

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

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

Outline

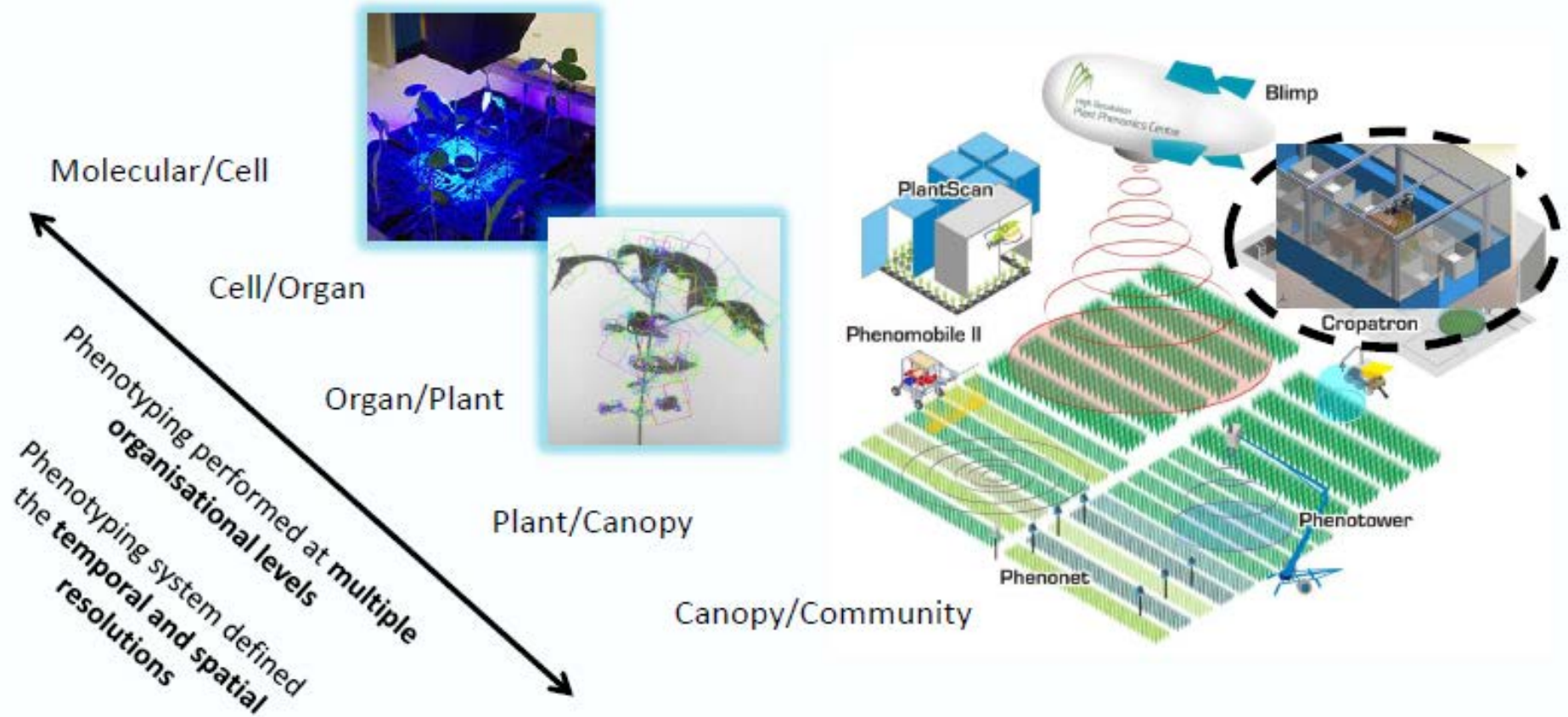
- Data Landscape
- Data integration challenges in the Life Sciences
- Ontologies/ Semantic Web Technologies
- Some example applications

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)

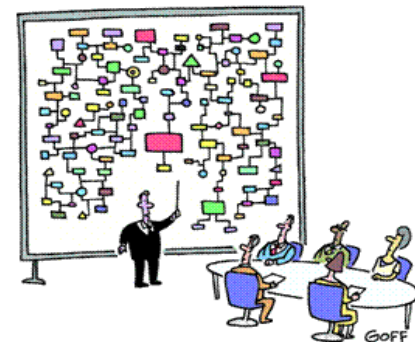


Which resources to use?

- 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."

