BioHackathon 2018 - Paris
http://bh2018paris.info/

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 information

Background

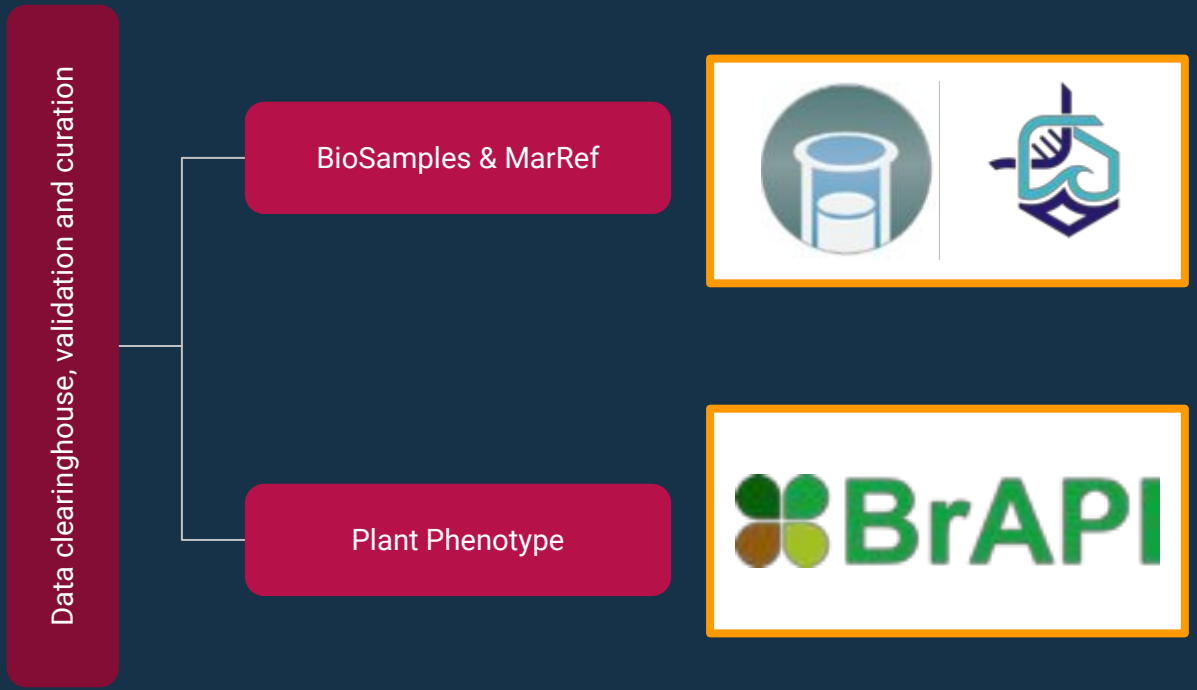
- Regular practice in life-sciences is for domain experts to manually curate and improve the quality of metadata associated with biological sample(s)
- Unfortunately this curated, high quality metadata often can't be embedded back into the original assay and results, reducing data FAIRness.

Goals of the hacking project

Goal and expected outcome

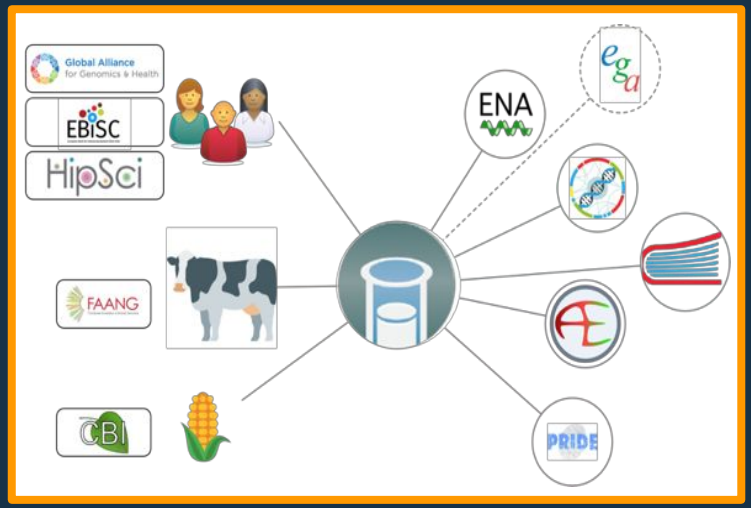
- **Expose metadata for programmatic access**
- **Programmatically validate the metadata against predefined schemas**
- **Store metadata in a central repository easily accessible from any resource**

