Institut de recherche pour le développement

Institut de biologie computationnelle

Ontology-based data management: lessons learned for ABD monitoring systems

Pierre Larmande
Institute of Computationnal Biology IRD
Pierre.larmande@ird.fr

Outline

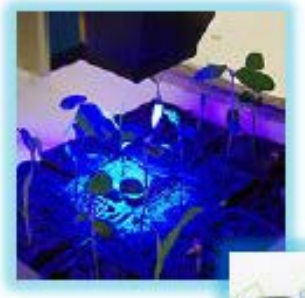
- Data Landscape
- Data integration challenges in the Life Sciences
- Ontologies/ Semantic Web Technologies
- Some example applications

Data landscape in the Life Sciences

- The availability of biological data has increased
- Advancements in:
  - computational biology
  - genome sequencing
  - high-throughput technologies
- Integrative approaches are necessary to understand the functioning of biological systems

Plant phenotyping

- Act of determining the quantitative or qualitative values of trait(s) of interest at any organisational level, in a given genomic expression state (MAGIC, NAM, RILs population) and a given environment
- Performed by plant phenotyping systems (from the manual measurement of leaf length to complex robotic systems with automated acquisition and measurement workflows)



Molecular/Cell

Cell/Organ

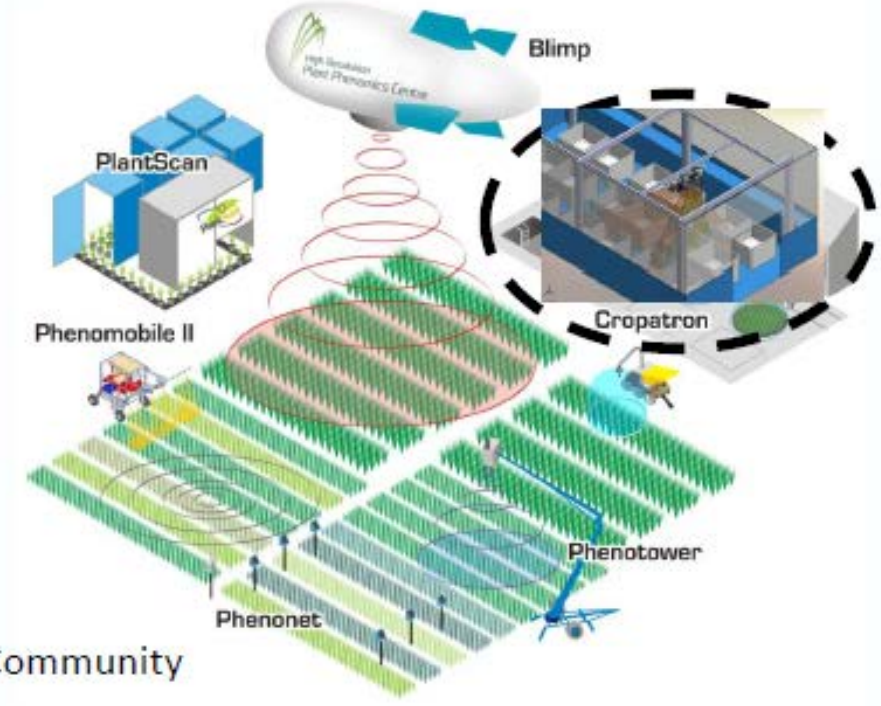
Organ/Plant

Plant/Canopy

Canopy/Community

Phenotyping performed at **multiple organisational levels**

Phenotyping system defined the **temporal and spatial resolutions**

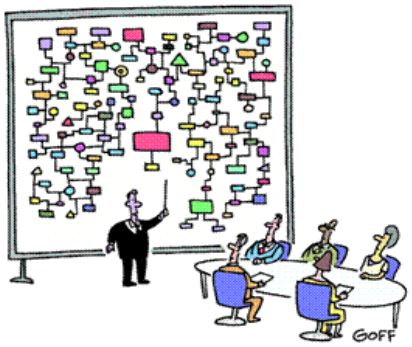


Which resources to use?

- Traditional information system
- Remote sensing networks/systems
- Social networks
- Scientific documents/publications
- …

Data integration challenges

- Lack of effective approaches to integrate data that has created a gap between data and knowledge
- Need for an effective method to bridge gap between data and underlying meaning
- Harvest the power of overlaying different data sets



"And that's why we need a computer."

Today's Web

- Today's Web content is suitable for human consumption

- Collection of documents
  - the existence of links that establish connections between documents

- Low on data interoperability and lacks semantics.

Standardization of data

- Drastic increase in data production.
- Standardization needed to manage and use these data
- Mainly used XML for standardizing data exchange.
  - SBML, CellML
- **M**inimum **I**nformation for **B**iological and **B**iomedical **I**nvestigations (MIBBI)
- **I**nvestigation, **S**tudy, and **A**ssay (ISA)

Ontologies

- Ontologies are formal representations of knowledge - definitions of concepts, their attributes and relations between them.

- To integrate data, improve machine interoperability and data analysis required a conceptual scaffold.

- Ontological terms used across databases
  - provide cross-domain common entry points in the description.
  - use to bring structured integration of various datasets.