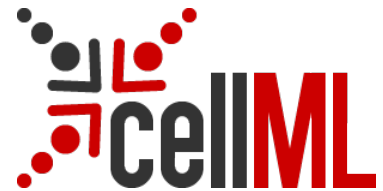
Today's Web

- 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
- **Minimum Information for Biological and Biomedical Investigations (MIBBI)**
- **Investigation, Study, and Assay (ISA)**



isatools



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 **Open Biomedical Ontologies (OBO)** initiative:
 - serves as an umbrella for well structured orthogonal ontologies.
 - Ontologies represented in OBO format and OWL



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 fieldbooks and breeding databases & repositories
- A visualization tool supporting curation of trait lists by a distributed community
- A discussion Forum

Crop Ontology Curation Tool

[Home](#) [About](#) [Users](#) [Feedback](#)



The Crop ontology is a service of the [Integrated Breeding Platform](#). Guidelines are available at the [Crop Ontology wiki](#); list of crop ontology codes and obo files are on the [GCP Pantheon](#). Check [Semantics for Biodiversity](#) web site. New icons appearing on the homepage next to each ontology, will let you download the ontology in [RDF/Turtle](#) format. Workshop on Crop Ontology and Phenotyping Data Interoperability, 31 March-4th April 2014, Montpellier <http://tiny.cc/rw51ax>

[Add New Terms](#) [API](#) [Help](#) [Agrtials](#) [Annotation Tool](#) [Register](#) [Login](#)

Latest

General Germplasm Ontology

FAO/IPGRI Multi-Crop Passport Descriptor 87 terms [BIOVERS](#)
FAO/IPGRI Multi-Crop Passport Descriptor

Germplasm 386 terms [SHRESTHA](#)
germplasm

ICIS germplasm method 166 terms [SHRESTHA](#)
ICIS germplasm methods

Phenotype and Trait Ontology

Banana 52 terms [IVANDENBERGH](#)
Banana beta version

Barley Trait Dictionary 76 terms [RPSVERMA](#)
ICARDA - Trait Dictionary Version Beta

Barley Trait POLAPGEN Ontology 148 terms [HCWI](#)
Barley Trait Ontology 6 June 2013 submitted by the Institute of Plant Genetics Poznan on behalf of Polapgen Consortium Poland

Cacao 8 terms [CACAONET](#)
DRAFT - Cacao Ontology

Cassava 205 terms [AAFOLABI](#)
IITA - Cassava Ontology - September 2014

Location and Environmental Ontology

Country and Location 1118 terms [SHRESTHA](#)
Describes official ISO 3166-1 alpha-2, alpha-3 and numeric country codes along with location names.

Crop Research 256 terms [SHRESTHA](#)
Describes experimental design, environmental conditions and methods associated with the crop study/experiment/trial and their evaluation.

Plant Anatomy & Development Ontology

Banana Anatomy 149 terms [CHANNELIERE](#)
Banana Anatomy

Plant Ontology 1710 terms [COOPERL](#)
The Plant Ontology describes plant anatomy and morphology and stages of development for all plants. The goal of the PO is to establish a semantic framework for meaningful cross-species queries across gene expression and phenotype data sets from plant genomics and genetics experiments.

TEST 130 terms [LVALETTE](#)
TEST

croponontology.org



Multilingual ontologies



[DOWNLOAD](#)
[SHOW ALL](#)
[EDIT](#)
English ↕

[Overall Panicle Weight](#)
[Permalink](#)
▼ **General**
0 Comments

Identifier CO_324:0000073

[Sorghum](#) ibfieldbook

[Grain weight over panicle](#)

[DOWNLOAD](#)
[SHOW ALL](#)
[EDIT](#)
French ↕

[Poids paniculaire](#)
[Permalink](#)
▼ **General**
0 Comments

[Sorghum](#) ibfieldbook

[Poids grain par panicle](#)

[Poids paniculaire](#) method_of

[Discret](#) scale_of

Identifier CO_324:0000073

Name of method Poids paniculaire

- [Poids des Panicules](#)
- [Poids de 100 Grains](#)
- [Poids des tiges](#)
- [Date de maturité](#)
- [Hauteur Plant](#)
- [Couverture du Grain](#)
- [Score Battage](#)
- [Moisissure des Grains](#)
- [Anthracnose du Grain](#)
- [Anthracnose Foliaire](#)

Describe how measured (method) Calculé = PGR/NbPAN

Growth stages Maturité

name Poids paniculaire

[Add a new attribute](#)

