Institut de recherche pour le développement

Institut de biologie computationnelle

Exposing the French agronomic resources as Linked Data

Pierre Larmande
Institute of Research for Development (IRD)
Head of data integration group at the Institute of Computational Biology
pierre.larmande@ird.fr

Institut Français de Bioinformatique (PIA)

PIA

Institut Français de Bioinformatique (PIA)

Mission générale : fournir des **ressources** de base en bioinformatique à la communauté des sciences de la vie

Directeur : J-F Gibrat, DR INRA

Responsable cellule e-infrastructure : C. Blanchet, IR CNRS

Infrastructure nationale de service en bioinformatique

– **Données** : Fournir un accès à des collections de données spécialisées à haute valeur ajoutée issues de l'expertise du laboratoire d'accueil

– **Outils** : Développer et mettre à disposition des outils et services en lignes pour analyser les données correspondant à l'expertise scientifique du laboratoire d'accueil

– **Appui** aux projets scientifiques et hébergement sur une infrastructure informatique

– **Infrastructure** : Mettre à disposition une infrastructure informatique

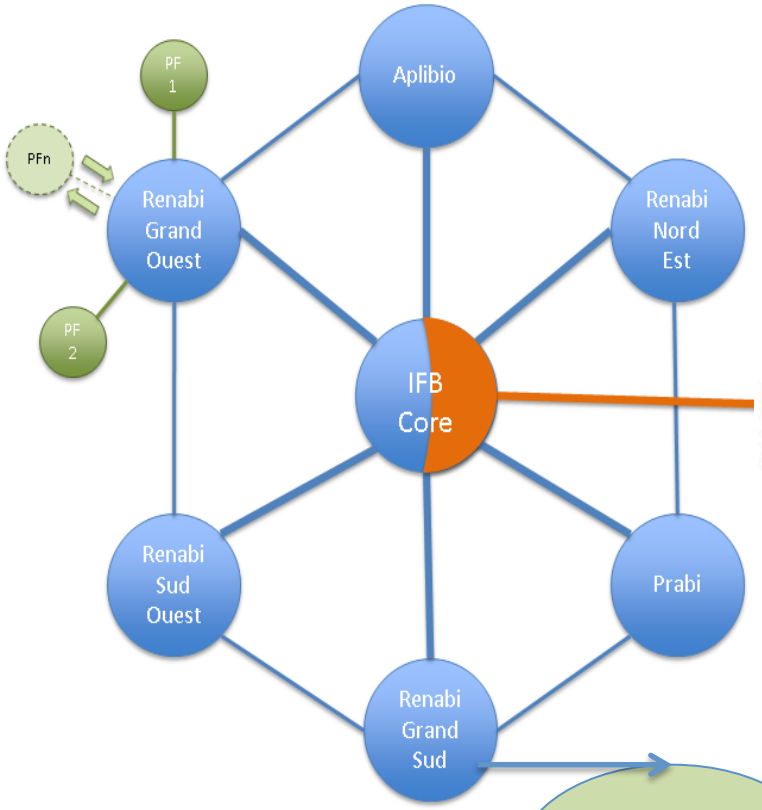
dédiée à l'analyse des données des sciences du vivant (matériel,données, outils)

– **Formations**

cf. http://france-bioinformatique.fr

Elixir

ELIXIR has been awarded €19 million from the EU to accelerate the implementation of Europe's life-science data infrastructure over the next four years.