Building an ontology

- Knowledge experts
- Documentations/research papers ….
- Reusing existing ontologies

Collaborative

NLP & text mining

Matching & NLP

Ontologies

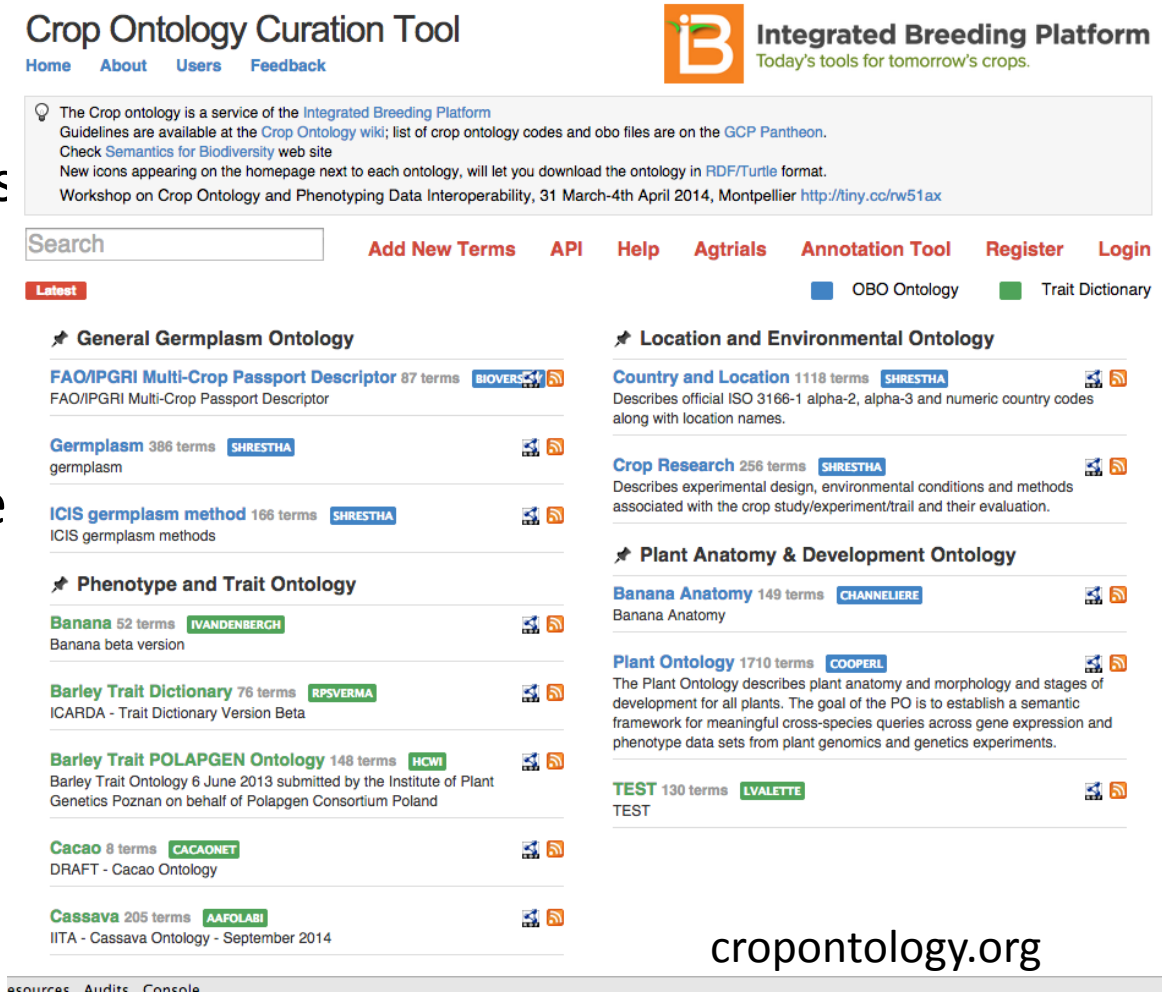
- The **O**pen **B**iomedical **O**ntologies **(OBO)** initiative:
  - serves as an umbrella for well structured orthogonal ontologies.
  - Ontologies represented in OBO format and OWL

OBO
open biomedical ontologies

Is a three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions.

Crop Ontology

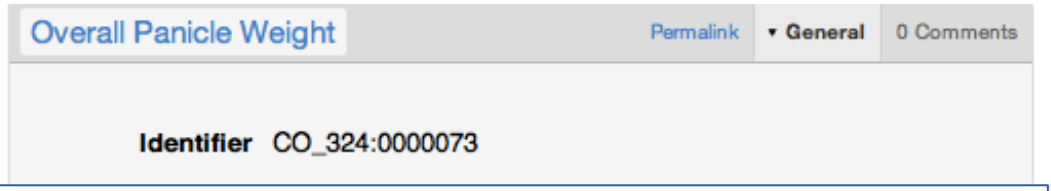
- Is an application ontology for fielbooks and breeding databases & repositories
- A visualization tool supporting curation of trait lists by a distribute community
- A discussion Forum