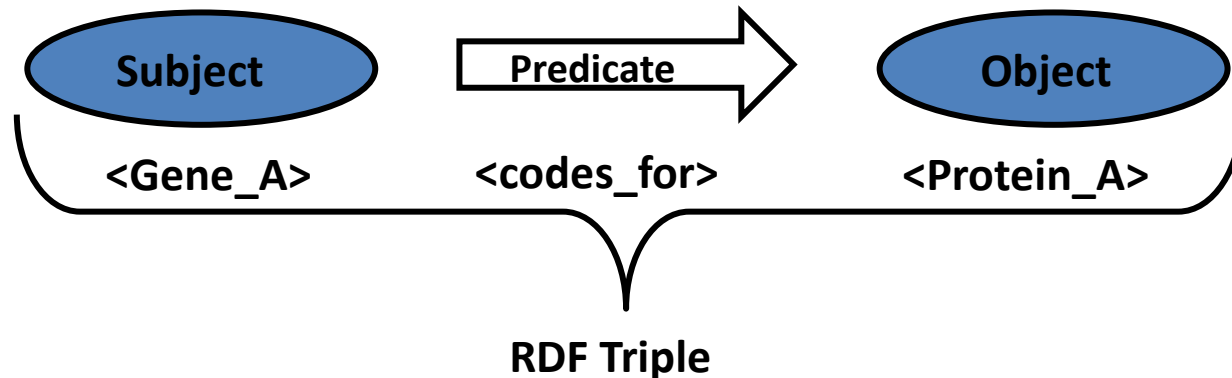
From Elizabeth Arnaud

Semantic Web Technology

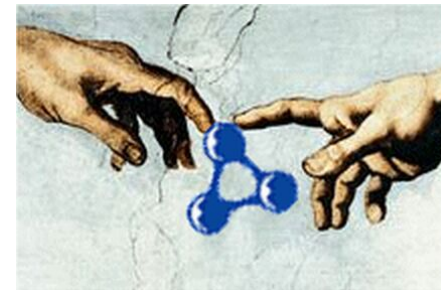
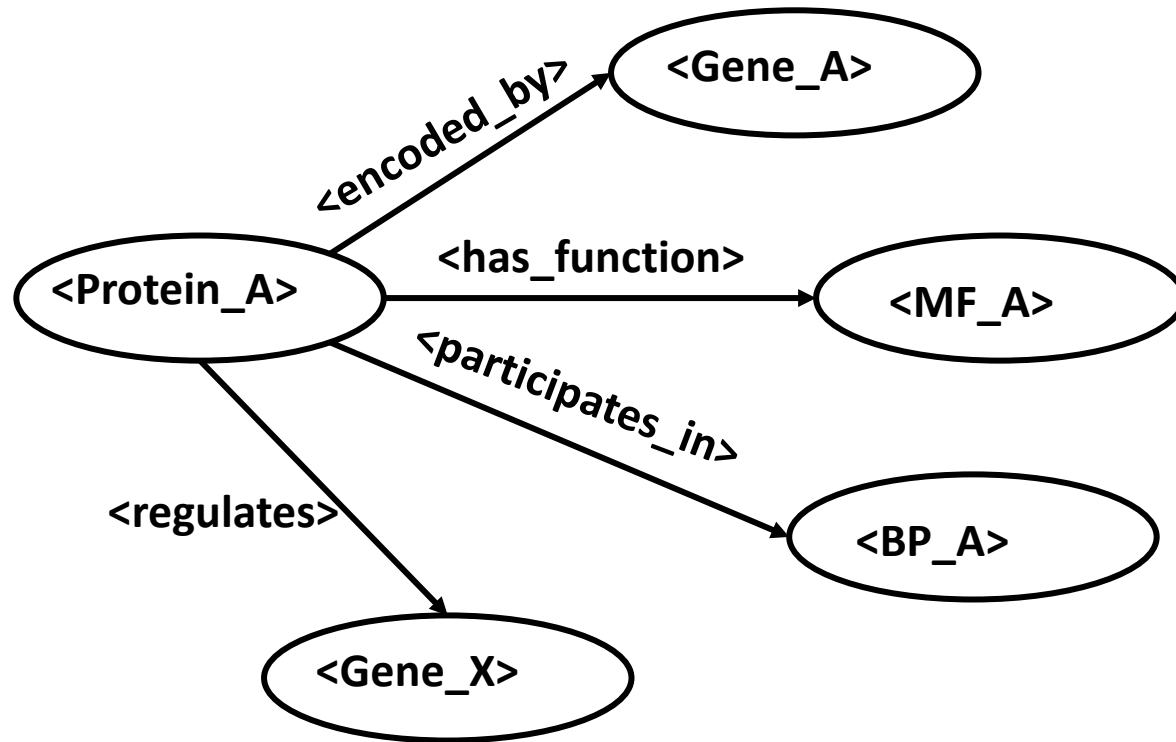
- An extension of the current Web technologies.
- Enables navigation and meaningful use of digital resources.
- Support aggregation and integration of information from diverse sources.
- Based on common and standard formats.