Use cases

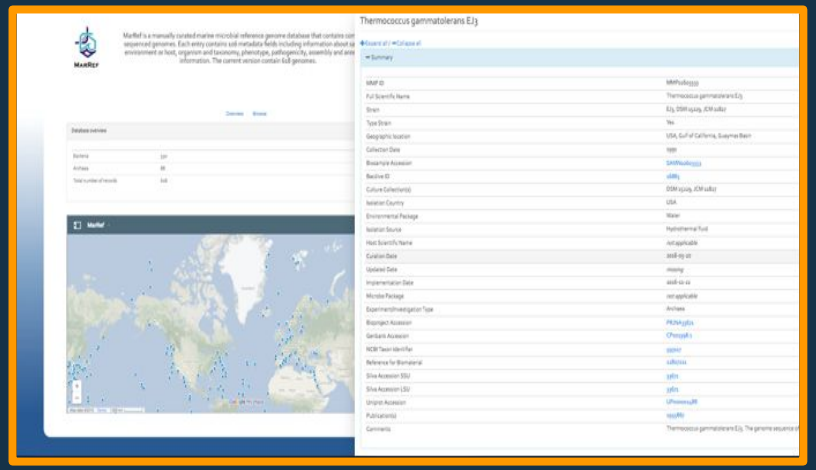

Data clearinghouse, validation and curation
BioSamples & MarRef
Plant Phenotype
BrAPI

BioSample, MarRef Use Case

- Biosamples: EMBL-EBI hub for sample metadata

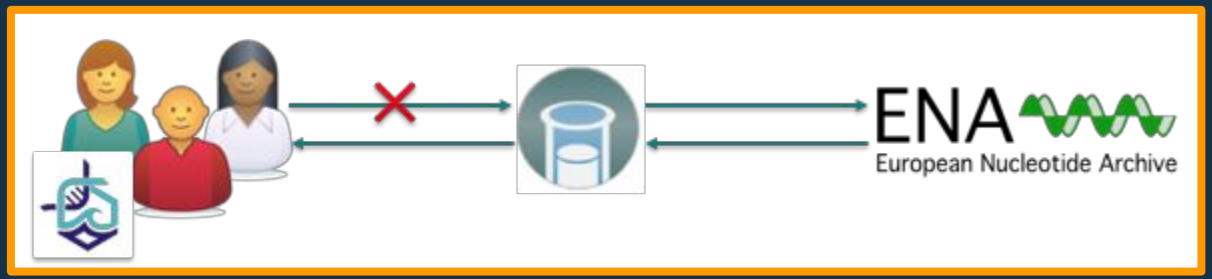


- Marine Metagenomic Portal: high-quality curated and freely accessible microbial genomics and metagenomics resources