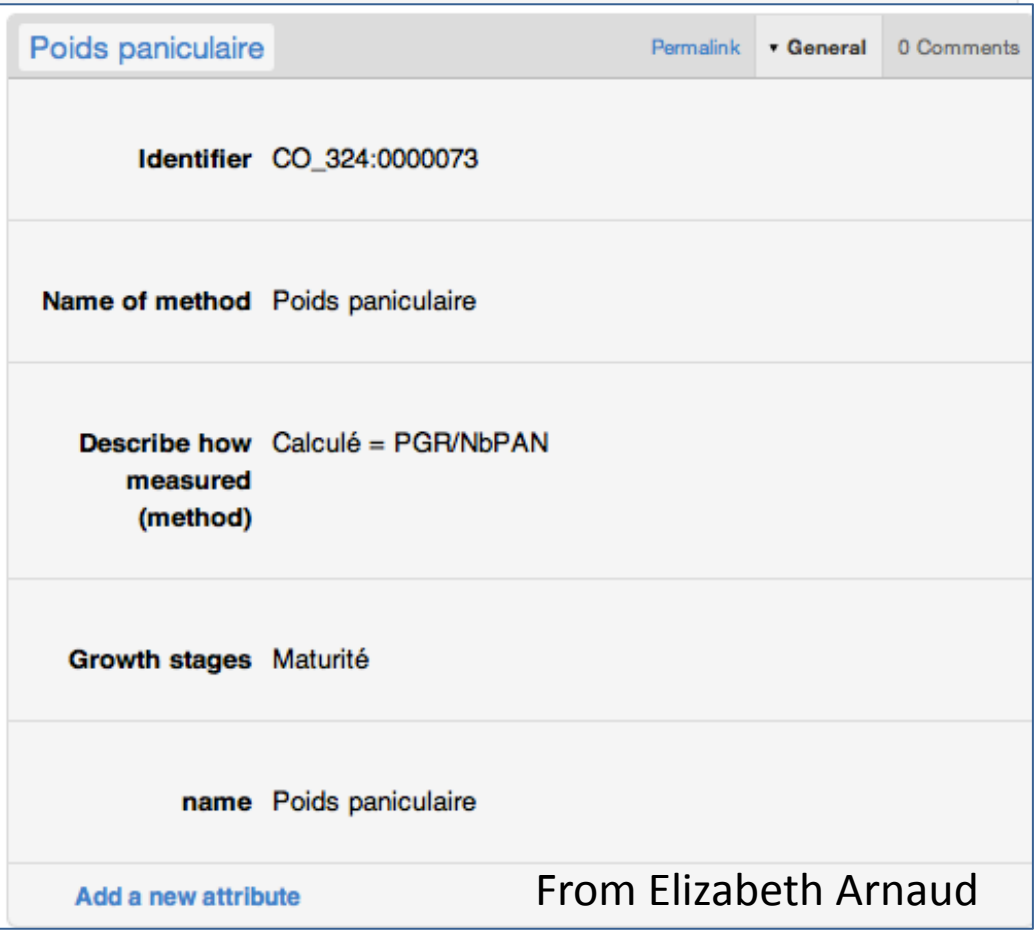


cropontology.org

From Elizabeth Arnaud

Multilingual ontologies





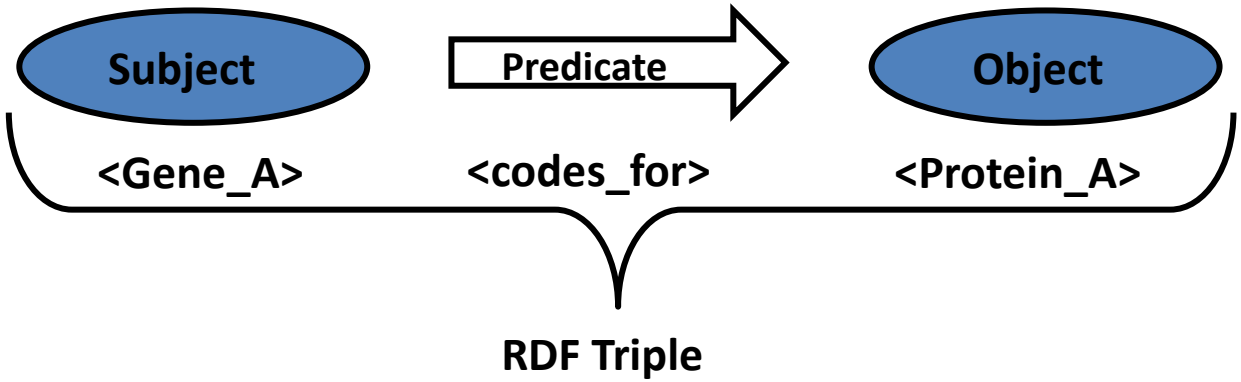
From Elizabeth Arnaud

Semantic Web Technology

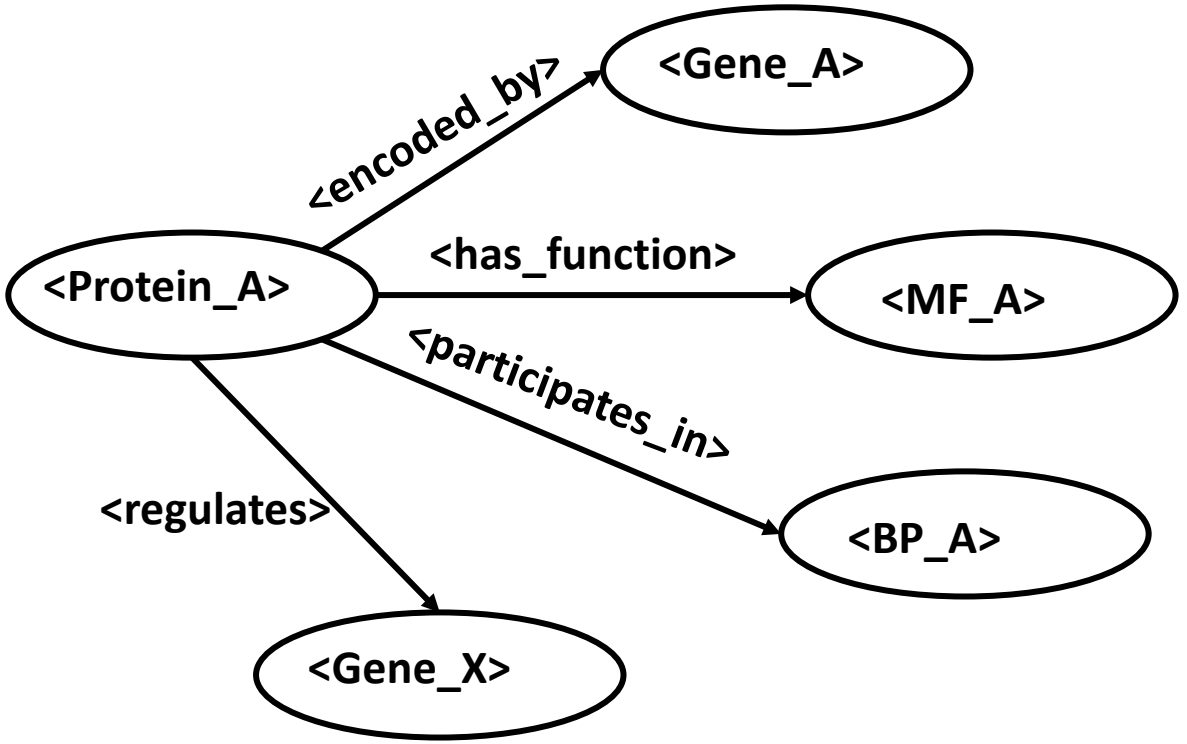
- An extension of the current Web technologies.
- Enables navigation and meaningful use of digital resources.
- Support aggregation and integration of information from diverse sources.
- Based on common and standard formats.