Resource Description Framework (RDF)

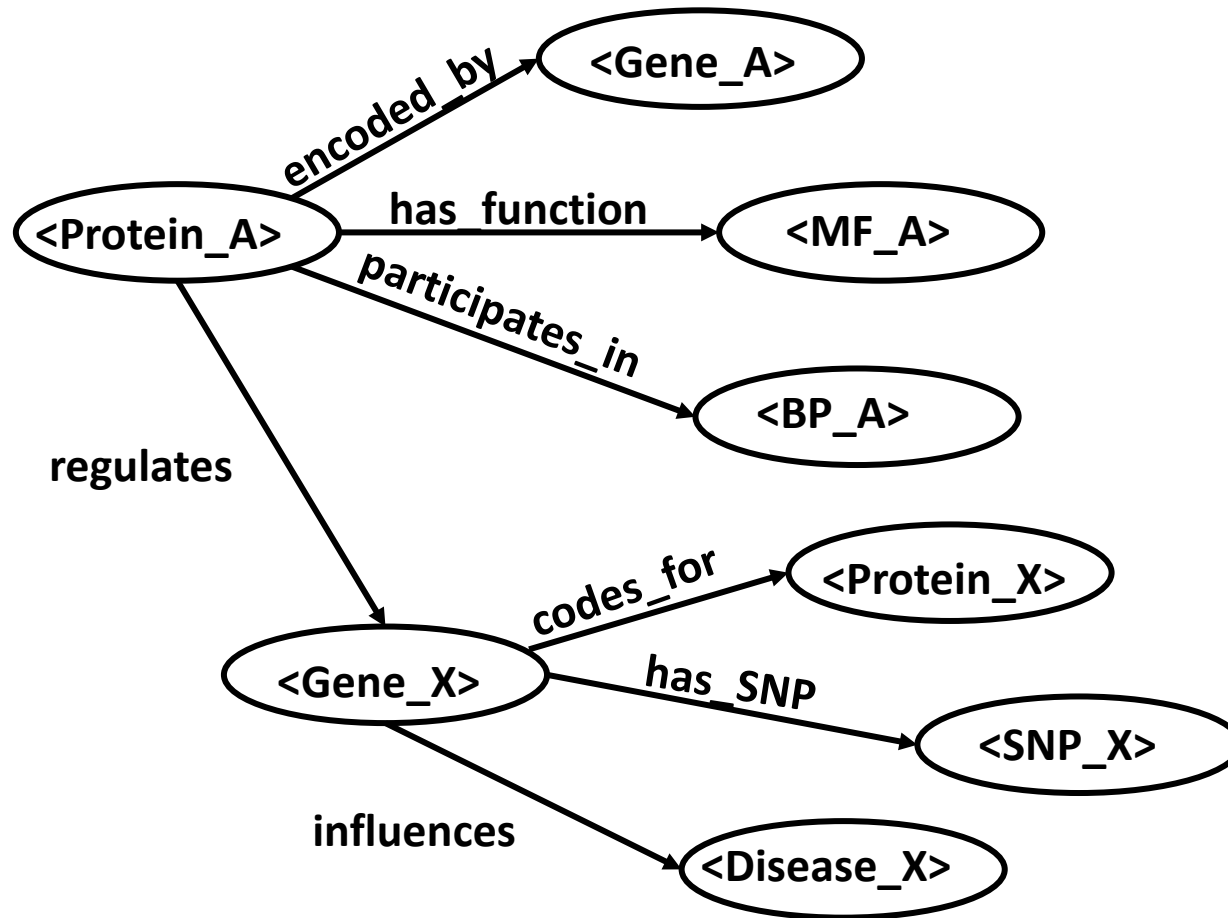
- Framework for representing information about resources on the Web
- Provides a labeled connection between two resources
- Uses Unique Resource Identifiers (URI)
- Statements take the form of triples:



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



- Can be joined with other graphs.
- Connected using shared URIs.