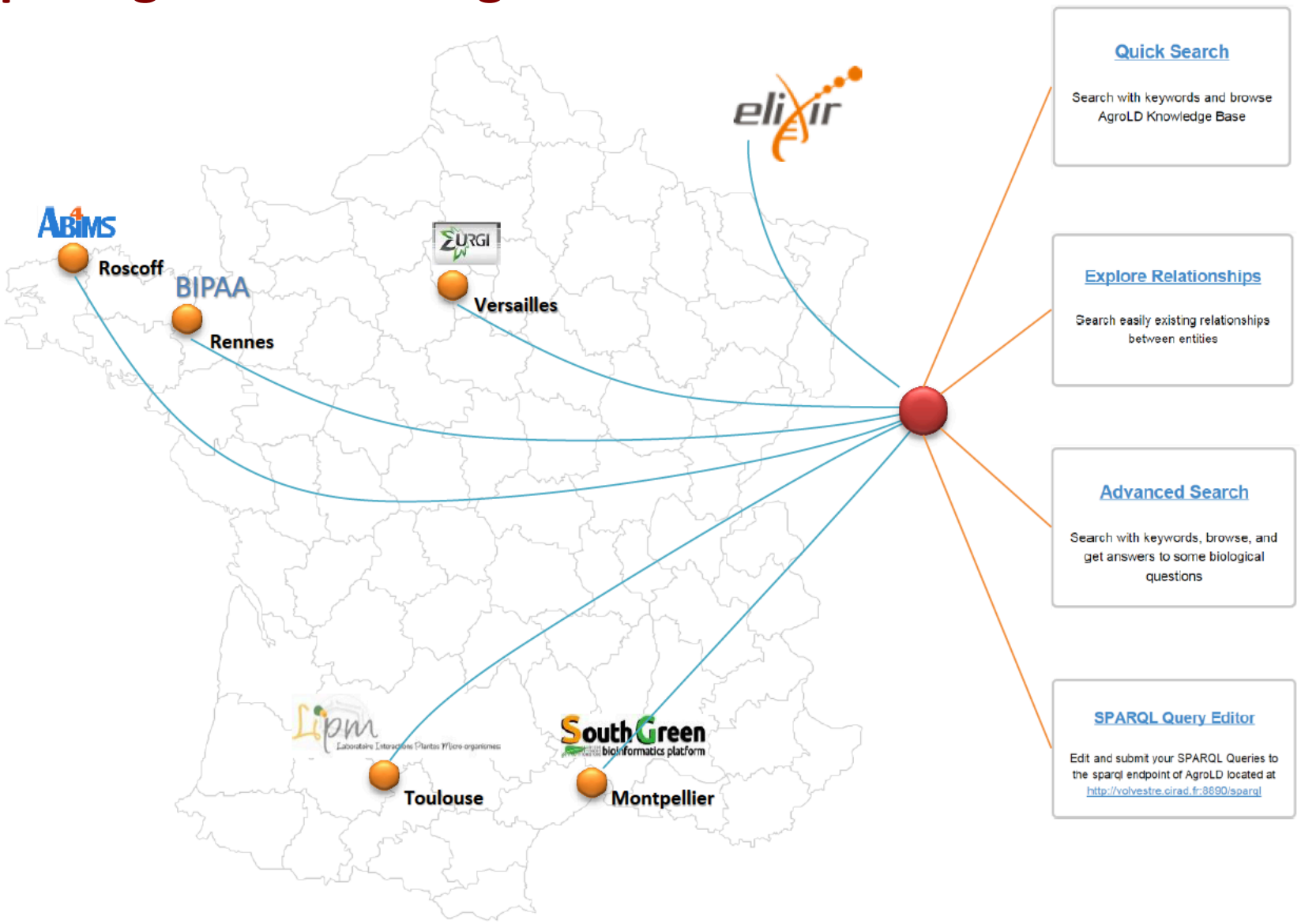


South Green

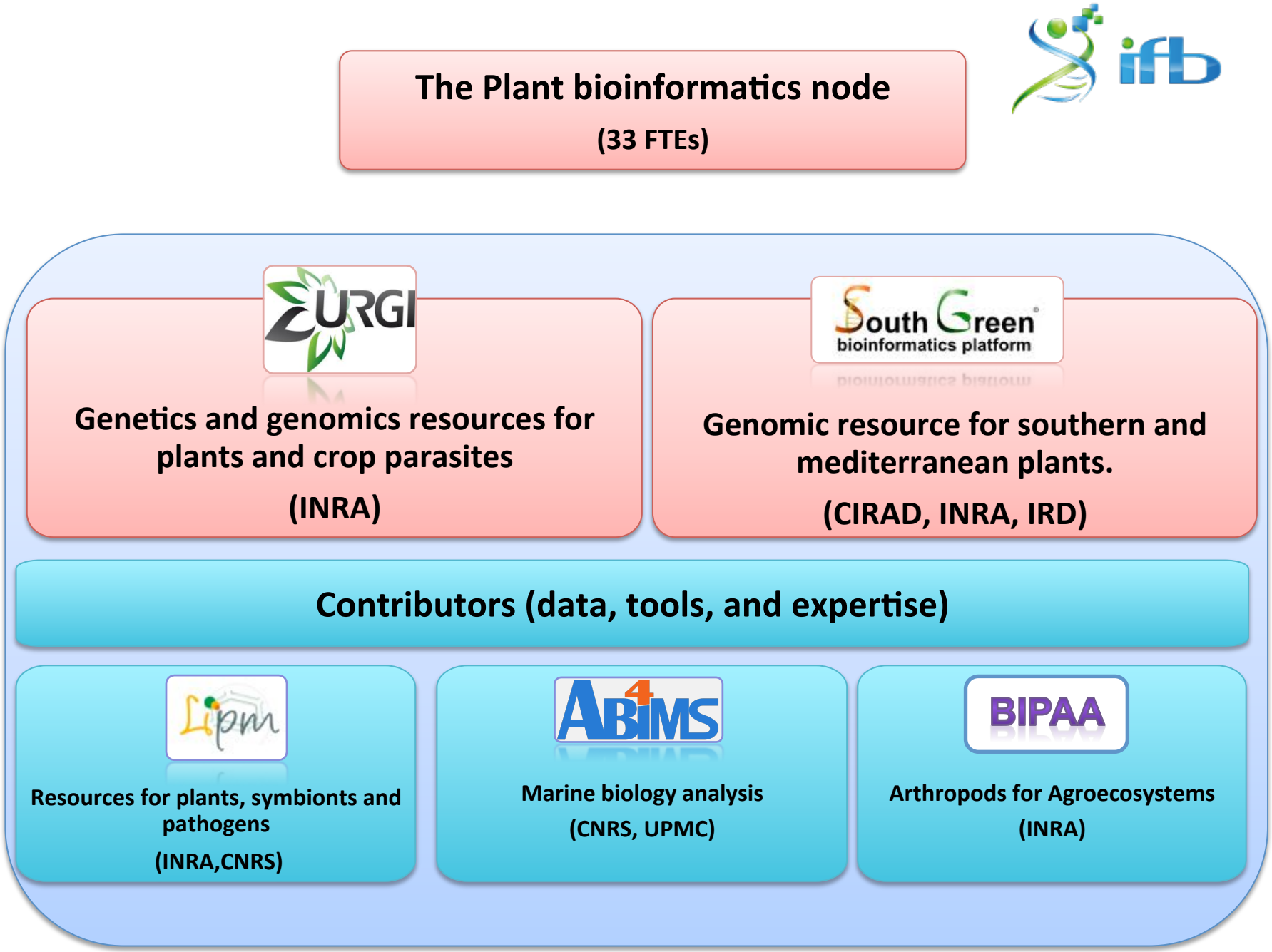
i-trop

- Rare disease
- Genomics for Human health
- Plant use case
- Marine biology use case

Exposing the French agronomic resources as Linked Data



IFB - 2015 – Plant node



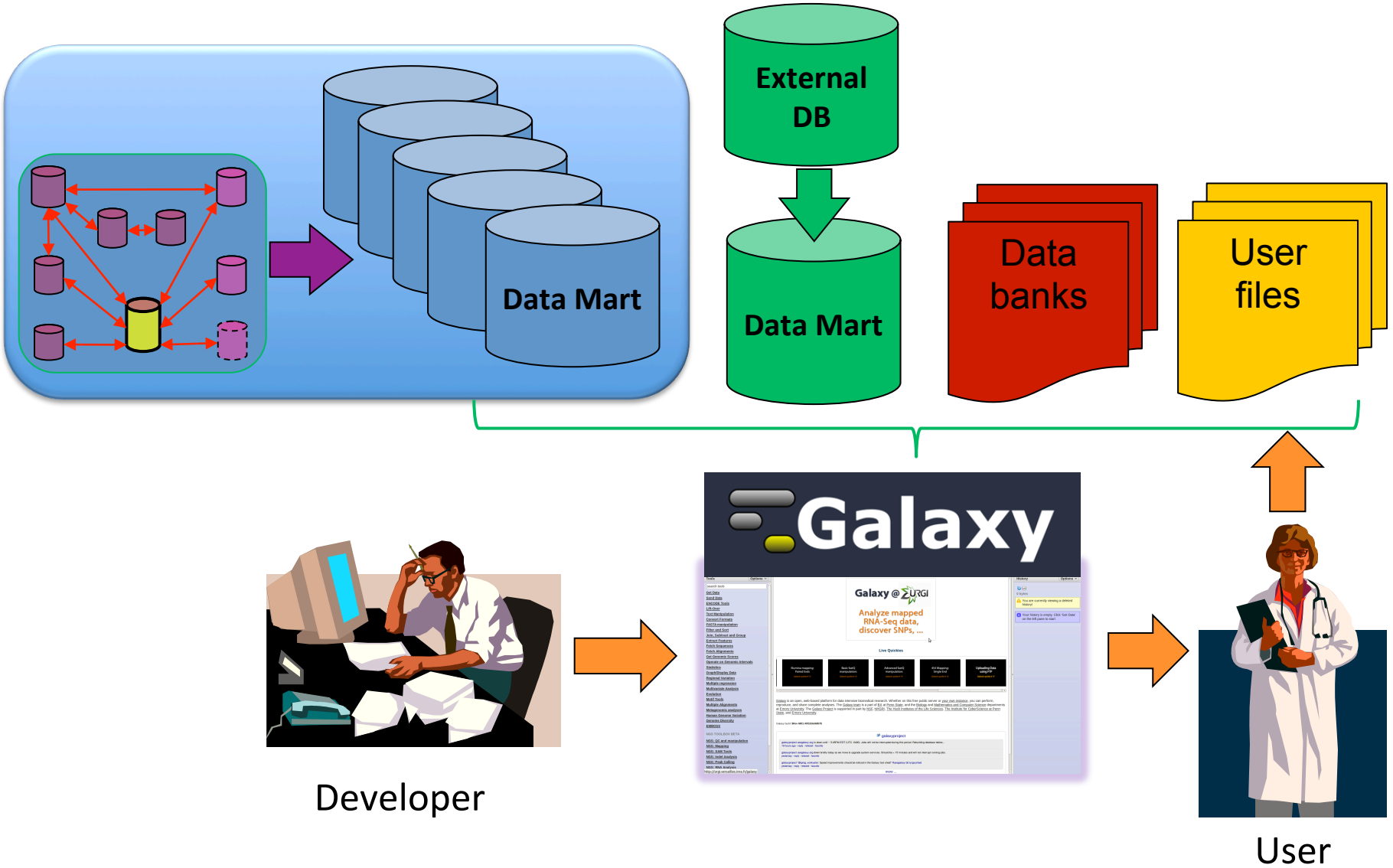
From H. Quesneville

Services

- analysis
- genomics annotation
- software hosting
- data repository
- data integration
- database design
- software engineering