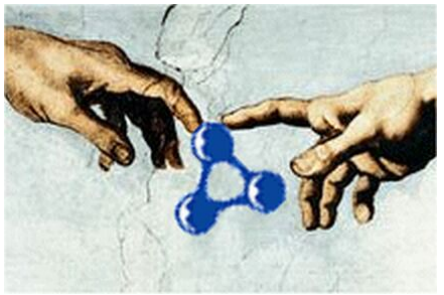
Resource Description Framework (RDF)

- Framework for representing information about resources on the Web
- Provides a labeled connection between two resources
- Uses Unique Resource Identifiers (URI)
- Statements take the form of triples:



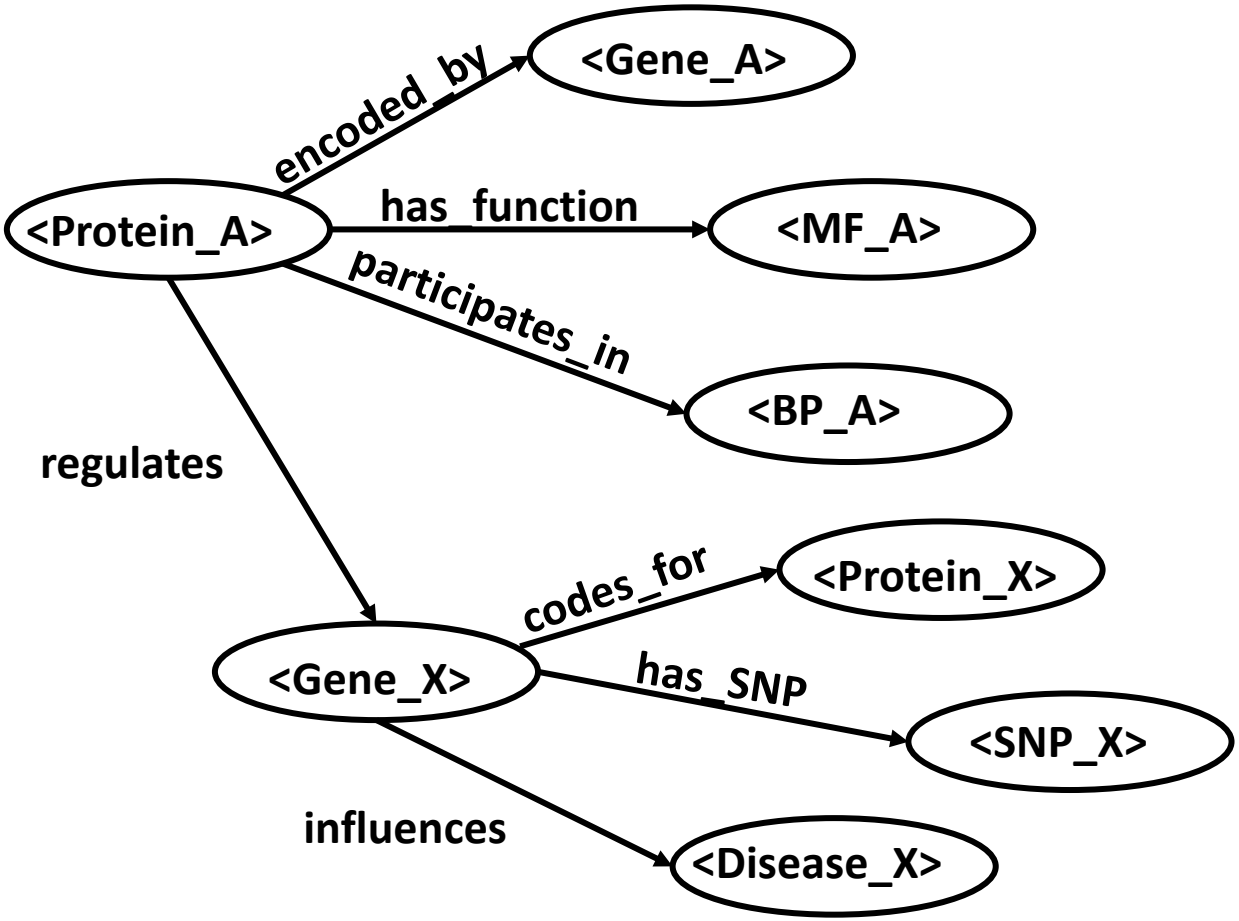
- Combining the triples results in a directed, labeled graph.

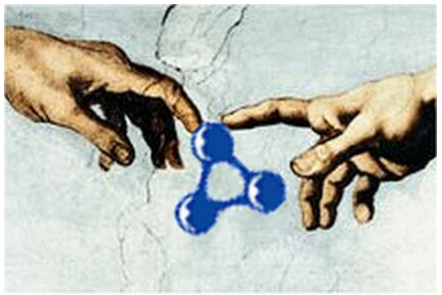




- Can be joined with other graphs.

- Connected using shared URIs.