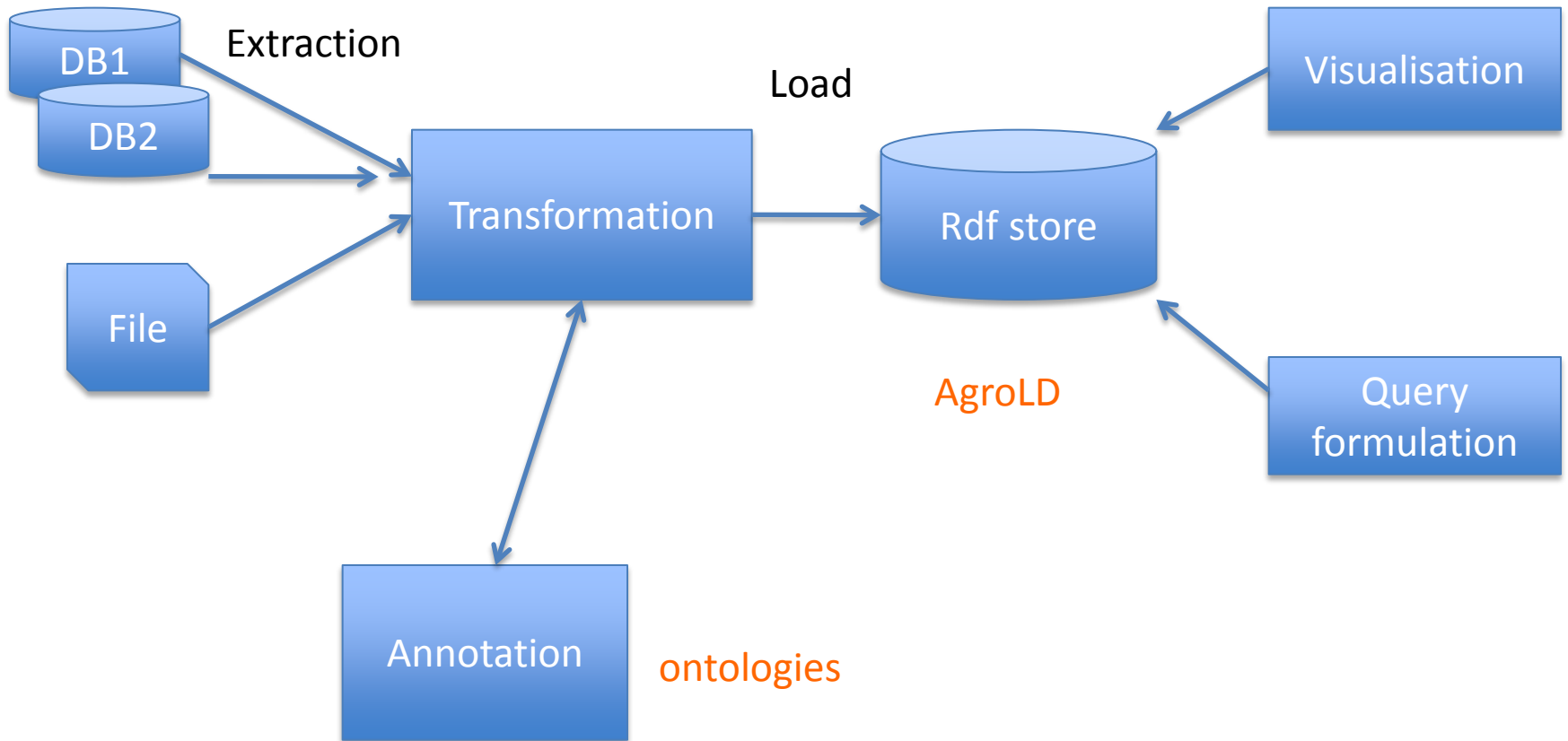
SPARQL

- 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

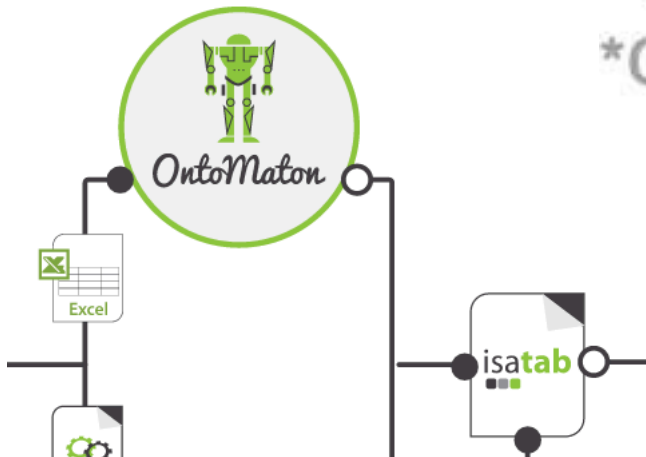


Google refine



Describe & curate your experiment with geographically distributed collaborators

talend*
*open data solutions



Bioportal an ontology repository

en, home, intro

en, home, twitter_buttonen, home, facebook_button [Cite Us](#)

Search all ontologies

Enter concept, e.g. Melanoma

[Advanced Search](#)

Find an ontology

Enter ontology name, e.g. NCI Thesaurus

[Browse Ontologies >](#)

Search resources

Enter a concept, e.g. Melanoma

[Advanced Resource Search](#)

