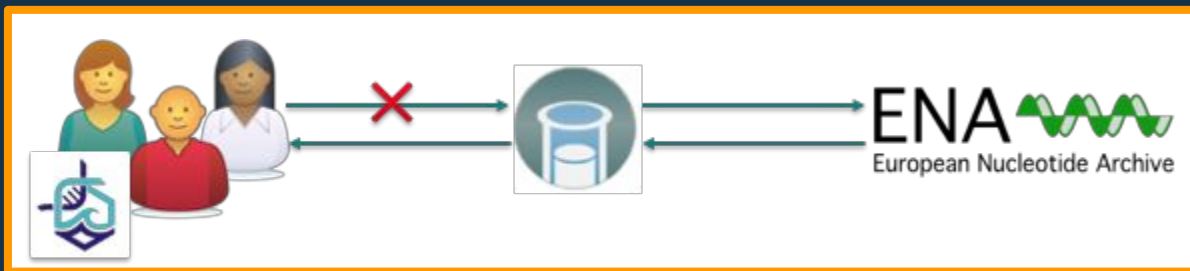
- Marine Metagenomic Portal: high-quality curated and freely accessible microbial genomics and metagenomics resources

The screenshot shows the Marine Metagenomic Portal (MarRef) interface. The top section displays the MarRef logo and a brief description: "MarRef is a manually curated marine microbial reference genome database that contains complete genomes. Each entry contains full metadata, including information about the environment, host, origin and taxonomy, phenotype, pathogenicity, assembly and annotation. The current version contains 1,000 genomes." Below this, there is a search bar and a table of results. The table lists various microbial genomes, including *Thermococcus gammatolerans* E13. The table columns include Name, Accession, and Date. The bottom section of the interface features a world map showing the locations of the sampled genomes.

Name	Accession	Date
<i>Thermococcus gammatolerans</i> E13	MG100000	2013-01-01
<i>Thermococcus gammatolerans</i> E13	MG100000	2013-01-01
<i>Thermococcus gammatolerans</i> E13	MG100000	2013-01-01

BioSample, Marine Metadata Use Case

- BioSamples stores sample metadata accessible to anybody
- Curators, like MarRef, re annotate the metadata based on literature and manual curation
- BioSamples, as well as associated services, will not be able to access that data easily and make sure the data is valid



MarRef compared to BioSamples

Lactococcus garvieae ATCC 49156	
+ Expand all - Collapse all	
Summary	
MMR ID	MMR0235969
Full Scientific Name	Lactococcus garvieae ATCC 49156
Strain	YT-3, ATCC 49156, DSM 6089, NCMB 13208
Type Strain	No
Geographic location	Japan
Collection Date	1976
Biosample Accession	SAMN0456969
Biocore ID	14881
Culture Collection(s)	ATCC 49156, DSM 6089, NCMB 13208
Isolation Country	Japan
Environmental Package	Host-associated
Isolation Source	Kidney
Host Scientific Name	Seniole guineensis
Curation Date	2016-05-02
Updated Date	missing
Implementation Date	2016-10-22
Microbe Package	not applicable
Experiment/Investigation Type	Bacteria
Bioproject Accession	PRJNA39896
Genbank Accession	AF002331.1
NCBI Taxon Identifier	420889
Source for Biomaterial	SRJ2713
Accession SSU	191896
Accession LSU	191896
16S Accession	U700008.2
Annotations	1010956
Comments	Previous name was Enterococcus seniole




Lactococcus garvieae ATCC 49156

Expand all Collapse all

Summary

MMF ID	MMF023569
Full Scientific Name	Lactococcus garvieae ATCC 49156
Strain	YT-3, ATCC 49156, DSM 6089, NCMB 13208
Type Strain	No
Geographic location	Japan
Collection Date	1976
Biosample Accession	SAMN02596969
Barcode ID	14882
Culture Collection(s)	ATCC 49156, DSM 6089, NCMB 13208
Isolation Country	Japan
Environmental Package	Host-associated
Isolation Source	Kidney
Host Scientific Name	Seriola quinqueradiata
Curation Date	2016-05-02
Updated Date	missing
Implementation Date	2016-10-22
Microbe Package	not applicable
Experiment/Investigation Type	Bacteria
Bioproject Accession	PRJNA392896
Genbank Accession	AF002332.1
NCBI Taxon Identifier	420889
Resource for Biomaterial	1882703
Accession SSU	192896
Accession LSU	192896
16S Accession	U00008822
16S Accession	192896
Comments	Previous name was Enterococcus sakaiioides