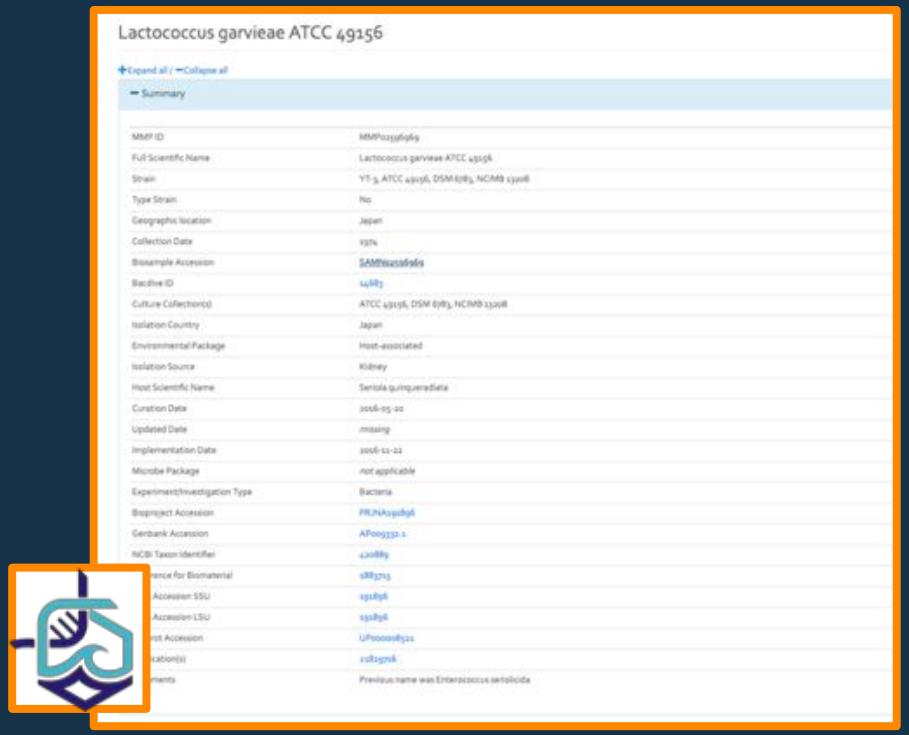


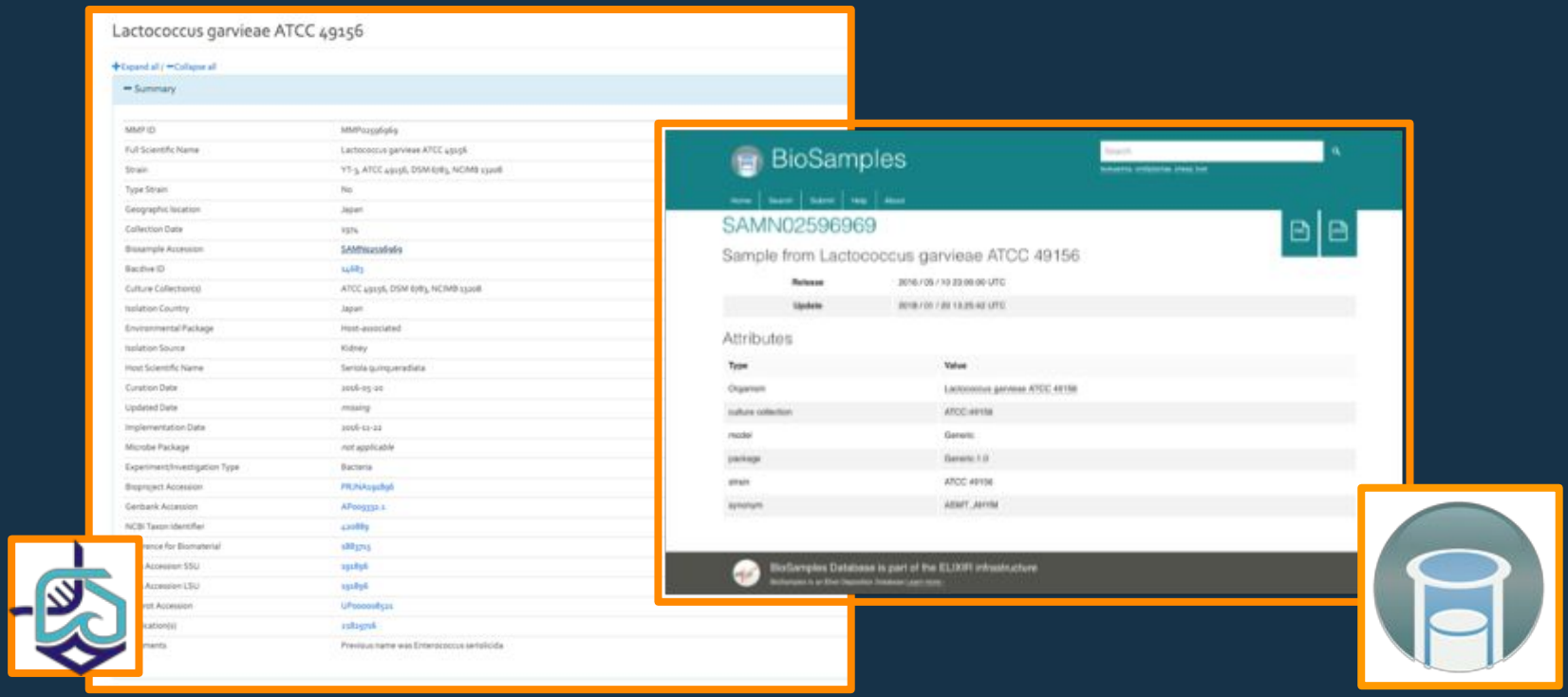
BioSample, Marine Metadata Use Case

- 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

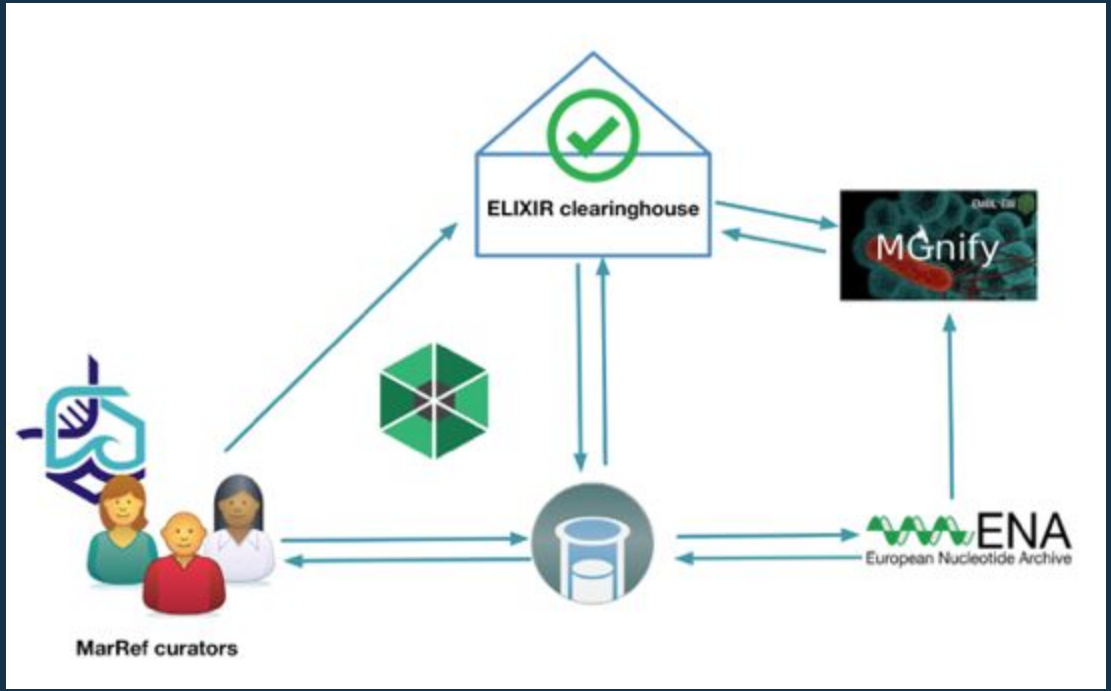



Lactococcus garvieae ATCC 49156
Expand all / Collapse all
Summary
BioSamples
SAMN02596969
Sample from Lactococcus garvieae ATCC 49156
Attributes
Type
Value
BioSamples Database is part of the ELIXIR infrastructure