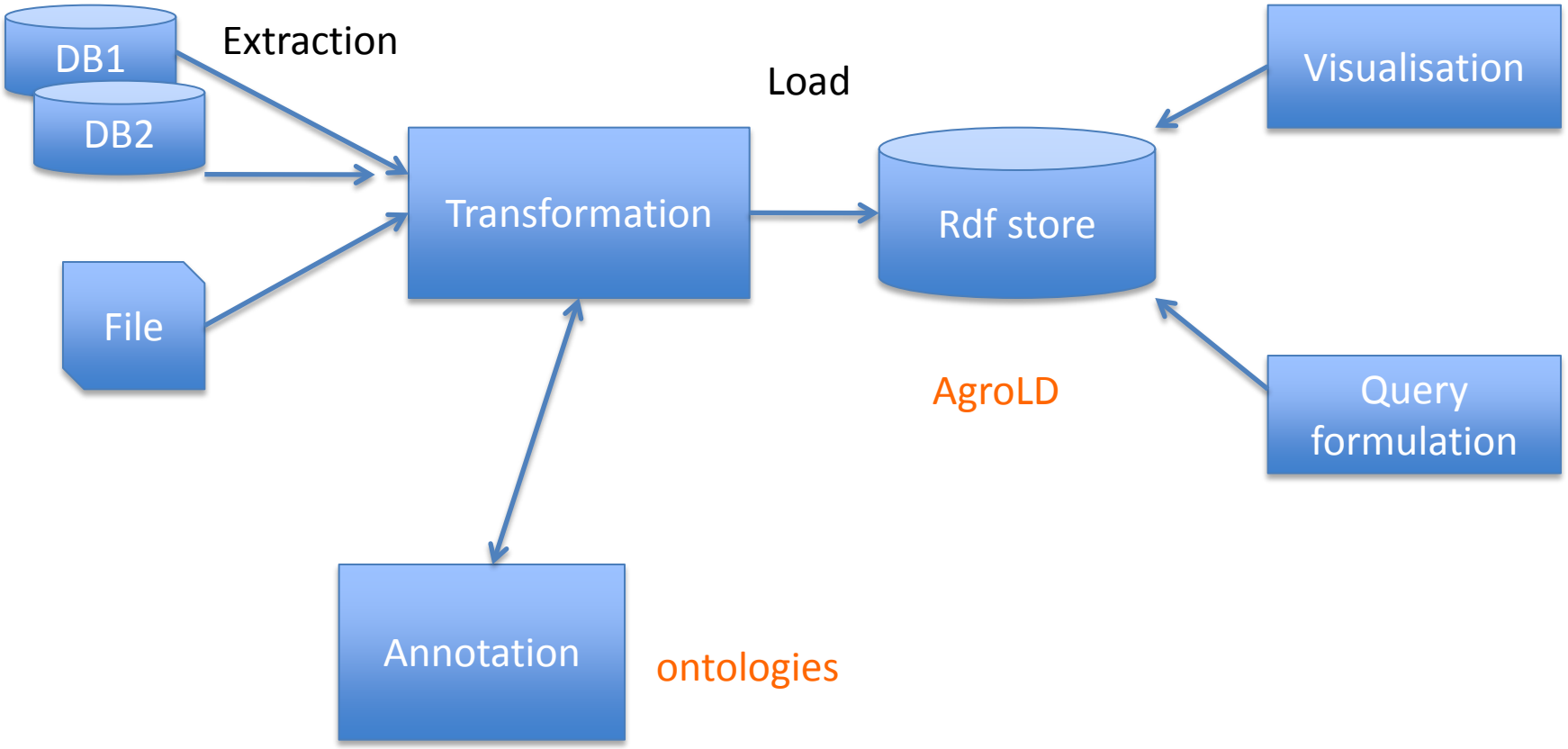


SPARQL

- Language which allows querying RDF models (graphs)
- Powerful, flexible
- Its syntax is similar to the one of SQL