BioSamples

[Home](#)
[Search](#)
[Submit](#)
[Help](#)
[About](#)

SAMN02596969

Sample from Lactococcus garvieae ATCC 49156

Release

2016/05/10 23:00:00 UTC

Update

2016/10/22 13:25:42 UTC





Attributes

Type	Value
Organism	Lactococcus garvieae ATCC 49156
Culture collection	ATCC 49156
Model	Genetic
Package	Genetic 1.0
Strain	ATCC 49156
Synonym	ABMT_49156

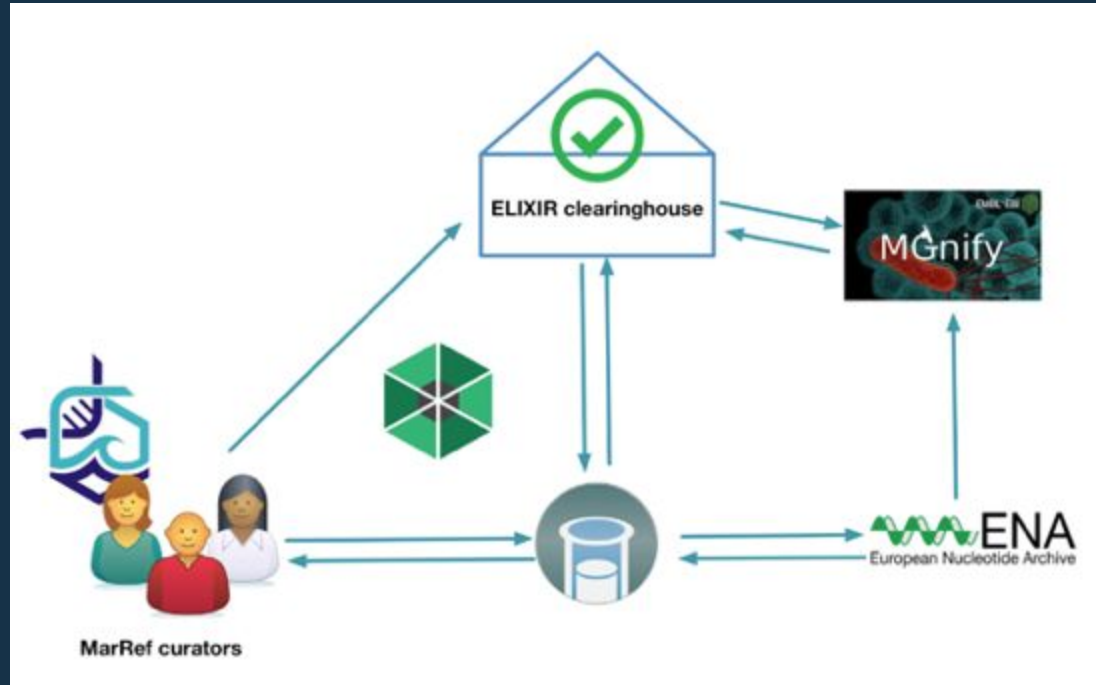
BioSamples Database is part of the ELIXIR infrastructure
[Accession is part of the ELIXIR infrastructure. Accession: 192896](#)