From H. Quesneville

Data Integration / data mining Workbench

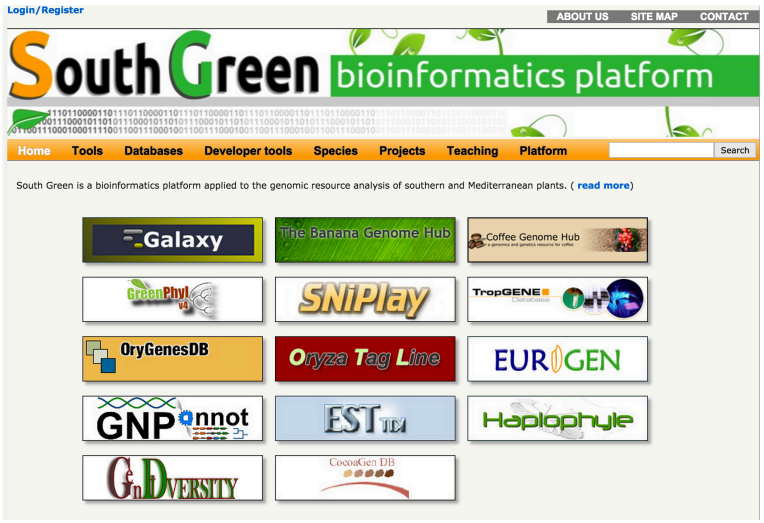


From H. Quesneville

Resources

urgi.versailles.inra.fr

www.southgreen.fr



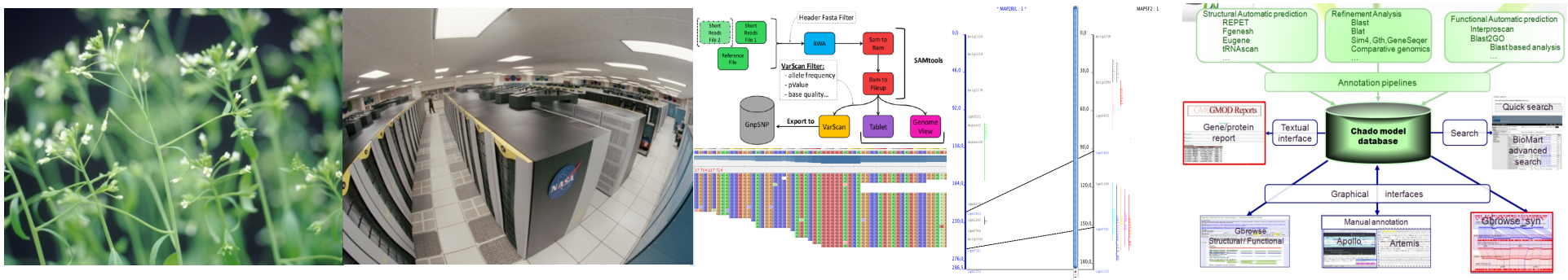
From H. Quesneville

IFB technical project

- WP1:Develop a RDF-based semantic interoperability between the Plant bioinformatics node databases
- WP2: Develop an intuitive Google-like search portal
- WP3: Setup the “IT-core hub” workflows under the Galaxy

From H. Quesneville

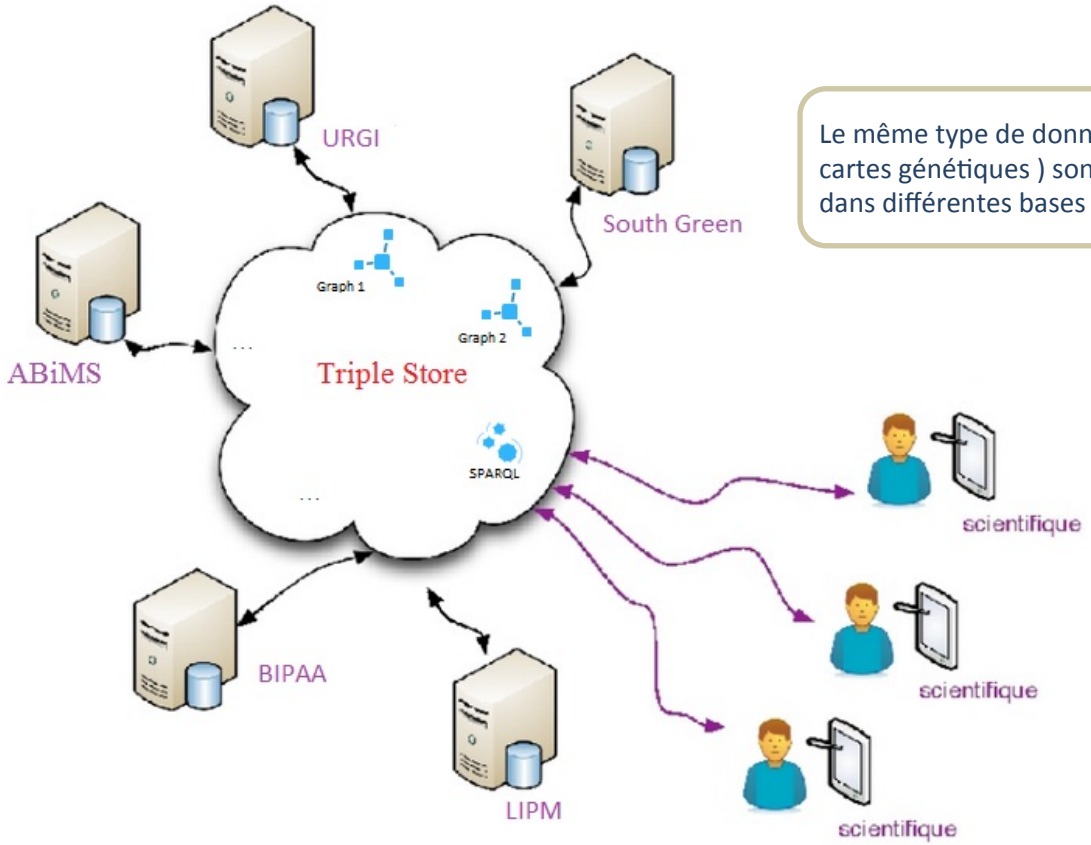
WP1 : RDF "store" permettant d'interconnecter sémantiquement l'ensemble des bases de données végétales des plateformes bioinformatiques IFB



IFB plant node: (URGI), South Green, IFB Grand Ouest (ABiMS), GenOuest (BIPAA) et LIPM.

From H. Quesneville

Develop a web semantic interoperability

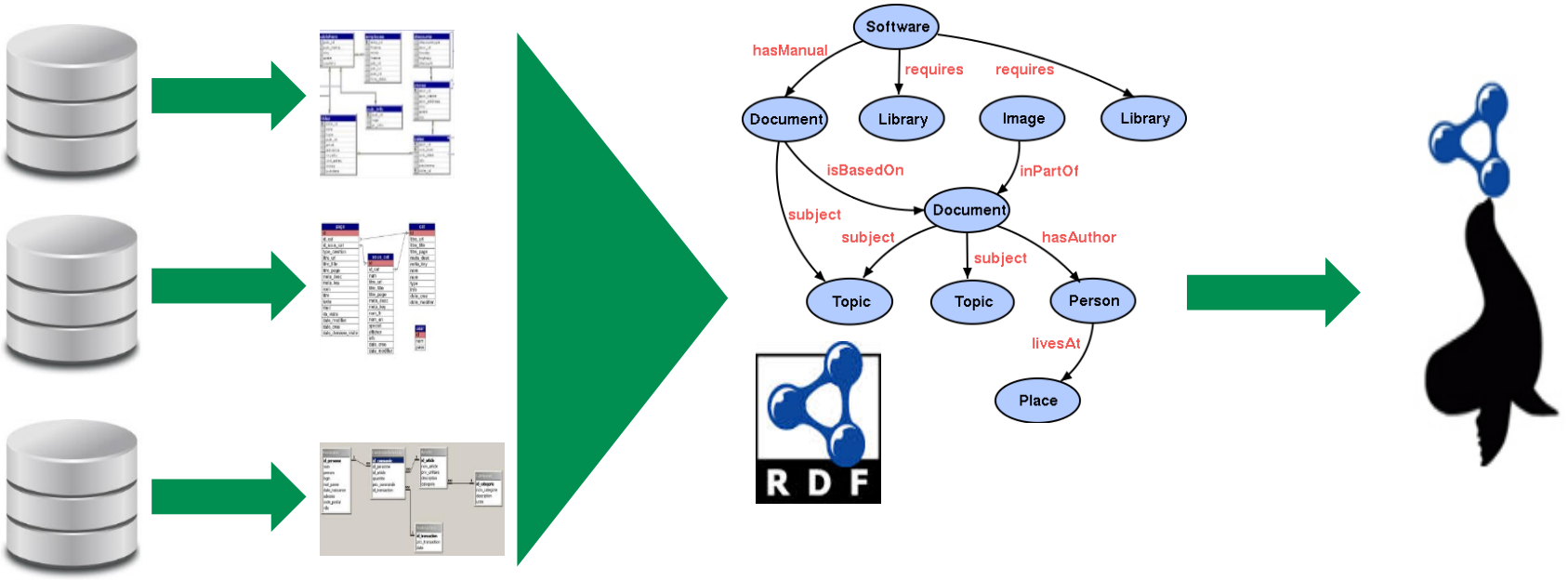


Le même type de données (par exemple SNP , collections génétiques , cartes génétiques) sont souvent représentées différemment dans différentes bases de données.