Most Viewed Ontologies

en, most_viewed

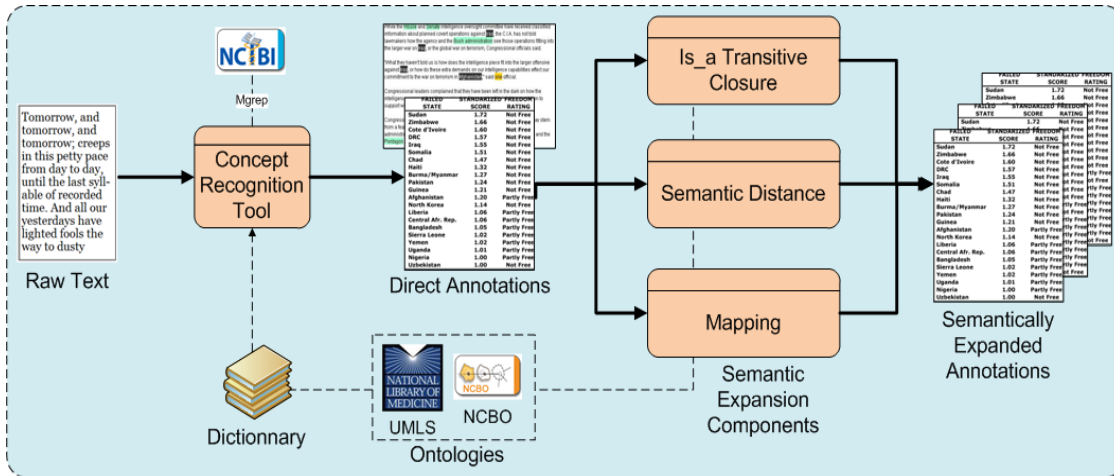
Latest Notes

No recent notes have been submitted

Latest Mappings

No recent mappings have been submitted

en, stats



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

From Clement Jonquet

Bioportal

Browse
en, ontologies, intro

FILTER BY CATEGORY
All Categories
All Groups
✓ Montpellier Computational Biology Institute (IBC)
Institut National de la Recherche Agronomique (INRA)
Semantic Indexing of French Biomedical Data Resources project (SIFR)
Crop Ontology Curation Tool (CROP)

FILTER BY GROUP (?)

FILTER BY TEXT

Submit New Ontology

Subscribe to all updates

ONTOLOGY NAME	VISIBILITY	CLASSES	NOTES	REVIEWS	PROJECTS	UPLOADED	CONTACT
EDAM bioinformatics operations, data types, formats, identifiers and topics EDAM	Public	2,824	0	0	0	04/16/2014	Clement Jonquet
Environment Ontology ENVO	Public	1,397	0	0	0	04/24/2014	Clement Jonquet
Gene Ontology GO	Public	40,481	0	0	0	03/18/2014	Clement Jonquet
Germplasm Ontology CO-GO	Public	386	0	0	0	04/16/2014	Clement Jonquet
Gramene Taxonomy Ontology GR-TAX	Public	58,585	0	0	0	04/24/2014	Clement Jonquet
Plant Ontology PO	Public	1,691	0	0	0	03/19/2014	Clement Jonquet
Plant Trait Ontology PTO	Public	1,326	0	0	0	04/24/2014	Clement Jonquet
Rice Trait Ontology CO-RTO	Public	0	0	0	0	07/08/2014	Clement Jonquet
Sequence Types and Features Ontology SO	Public	2,021	0	0	0	04/24/2014	Clement Jonquet
Wheat Trait Ontology CO-WTO	Public	640	0	0	0	01/08/2015	Clement Jonquet

Showing 1 to 10 of 10 entries (filtered from 36 total entries)