What we want to build

	Integrate BioSchemas into MarRef and extract the metadata using the BioSchemas crawler	<ul style="list-style-type: none">•Python•Go	Day 1
	Produce a JSON schema to validate the data	<ul style="list-style-type: none">•JSON schemas•Validation•Ontology	Day 1 – Day 2
	Feedback between BioSamples and MarRef	<ul style="list-style-type: none">•JSON API•Python•Java	Day 2 - Day 3
	Work on a repository for exported/validated metadata – Elixir clearinghouse	<ul style="list-style-type: none">•MongoDB•Java•Python	Day 3 – Day 4

Workflow



Plant Use Case

Elixir Plant data search: community data portal

Filters 🔍

sources

- ☐ **URGI GnpIS** (67878)
- ☐ **CIRAD TropGENE** (727)
- ☐ **VIB PIPPA** (679)
- ☐ **IBET BioData** (61)
- ☐ **NIB** (8)

types

- ☐ **Germplasm** (68392)
- ☐ **Phenotyping Study** (961)

URGI 🏠

GnpIS 🌐

Home / Global Search

Taxon / Germplasm

Phenotyping

Polymorphism

Association

SEARCH FORM

Germplasm **Variable**

Crop
(common name, species, genus, subtaxa & synonyms)

Germplasm list
(panel, collection & population)

Accession
(accession name, number & synonyms)

* Only the first terms corresponding to the current search are suggested

RESULTS

Germplasm **CIRAD TropGENE** **AUGUSTO**
"Oryza sativa L." is a Oryza sativa L. (rice) accession (number:)

Germplasm **IBET BioData** **SWARNA**
"SWARNA" is a Oryza sativa (Rice) accession (number: "IRGC")

Phenotyping Study **NIB** **Proteomics**
"Proteomics" is a phenotyping study conducted from 2010-01-01 to

Phenotypes

Winter wheat (*Triticum aestivum* L) phenotypic data from the multiannual, multilocal field trials of the INRA Small Grain Cereals Network.

François-Xavier Oury, Emmanuel Heumez, Bernard Rolland, Jérôme Auzanneau, Pierre Bérard, Maryse Brancourt-Huimel, Xavier Charrier, Hubert Chiron, Camille Depatureaux, Laurent Falchetto, Olivier Gardet, Stéphane Gilles, Alex Giraud, Christophe Lecomte, Jean-Yves Morlais, Pierre Pluchard, Didier Tropée, Maxime Trotet, Patrice Walczak, Gérard Doussinault, Michel Rousset, Gilles Charmet

[Query dataset as a semantic graph.](#)

[Or download the dataset as RDF archive.](#)

[Abstract](#)

Published 2015 by INRA

[Back to Form](#)

[Search parameter\(s\):](#)

Geolocation

DATA SETS: 4

Network Data Set :
[INRA Wheat Network BRC accession \(A series\)](#)

Network Data Set :
[INRA Small Grain Cereals Network](#)
DOI:<http://dx.doi.org/10.15454/1.4489666216568333E12>

Network Data Set :
[INRA Wheat Network not BRC accession \(B and C series\)](#)



Origin site Collecting site Evaluation site

Phenotyping campaign(s)

2000 × 2001 × 2002 × 2003 × 2004 × 2005 × 2006 × 2007 × 2008 ×
2009 × 2010 × 2011 × 2012 × 2013 × 2014 × 2015 ×