What we want to build

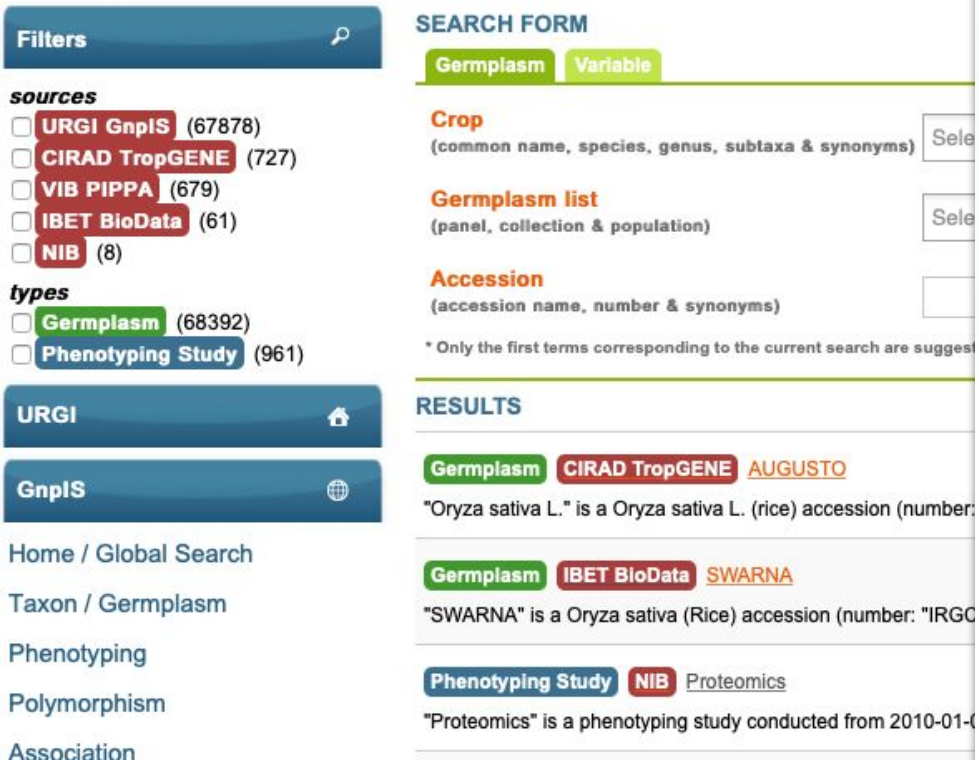
	Integrate BioSchemas into MarRef and extract the metadata using the BioSchemas crawler	•Python •Go	Day 1
	Produce a JSON schema to validate the data	•JSON schemas •Validation •Ontology	Day 1 – Day 2
	Feedback between BioSamples and MarRef	•JSON API •Python •Java	Day 2 - Day 3
	Work on a repository for exported/validated metadata – Elixir clearinghouse	•MongoDB •Java •Python	Day 3 – Day 4