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

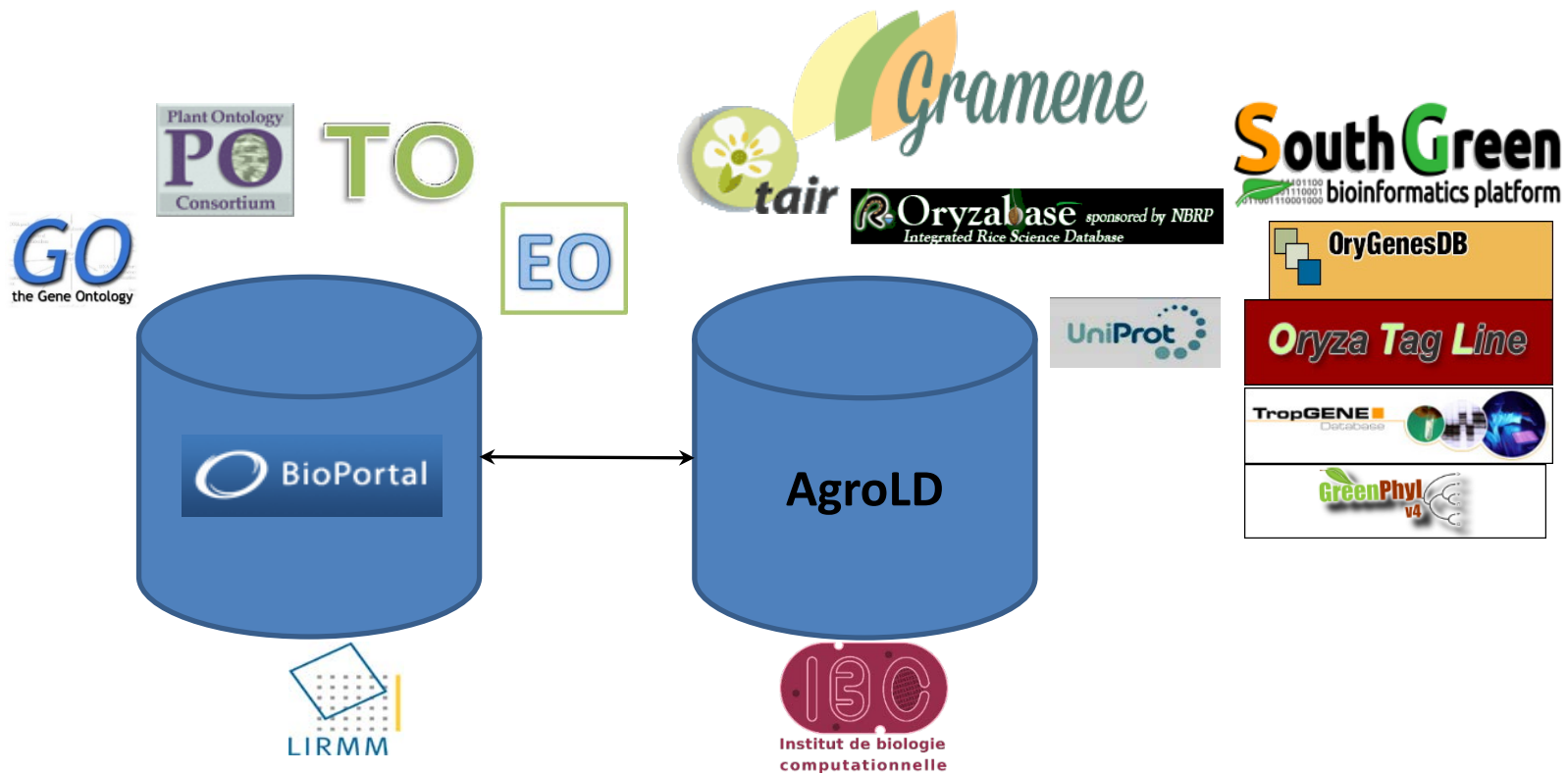
Agronomic Linked Data (AgroLD)