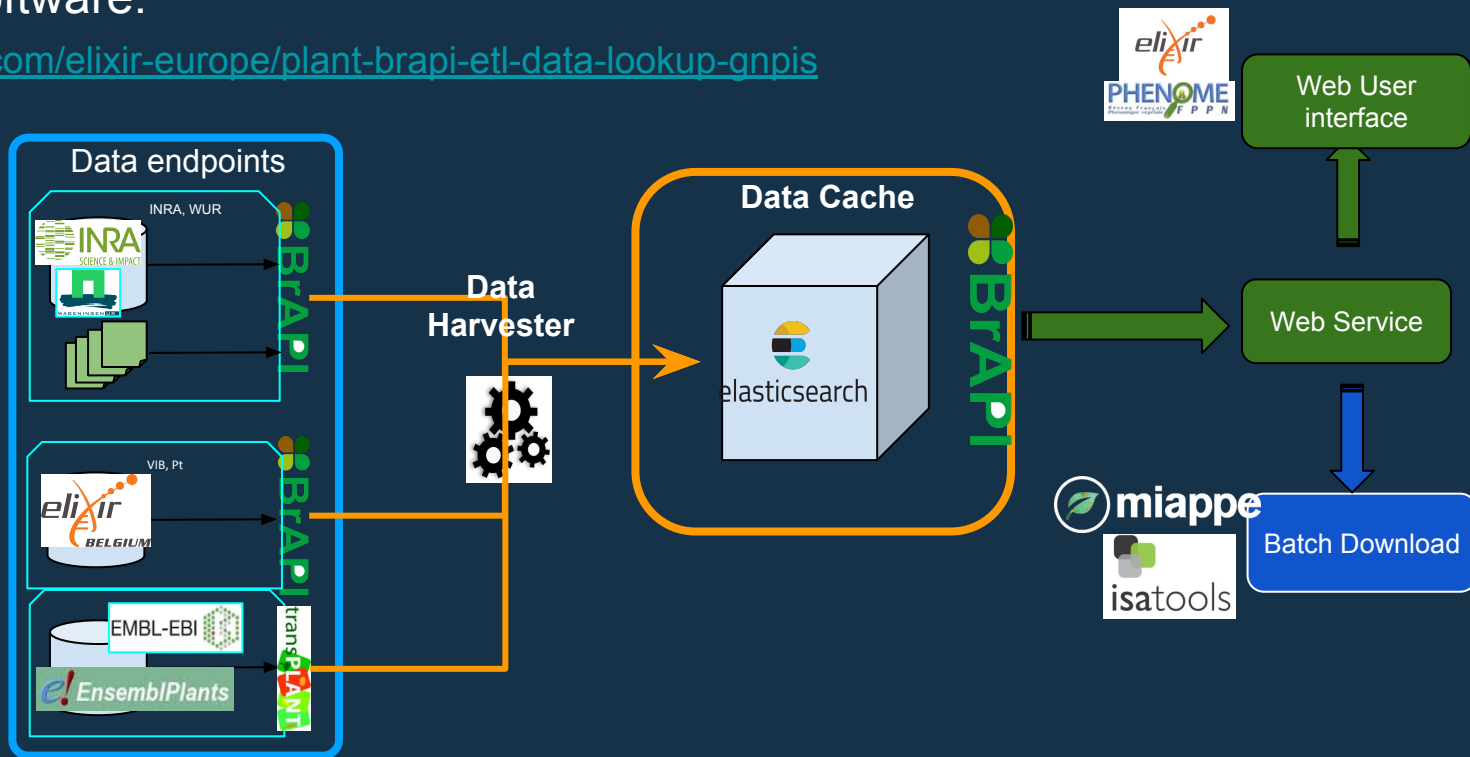
[remove all](#) [add all](#)

[Trial list](#)

[Phenotypic data](#)