Some results

Multi-scale integration



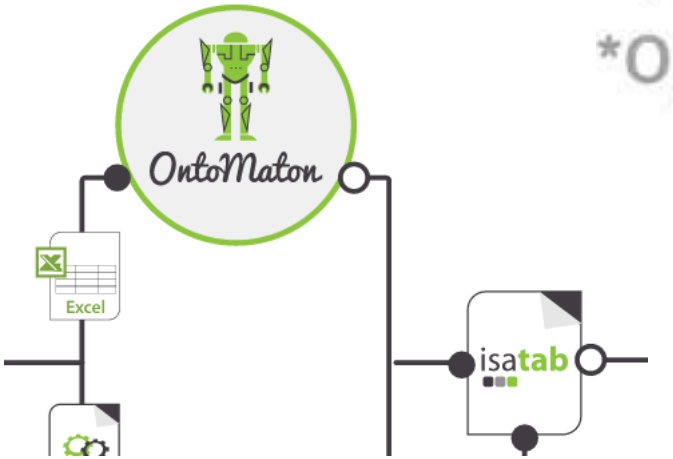
Data & schema extraction

Google refine

talend*

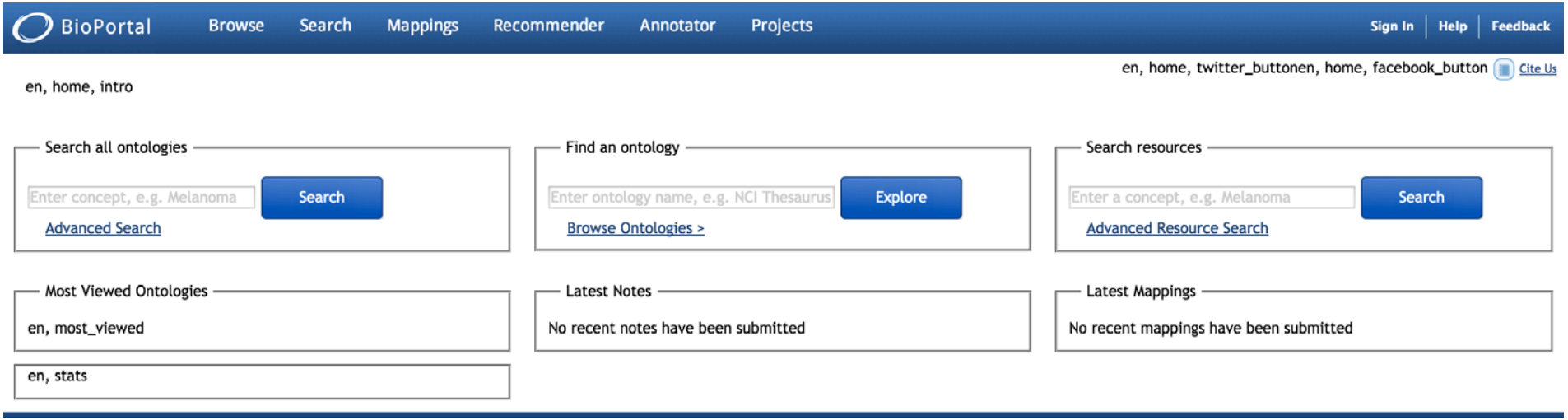
*open data solutions

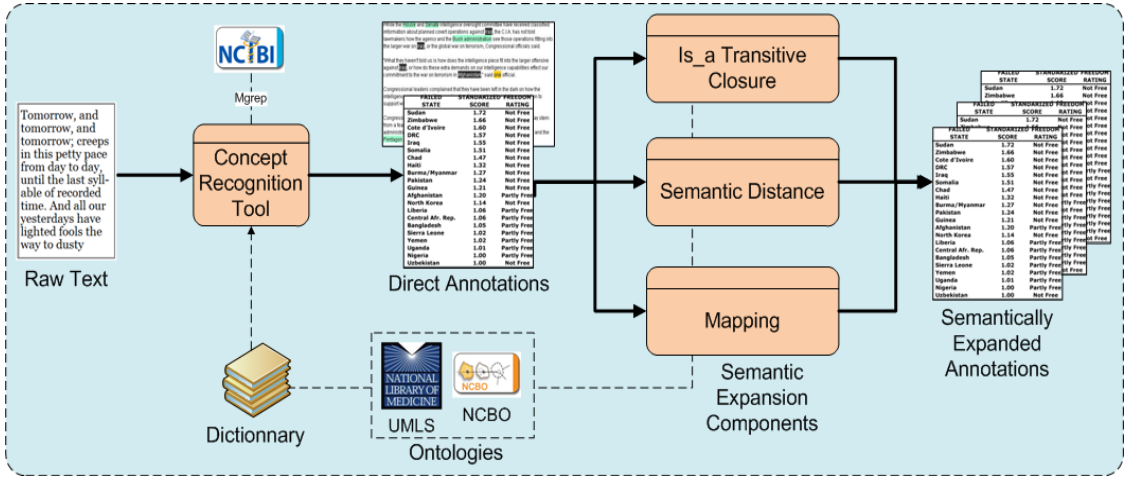
Describe & curate your experiment with geo-graphically distributed collaborators



python™

Bioportal an ontology repository

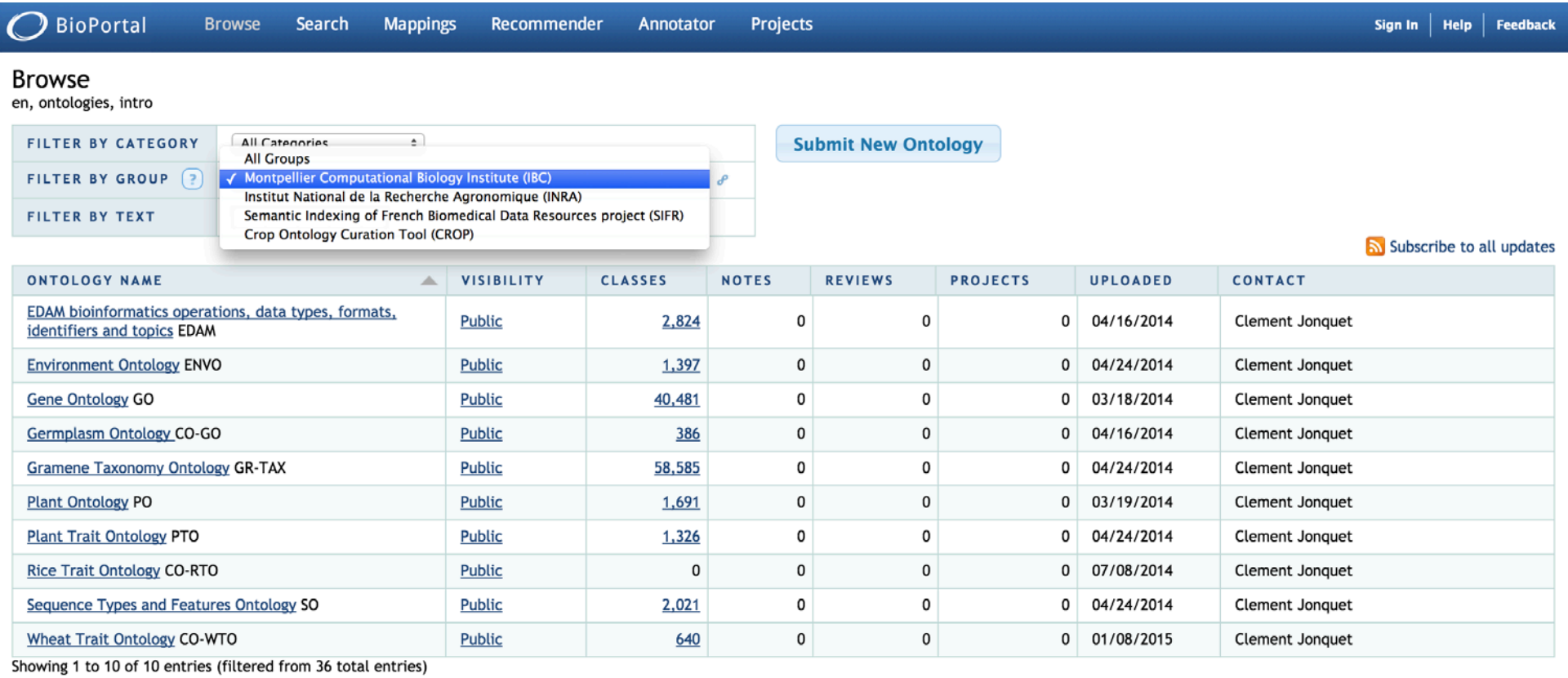




1. direct annotations
1. semantically expanded
1. aggregated and scored

From Clement Jonquet

Bioportal



- Customization
- Groups management
- Restful API
- Enhance communication between websmatch
- applying to crop ontology management