- 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

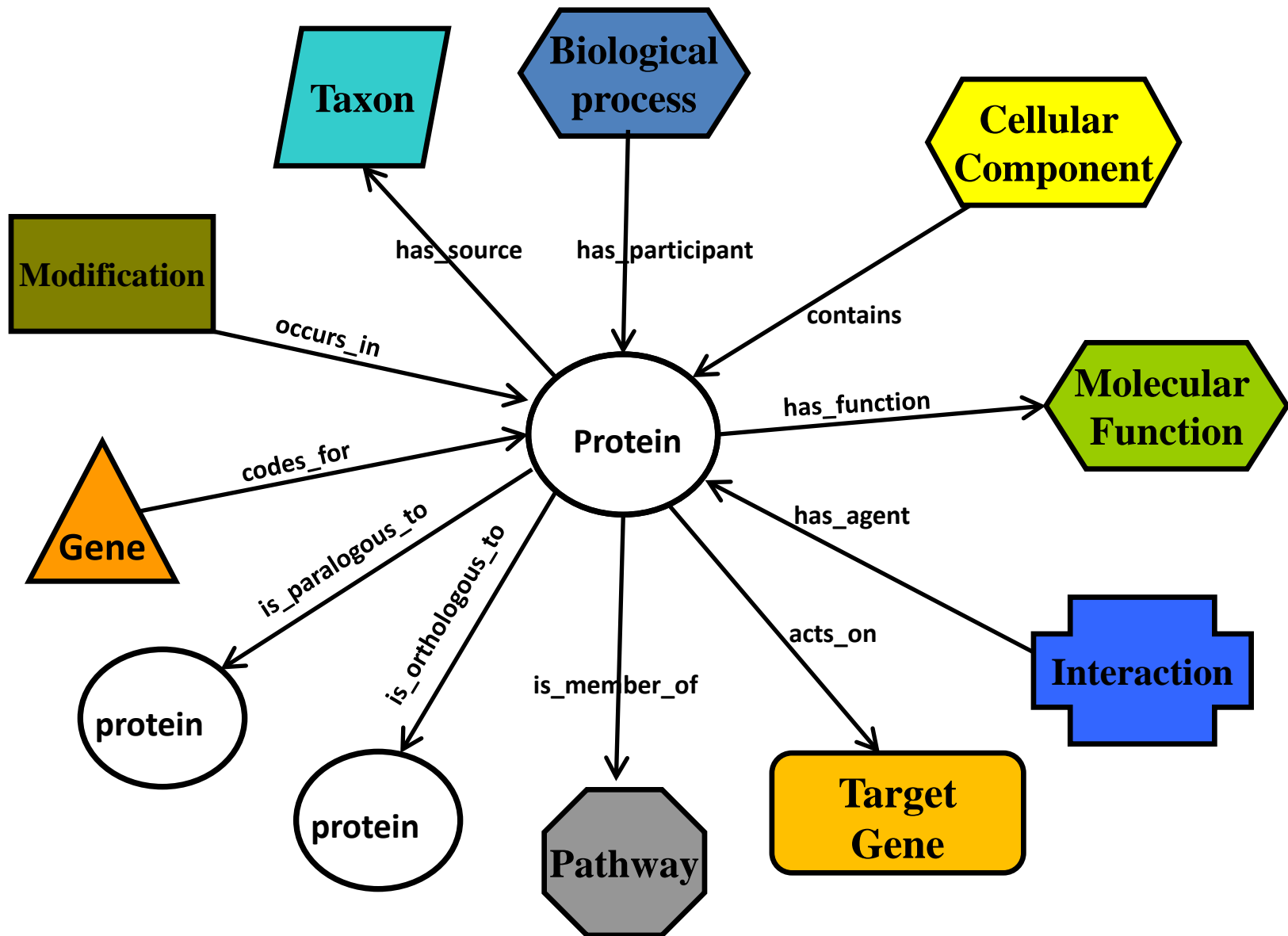


AgroLD – Phase I

- 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

Top ThesauForm - Traits Of Plant

HOME FACETED SEARCH BROWSE HIERARCHY

Organ	Chemical Compound	Size	Flux	Biological and ecological properties	Expression Basis
Bark (+)	Aluminium (1)	Area (3)	Absorption rate (0)	Environmental preference (0)	By area (22)
Branch (+)	Calcium (1)	Density (0)	Decomposition rate of litter (0)	Trait (22)	By length (+)
Bud (+)	Carbon (1)	Length (0)	Growth (0)		By mass (+)
Conduit (+)	Chlorine (1)	Mass (0)	Growth rate (0)		By volume (+)
Cotyledon (+)	Cobalt (1)	Volume (0)	Phenophase (0)		
Dispersule (+)	Copper (1)		Photosynthesis (0)		
Flower (+)	Iron (1)		Respiration rate (0)		
Fruit (+)	Magnesium (1)		Water flux (0)		
Leaf (22)	Manganese (0)				
Leaflet (+)	Molybdenum (1)				
Litter (+)	Sodium (1)				
Mesophyll (+)	Nickel (1)				
Parenchyma (+)	Nitrogen (1)				
Phloem (+)	Phosphorus (1)				
Propagule (+)	Potassium (0)				
Resprout (+)	Silicon (1)				

activated filters

Facets (filters)

22 Results Sort by: Name Measurement Type Organ Chemical Compound Biological and ecological properties Deselect all filters

Leaf cobalt content area	Leaf area	Leaf iron content area	Leaf molybdenum content area	Leaf carbon content area
Trait , Leaf , By area , Cobalt	Trait , Leaf , Area , By area	Trait , Leaf , By area , Iron	Trait , Leaf , By area , Molybdenum	Trait , Leaf , By area , Carbon
Leaf mass per area	Specific leaf area	Leaf nickel content area	Leaf aluminium content area	Leaf lamina cell area
Trait , Leaf , By area	Trait , Leaf , By area	Trait , Leaf , By area , Nickel	Trait , Leaf , By area , Aluminium	Trait , Leaf , Area , By area
Leaf nitrogen	Leaf phosphorus	Leaf boron content	Leaf area per	Leaf mesophyll

Results



Thanks !