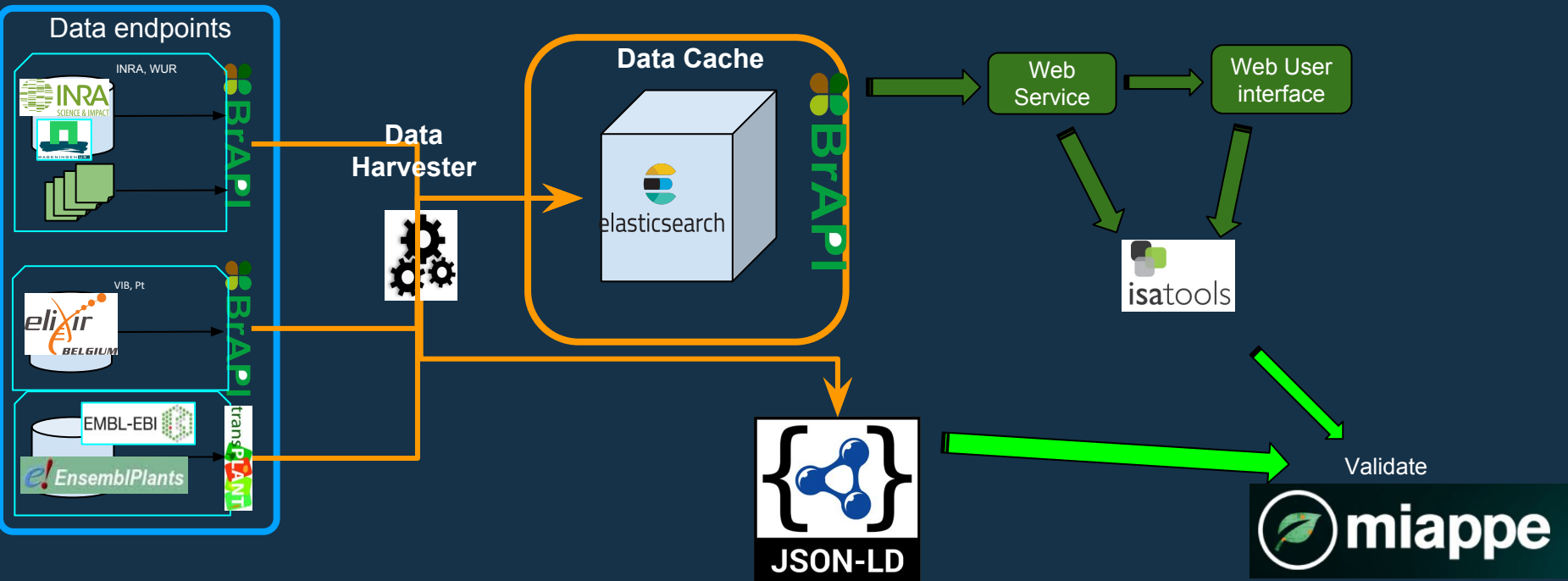
Elixir Plant data search: distributed system

- Phenotype through Breeding API
- Generic WheatIS/transPlant for all other data types
- Open source software.
 - <https://github.com/elixir-europe/plant-brapi-etl-data-lookup-gnpis>


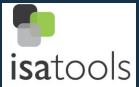




Elixir Plant data search: Hackathon objectives

- Endpoints need to be validated
- Datasets must really comply to MIAPPE BrAPI specifications
- Datasets must be made available as ISA-Tab for further analysis



What we want to build

 <p>JSON-LD</p>	<p>BrAPI v1.2 JSON-LD context BrAPI v1.2 JSON Schemas BrAPI 2 JSON-LD</p>	<ul style="list-style-type: none"> •Python •JSON schemas 	<p>Day 1 - Day2</p>
 <p>isatools</p>	<p>Validate BrAPI2ISA on all uses cases</p> <ul style="list-style-type: none"> - Single experiment - Phenotyping network - Perennial plants 	<ul style="list-style-type: none"> •Java •REST API 	<p>Day 2 – Day 3</p>
 <p>BrAPI</p>	<p>Integrate BrAPI 2 ISA as a service</p>	<ul style="list-style-type: none"> •JSON API •Python •Java 	<p>Day 3 - Day 4</p>
 <p>miappe</p>	<p>Validate datasets</p> <ul style="list-style-type: none"> - Ontologies & JSON-LD - JSON-Schemas - ISA framework 	<ul style="list-style-type: none"> •Ontologies •Java •Python 	<p>Day 3 – Day 4</p>

Post-biohackathon perspectives

- Integration of BrAPI 2 ISA in Elixir Plant Data Search
- Data validator proposed to the whole Plant community
- Semantic capabilities on BrAPI endpoints
- [...]

We want you

- Developers interested in Bioschemas applications
- Developers with knowledge on any of JavaScript, Java, GO, Python, data indexing tools...
- Developers with knowledge on MongoDB, JSON and JSON Schema
- Data resource developers or owners
- Curators or data validators
- Ontologists



Acknowledgements