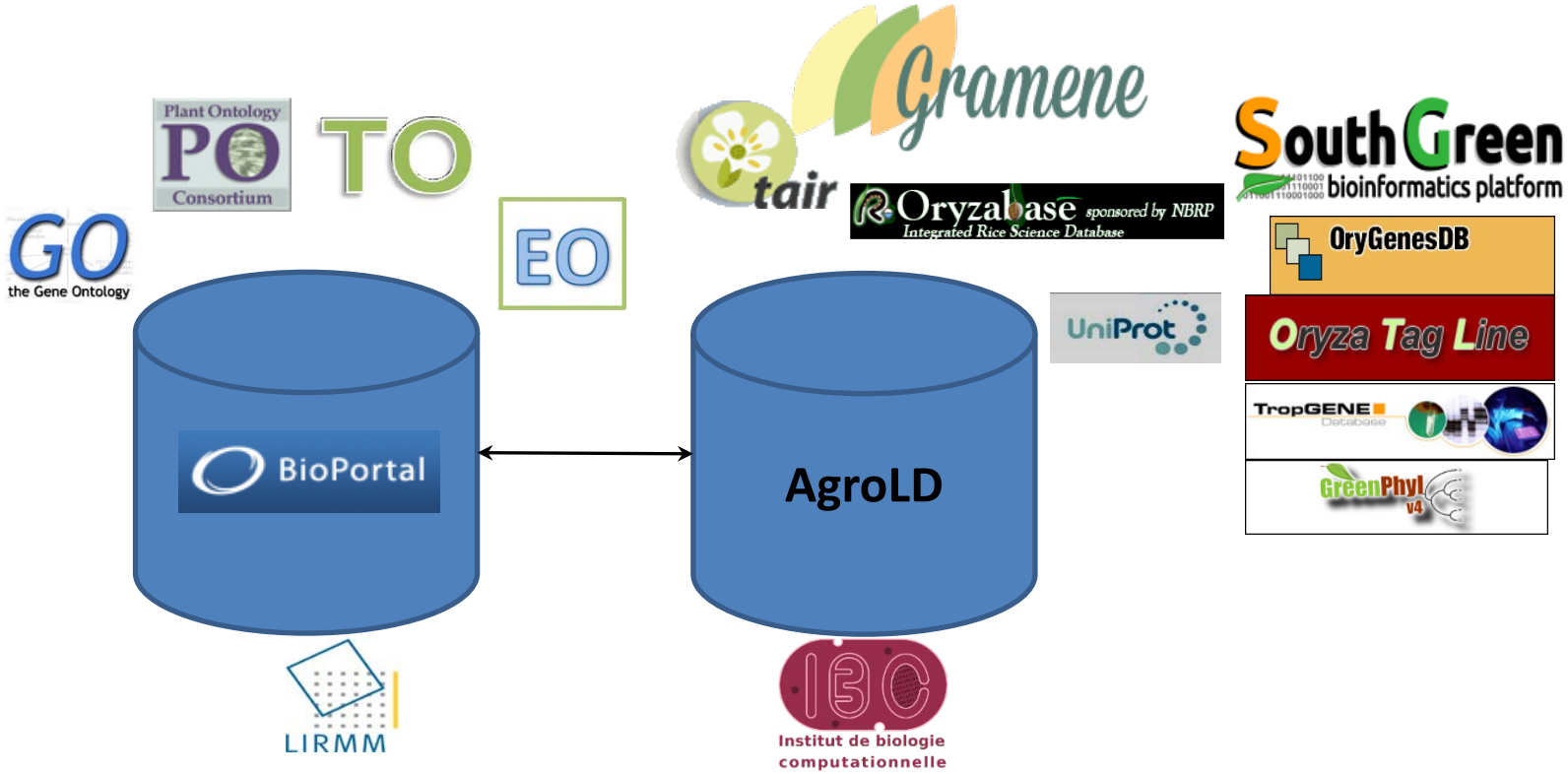
Agronomic Linked Data (AgroLD)

- Semantic web based system that captures knowledge pertaining to plant data

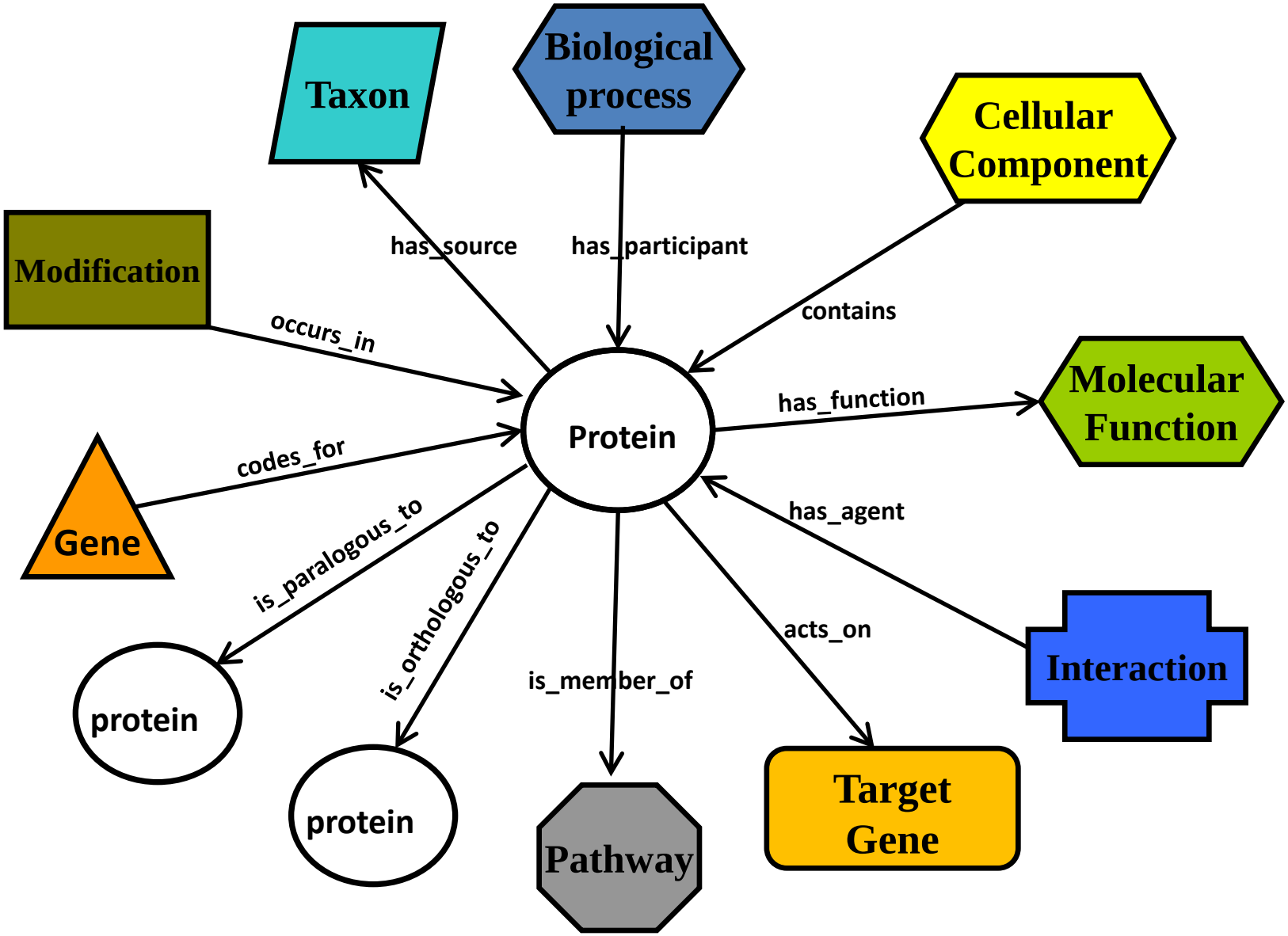
- Aim:
  - Capability to answer complex real life questions
  - Efficient information integration / retrieval
  - Easy extensibility