Développement d'une application Web interrogeant les différentes sources de données.

From N. El Hassouni

Develop a web semantic interoperability



Ontology based annotation of database schemas (GO, PO, TO, CO, …)

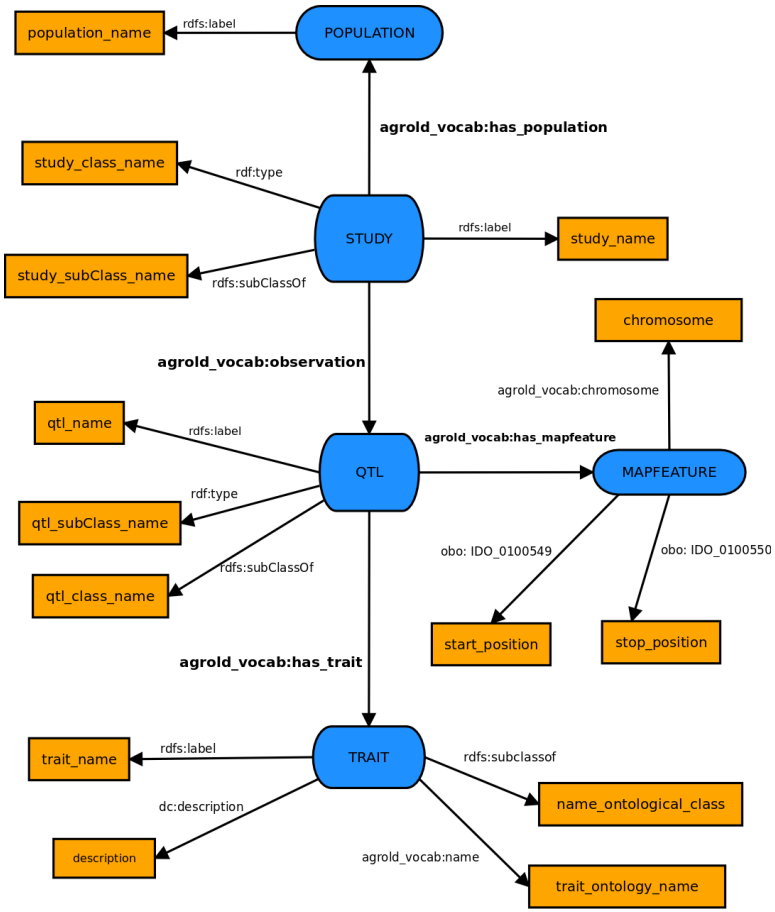
RDF modeling of the databases schemas

RDF triple store storage

Query of dispersed data for data integration through Semantic web services (SPARQL, web user interface)

From N. El Hassouni

Le Modèle RDF présente l'organisation générale des données.

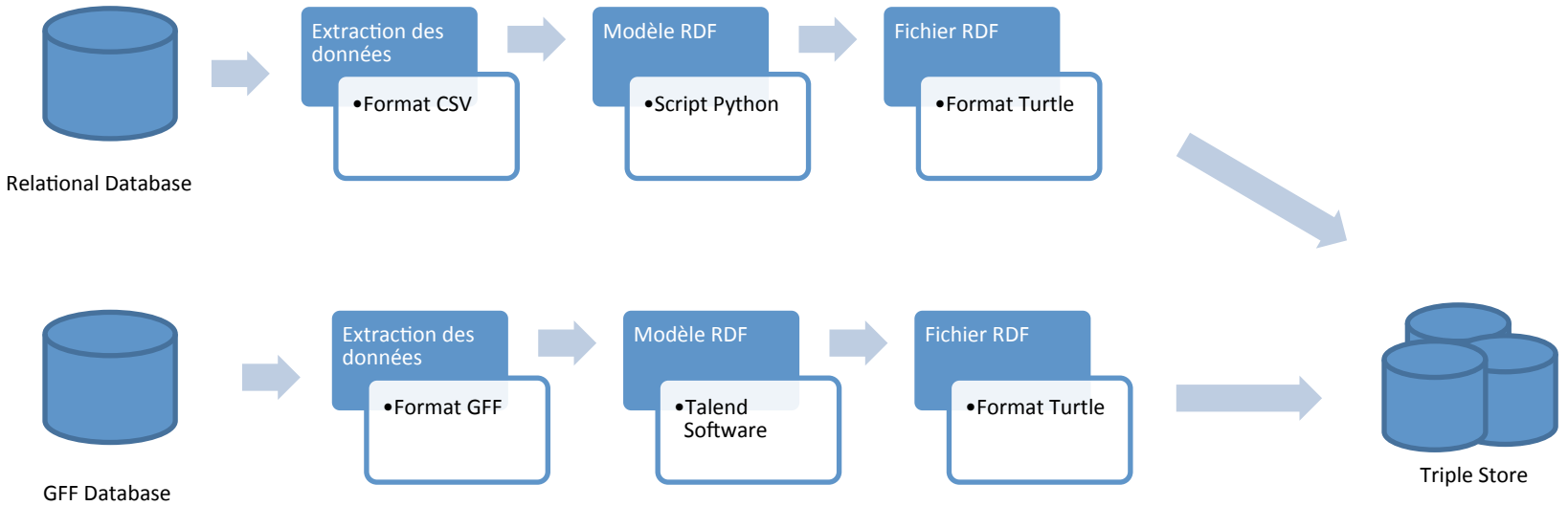


Un Graphe nommé est un concept clé de l'architecture du Web sémantique dans lequel un ensemble de **ressources déclarées sont identifiées en utilisant une URI**.

Exemple d'URI:
http://www.southgreen.fr/tropgene.trait/35

From N. El Hassouni

Démarche



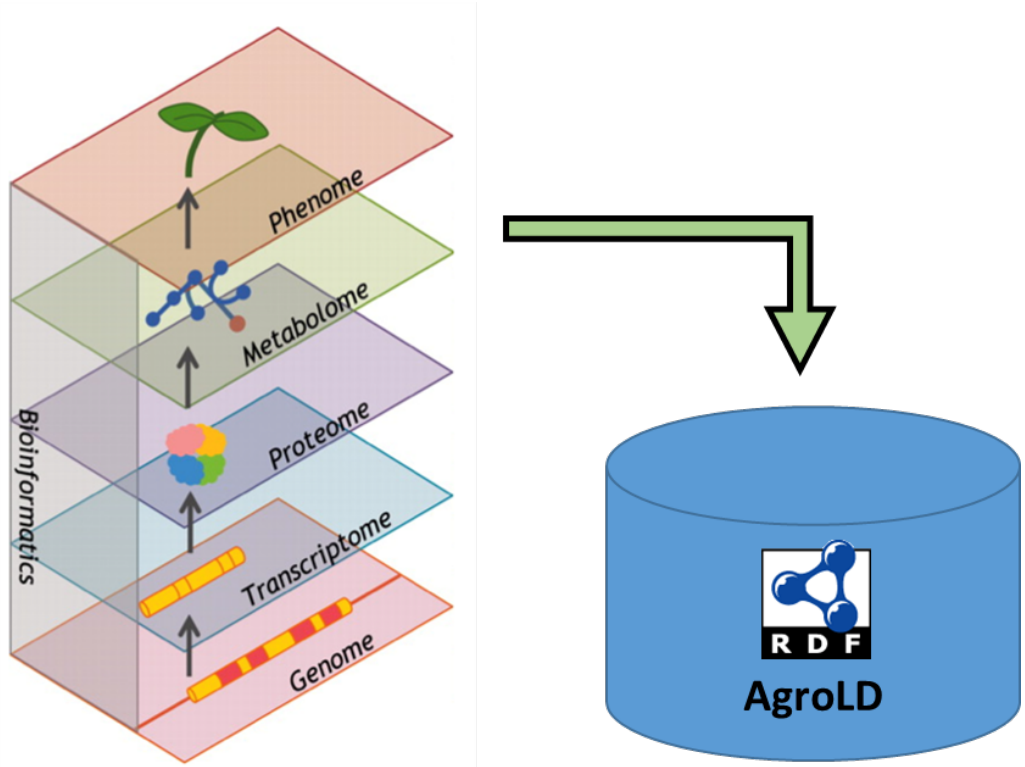
...