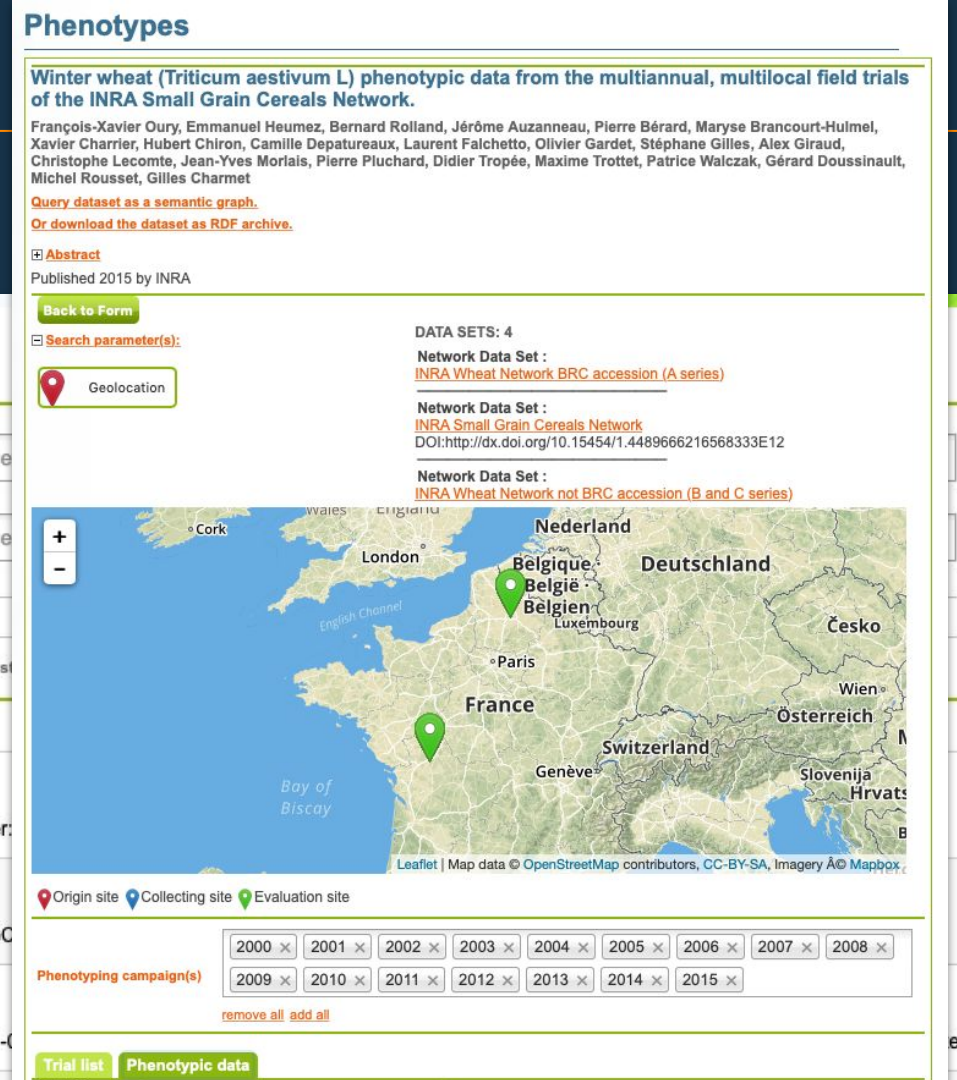
Workflow


ELIXIR clearinghouse
MGnify
ENA
European Nucleotide Archive
MarRef curators

Plant Use Case

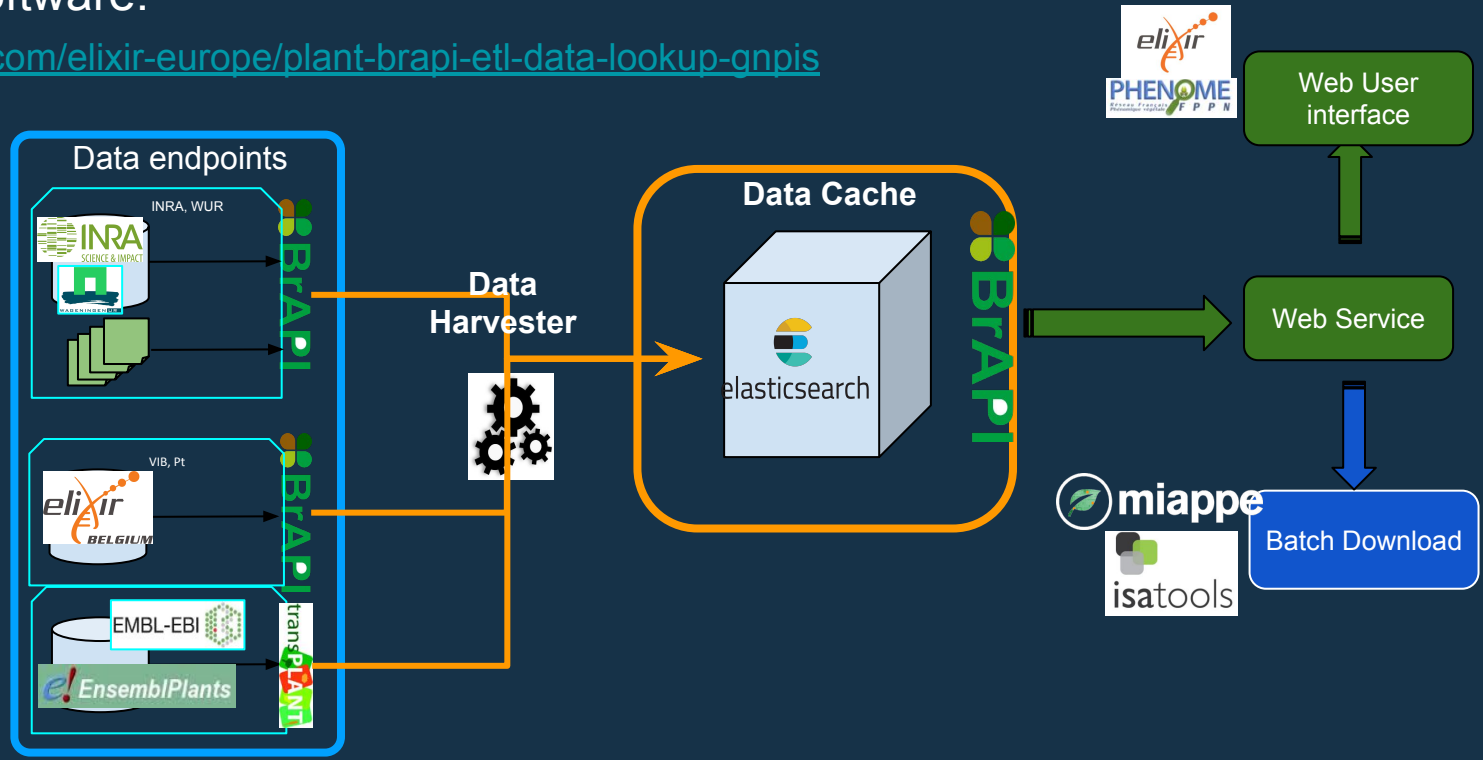
Elixir Plant data search: community data portal