AgroLD – Phase I

- AgroLD will be developed in phases –
  - **Phase I**: includes data on *Oryza* sps. and *Arabidopsis thaliana*
  - SPARQL endpoint: **agrold.org**



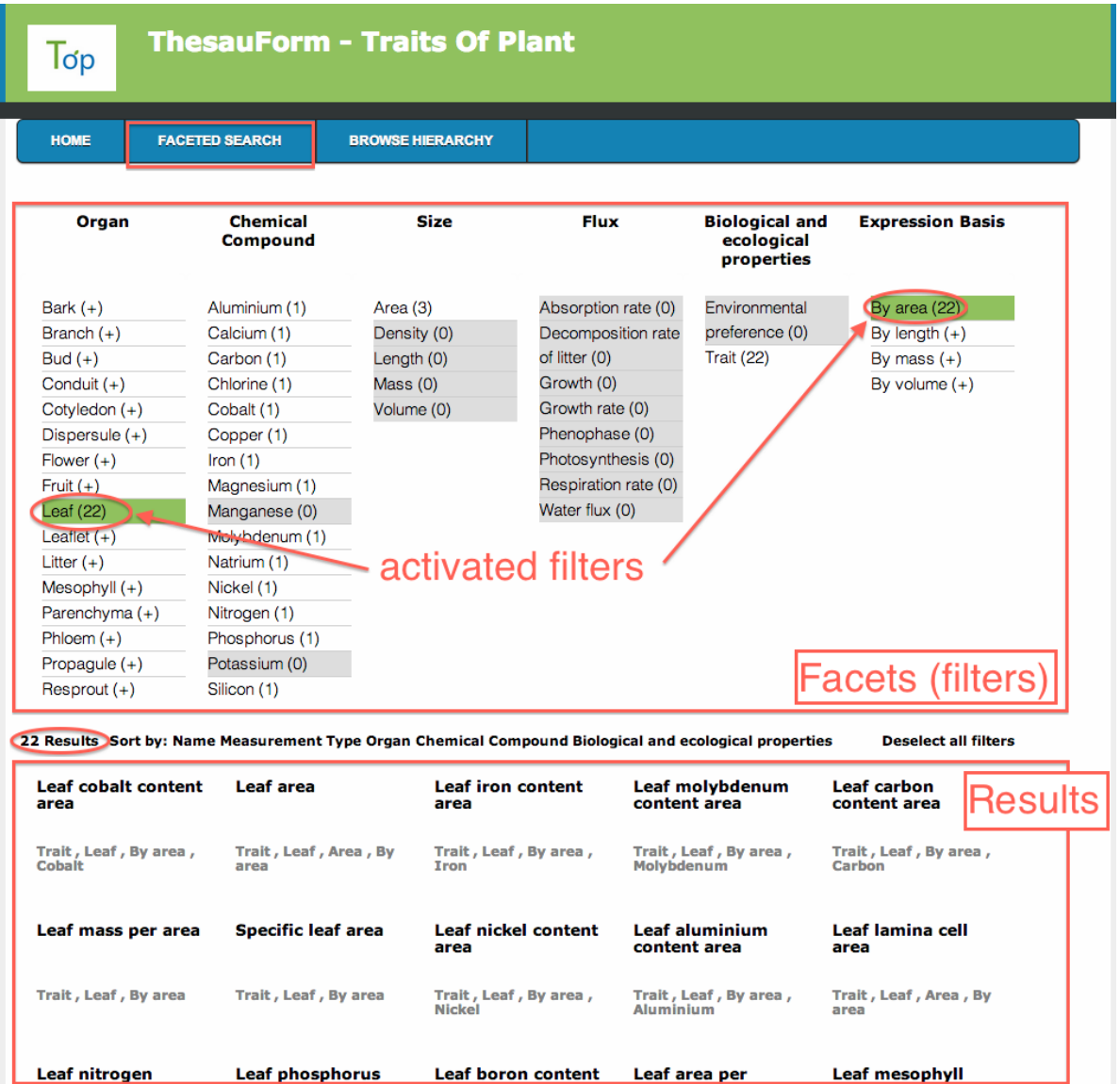
Knowledge representation in AgroLD



Visualization of the collections of traits: the concept of facets

Marie-Angélique Laporte, Luca Matteis, Harold Duruflé & Elizabeth Arnaud

Faceted search benefits



- Facilitate the thesaurus appropriation by the end-users :
  - Reorganize in an intuitive way the thesaurus terms


Thanks !