Pipelines d'extraction des données et de transformation en RDF

From N. El Hassouni

Agronomic Linked Data (AgroLD)

- RDF knowledge base that integrates data from a variety of plant resources.
- Integrate information at different levels.



AgroLD
(www.agrold.org)

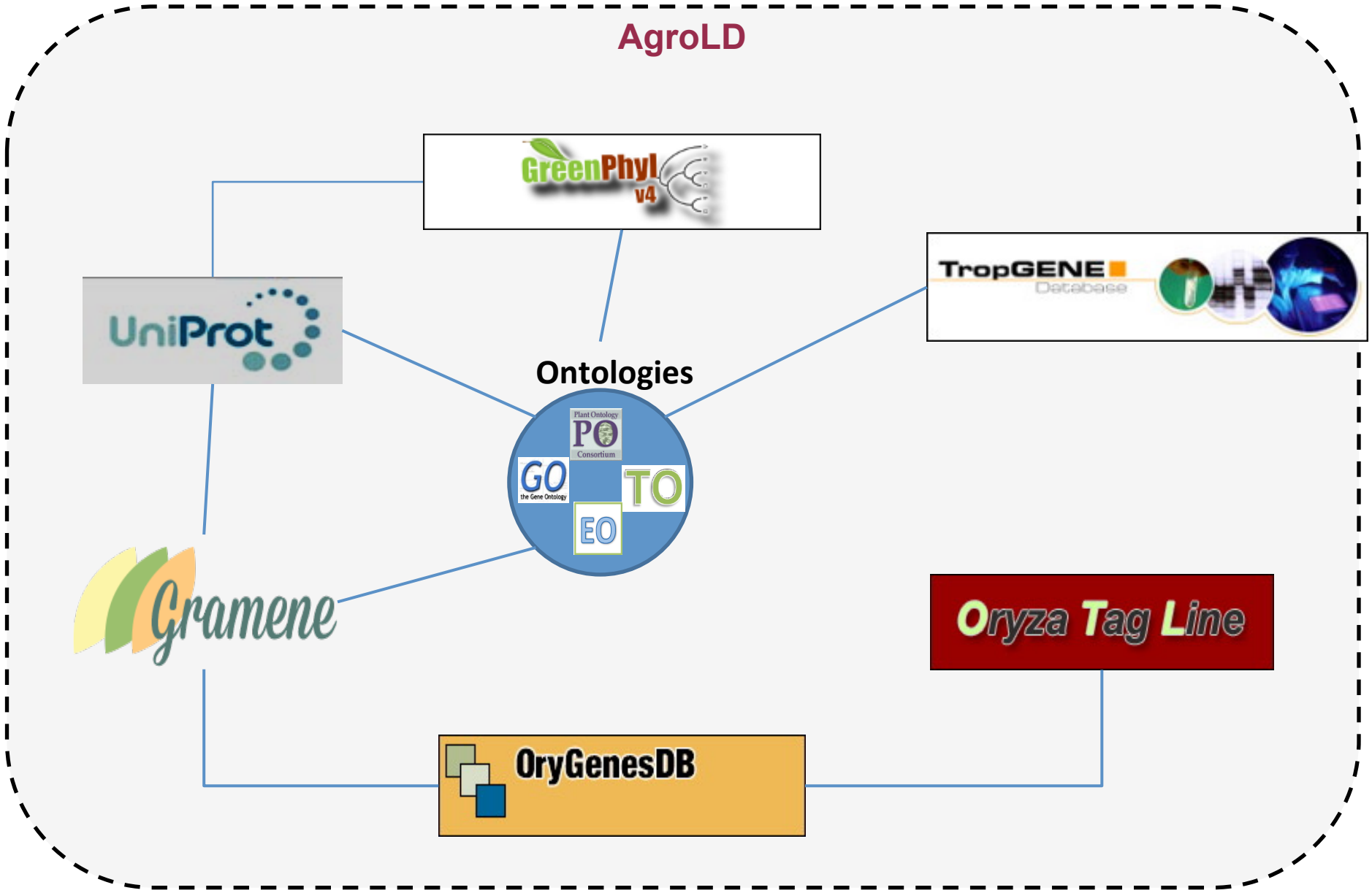
- AgroLD is developed in phases:

- Phase I: includes information on:
  - **Arabidopsis thaliana**
  - **Wheat (Triticum spp.)**
    - Triticum aestivum
    - Triticum urartu
  - **Rice (Oryza spp.)**
    - Oryza barthi
    - Oryza brachyantha
    - Oryza Sativa
    - Oryza glaberimma
  - **Sorghum (Sorghum bicolor)**
  - **Maize/Corn (Zea mays)**