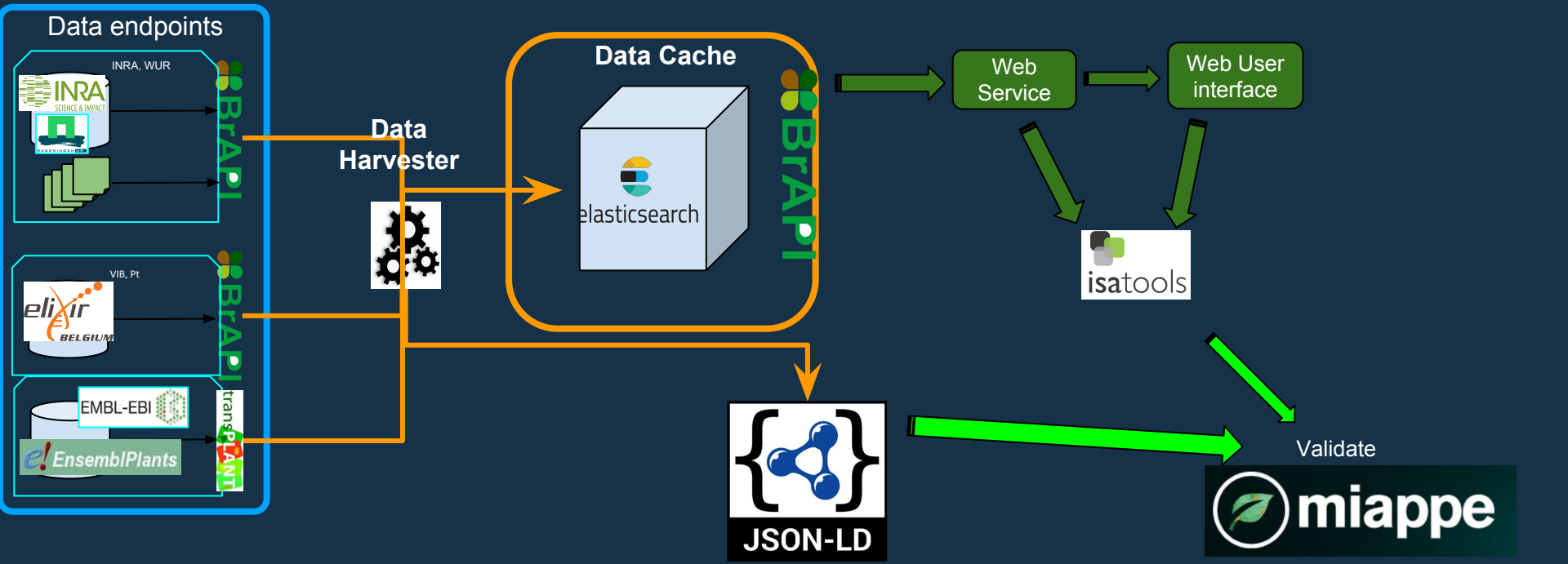
Elixir Plant data search: distributed system

- 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

	BrAPI v1.2 JSON-LD context BrAPI v1.2 JSON Schemas BrAPI 2 JSON-LD	•Python •JSON schemas	Day 1 - Day2
	Validate BrAPi2ISA on all uses cases - Single experiment - Phenotyping network - Perenial plants	•Java •REST API	Day 2 – Day 3
	Integrate BrAPi 2 ISA as a service	•JSON API •Python •Java	Day 3 - Day 4
	Validate datasets - Ontologies & JSON-LD - JSON-Schemas - ISA framework	•Ontologies •Java •Python	Day 3 – Day 4

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