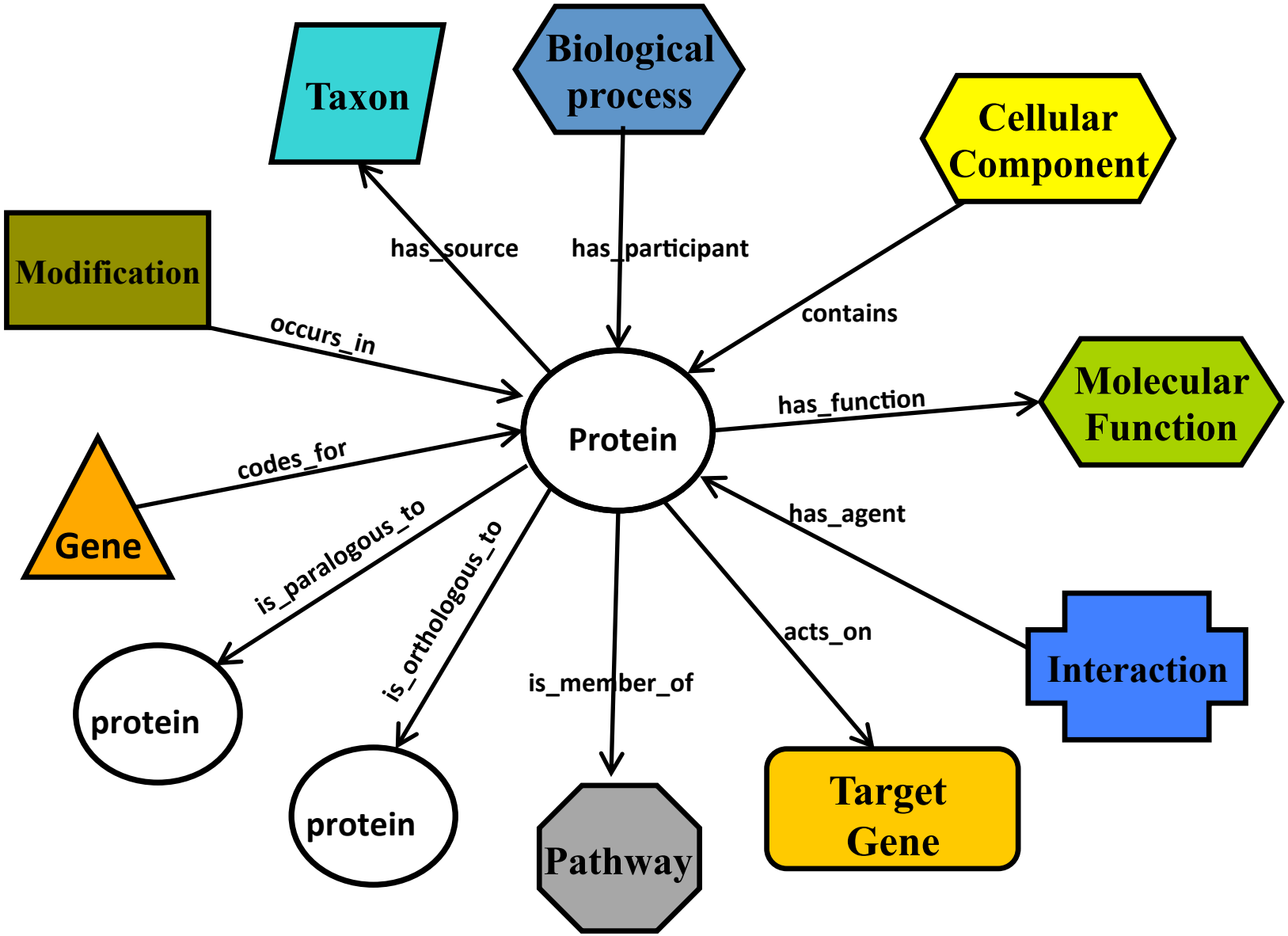
Information in AgroLD

- Integrates information from:
  - **Ontologies:** Gene Ontology (GO), Sequence Ontology (SO), Plant Ontology (PO), Plant Trait Ontology (TO), Plant Environment Ontology (EO), NCBI Taxonomy
  - **Information sources:**
    - **Ontology association:** GOA, Gramene (TO, PO and EO)
    - **Gene/Protein information:** OryGenesDB, Gramene, UniPort
    - **QTL information:** TropGeneDB, Gramene
    - **Pathway information:** Gramene - Cyc
    - **Phenotype information:** Oryza Tag Line
    - **Homology prediction:** GreenPhylDB

Knowledge in AgroLD



Knowledge representation in AgroLD



Agronomic Linked Data (AgroLD)

Home | Search | Documentation | About | Please send us your feedback!

The Agronomic Linked Data (AgroLD) Project

At the Institute of Computational Biology (IBC), we are involved in developing methods to aid data integration and knowledge management within the plant biology domain to improve information accessibility of heterogeneous data. Among others, a solution for the data integration challenges is offered by the Semantic Web technologies. The semantic web has emerged as one of the most promising solutions for high scale integration of distributed resources. This is made possible by a stack of technologies such as the Resource Description Framework (RDF), RDF Schema (RDFS), Web Ontology Language (OWL) and the SPARQL Query Language (SPARQL) proposed by the World Wide Web Consortium (W3C). RDF forms the basis of the stack allows modeling information as a directed graph composed of triples that can be queried using SPARQL.

AgroLD is a RDF knowledge base that consists of data integrated from a variety of plant resources and ontologies. The aim of the Agronomic Linked Data (AgroLD) project is to provide a portal for bioinformatics and domain experts to exploit the homogenized data models towards efficiently generating research hypotheses.

Quick Search

Search with keywords and browse AgroLD Knowledge Base

Advanced Search

Search with keywords, browse, and get answers to some biological questions

Explore Relationships

Search easily existing relationships between entities

SPARQL Query Editor

Edit and submit your SPARQL Queries to the sparql endpoint of AgroLD located at http://volvestre.cirad.fr:8890/sparql

www.agrold.org

Search and browse AgroLD

Search examples: ontological concepts - 'plant height' or 'regulation of gene expression'; gene names - 'GRP2' or 'TCP12'.

Search Text **Plant height** Search

Entity	Title	Named Graph	
http://www.identifie...gramene.qtl/AQFS209	PTHT	http://www.southgree...old/qtl.annotations	**plant height**.
http://www.identifie...gramene.qtl/AQBK027	BNL6.32	http://www.southgree...old/qtl.annotations	**plant height**.
http://www.identifie...gramene.qtl/AQCN007	BNL6.32	http://www.southgree...old/qtl.annotations	**plant height**.
http://www.southgree...notype/ALGB06_otl_6		http://www.southgreen.fr/agrold/otl	... dark green leaves; wide flag leaves; compact **plant** 1 **plant**.
http://www.southgree...notype/AEVD11_otl_2		http://www.southgreen.fr/agrold/otl	Decreased **height** 30 ; late flowering; compact **plant**; erect leaves 1 **plant**.
http://www.identifie...gramene.qtl/AQEA058	BNL6.32	http://www.southgree...old/qtl.annotations	**plant height**.
http://www.identifie...gramene.qtl/AQEA071	BNL6.32	http://www.southgree...old/qtl.annotations	**plant height**.
http://www.identifie...ramene.qtl/AQFS1461	PTHT	http://www.southgree...old/qtl.annotations	**plant height**.
http://www.identifie...gramene.qtl/AQFS183	PTHT	http://www.southgree...old/qtl.annotations	**plant height**.

OpenLinks Faceted search

String matching on literal + ranking based on occurrence number

Sparql query editor

```
BASE <http://www.southgreen.fr/agrold/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX obo:<http://purl.obolibrary.org/obo/>
PREFIX uniprot:<http://purl.uniprot.org/uniprot/>
PREFIX vocab:<vocabulary/>
PREFIX graph:<gramene.cyc>
PREFIX pathway:<biocyc.pathway/CALVIN-PWY>

SELECT DISTINCT ?gene ?name ?taxon_name
WHERE {
 GRAPH graph: {
  ?gene vocab:is_agent_in pathway:.
  ?gene rdfs:label ?name.
  ?gene vocab:taxon ?taxon_name.
 }
```

Execution timeout 20000 milliseconds *(values less than 1000 are ignored)* Results Format RDF/XML Download Results

Filename to Save As: query.sparql Save Query Choose File No file chosen Load Selected Query File

2. Search terms by label (select)
3. List relation types in a given graph (select)
4. Retrieve the local neighbourhood of Oryza sativa japonica protein: **IAA16** - Auxin-responsive protein (UniProt accession:P0C127) (select)
5. Identify Wheat proteins that are involved in root development. (select)
6. Retrieve genes that participate in a given pathway: **Calvin cycle** (select)
7. Retrieve Proteins associated with a given QTL: **DTHD** (days to heading) (select)
8. Get the ID corresponding to the ontology term "**homoaconitate hydratase activity**" (select)
9. Get the name of the ontological element that has the ID "**GO:0003824**" (select)
10. Get the level **4** ancestor of **GO:0004409** (select)
11. Get the level **2** descendence of **GO:0003824** (select)
12. Get protein ids associated with the ontological id **GO:0003824** (select)
13. Get QTL ids associated with the ontological id **EO:0007403** (select)
14. Describe **uniprot:P0C127** (select)

Results

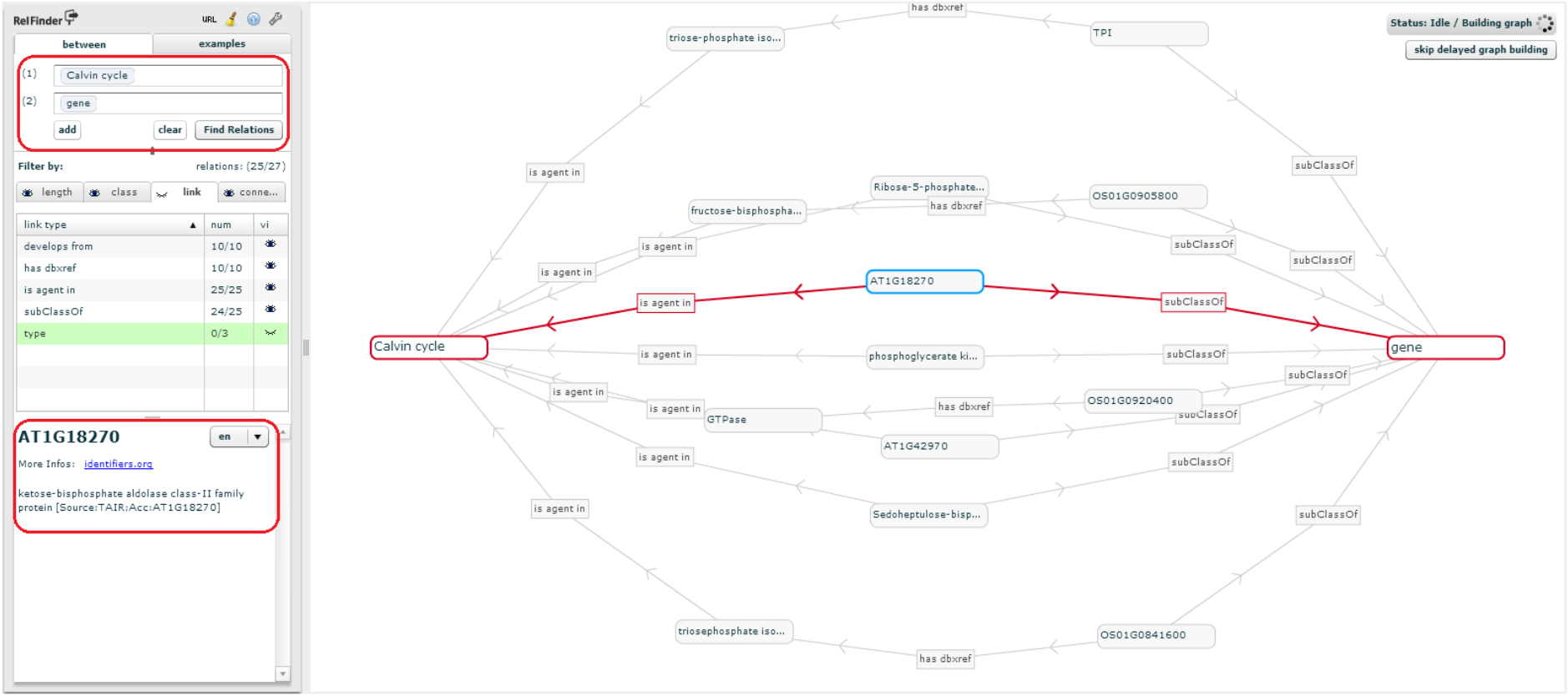
Raw Response | Table | Pivot Table

Search: Show 50 entries

	gene	name	taxon_name
1	http://identifiers.org/ensembl.plant/AT1G18270	fructose-bisphosphate aldolase	obo:NCBITaxon_3702
2	http://identifiers.org/ensembl.plant/AT1G42970	glyceraldehyde-3-phosphate dehydrogenase	obo:NCBITaxon_3702
3	http://identifiers.org/ensembl.plant/AT1G43670	fructose-1,6-bisphosphatase	obo:NCBITaxon_3702

Visualisation of queries

Search > Explore



Based on RelFinder [Heim et al, 2009]

Web Services API

AgroLD API 1.0 Interactive Documentation

This page provides information about the RESTful API (version 1.0) of AgroLD.

More documentations
Find out more about Swagger

http://swagger.io

Contact the developer

IBC Montpellier

gene : Services over genes — Show/Hide | List Operations | Expand Operations

POST	/genes{_format}	Retrieve complete URI and description of all genes from AgroLD in JSON format
POST	/genes/publications/byId	Get publications of a gene
POST	/genes/byKeyword{_format}	Retrieve genes with the URI or the name or the description containing the given keyword
POST	/genes/encodingProtein{_format}	Get URIs, ids, and name of genes encoding a protein given its ID
POST	/genes/participatingInPathway{_format}	Get URIs, ids, and name of genes participating in a pathway given its ID

graphs : General Services — Show/Hide | List Operations | Expand Operations

ontologies : Services over ontologies used in AgroLD — Show/Hide | List Operations | Expand Operations

pathway : Services over pathways — Show/Hide | List Operations | Expand Operations

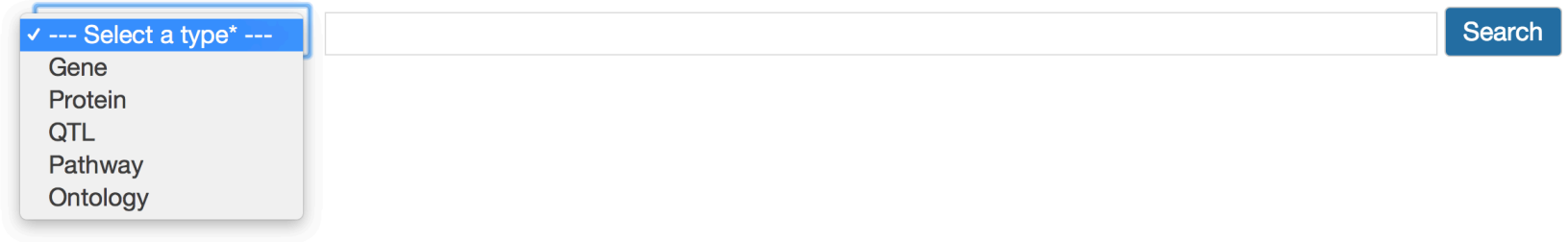
protein : Services over proteins — Show/Hide | List Operations | Expand Operations

qtl : Services over QTLs — Show/Hide | List Operations | Expand Operations

Advanced form-based search

Search examples: ontological concepts - 'plant height' or 'regulation of gene expression'; gene names - 'GRP2' or 'TCP2'.

QTL ID: 'AQAA003' ; protein name: 'TBP1'



Search **protein** with keyword " **TBP1** "

Search: Show 30 entries

	Id	Name	Description	URI
1	Q9LL45 **(display)**	TBP1	Telomere-binding protein 1	http://purl.uniprot.org/uniprot/Q9LL45 **(in Sparql)**
2	Q9LL45 **(display)**	TBP1_ORYSJ	Telomere-binding protein 1	http://purl.uniprot.org/uniprot/Q9LL45 **(in Sparql)**
3	AT3G13445.1 **(display)**	"AT3G13445.1"^^xsd:string	"| Symbols: TBP1, TFIID-1 | TATA binding protein 1 | chr3:4380317-4381869 FORWARD LENGTH=200"^^xsd:string	http://www.southgreen.fr/agrold/greenphyl.sequence/AT3G13445.1 **(in Sparql)**
4	P46465 **(display)**	TBP1	26S protease regulatory subunit 6A homolog	http://purl.uniprot.org/uniprot/P46465 **(in Sparql)**

Showing 1 to 4 of 4 entries

Results are combined with external services

PROTEIN : Q9LL45 / TBP1

Telomere-binding protein 1

URI: *http://purl.uniprot.org/uniprot/Q9LL45*

is encoded by

Search: Show 30 entries

	Id	Name	Description	URI
1	OS02G0817800 **(display)**	TBP1	Telomere-binding protein 1 [Source:UniProtKB/Swiss-Prot;Acc:Q9LL45]	http://identifiers.org/ensembl.plant/OS02G0817800 **(in Sparql)**

Showing 1 to 1 of 1 entries

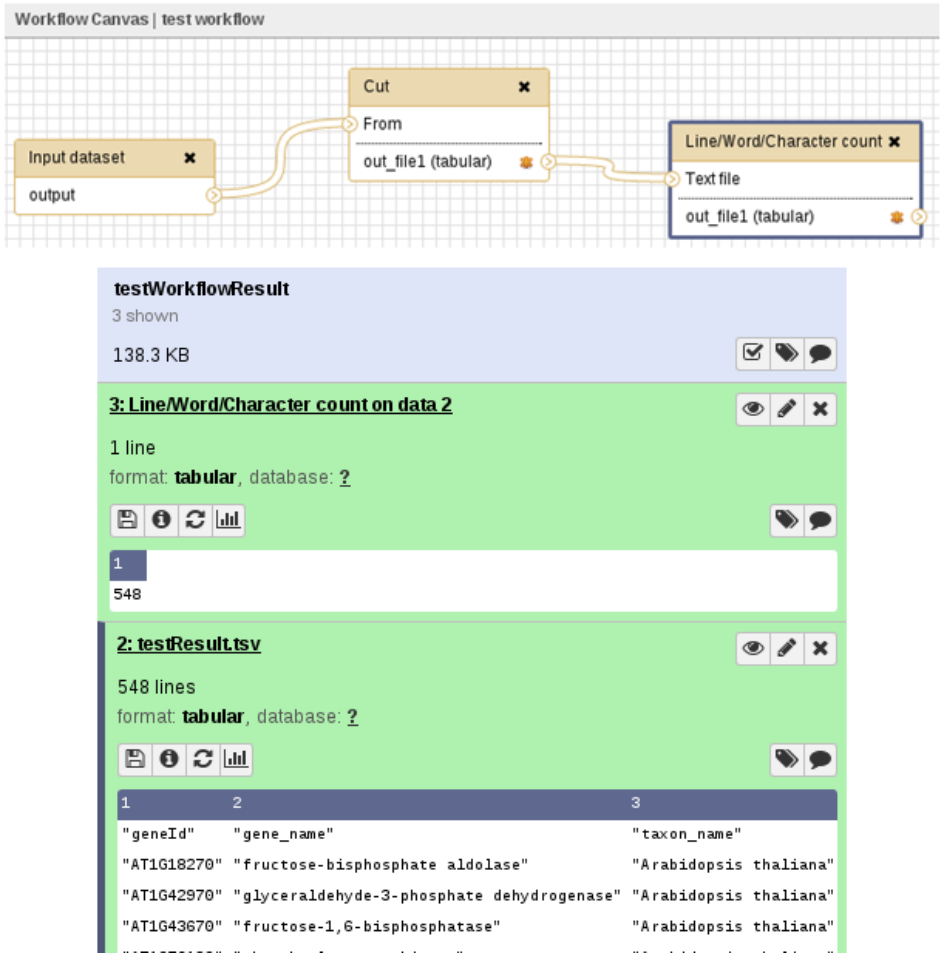
QTL associations +

Ontology associations +

Publication

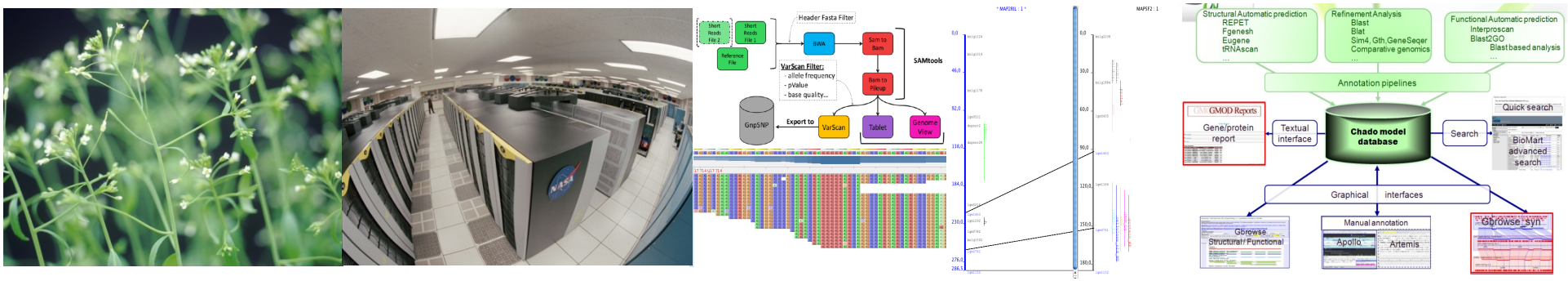
1. Yu EY, Kim SE, Kim JH, Ko JH, Cho MH, Chung IK., " **Sequence-specific DNA recognition by the Myb-like domain of plant telomeric protein RTBP1.** ", *J Biol Chem*, 2000
   More at: http://www.ncbi.nlm.nih.gov/pubmed/10811811
2. International Rice Genome Sequencing Project., " **The map-based sequence of the rice genome.** ", *Nature*, 2005
   More at: http://www.ncbi.nlm.nih.gov/pubmed/16100779
3. ice Annotation Project, Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, ..., " **The Rice Annotation Project Database (RAP-DB): 2008 update.** ", *Nucleic Acids Res*, 2008
   More at: http://www.ncbi.nlm.nih.gov/pubmed/18089549
4. Hong JP, Byun MY, Koo DH, An K, Bang JW, Chung IK, An G, Kim WT., " **Suppression of RICE TELOMERE BINDING PROTEIN 1 results in severe and gradual developmental defects accompanied by genome instability in rice.** ", *Plant Cell*, 2007
   More at: http://www.ncbi.nlm.nih.gov/pubmed/17586654
5. Ko S, Yu EY, Shin J, Yoo HH, Tanaka T, Kim WT, Cho HS, Lee W, Chung IK., " **Solution structure of the DNA binding domain of rice telomere binding protein RTBP1.** ", *Biochemistry*, 2009
   More at: http://www.ncbi.nlm.nih.gov/pubmed/19152316

Galaxy Wrapper available for AgroLD



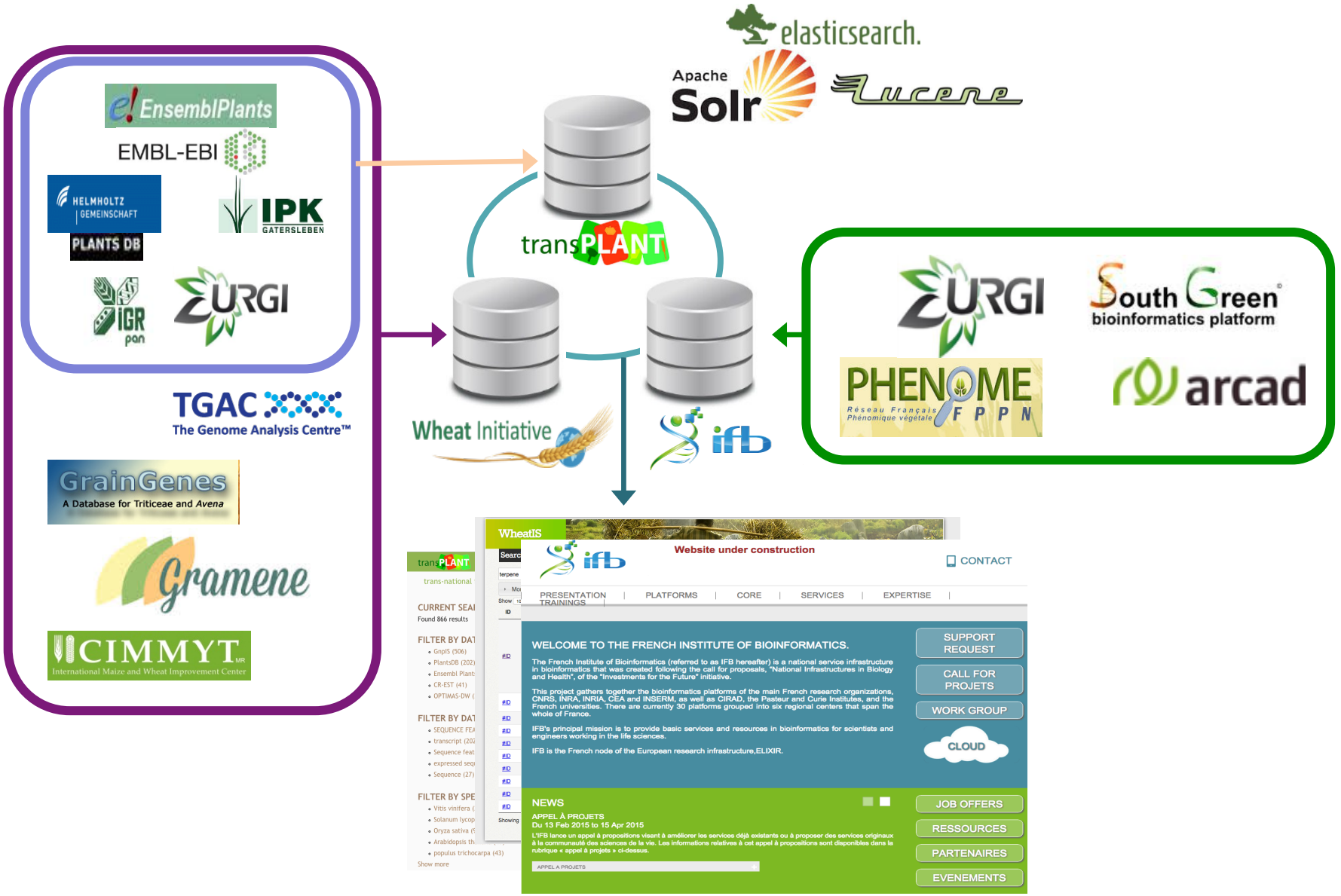
From Gildas Tagny, M2 internship

WP2: Develop an intuitive Google-like search portal

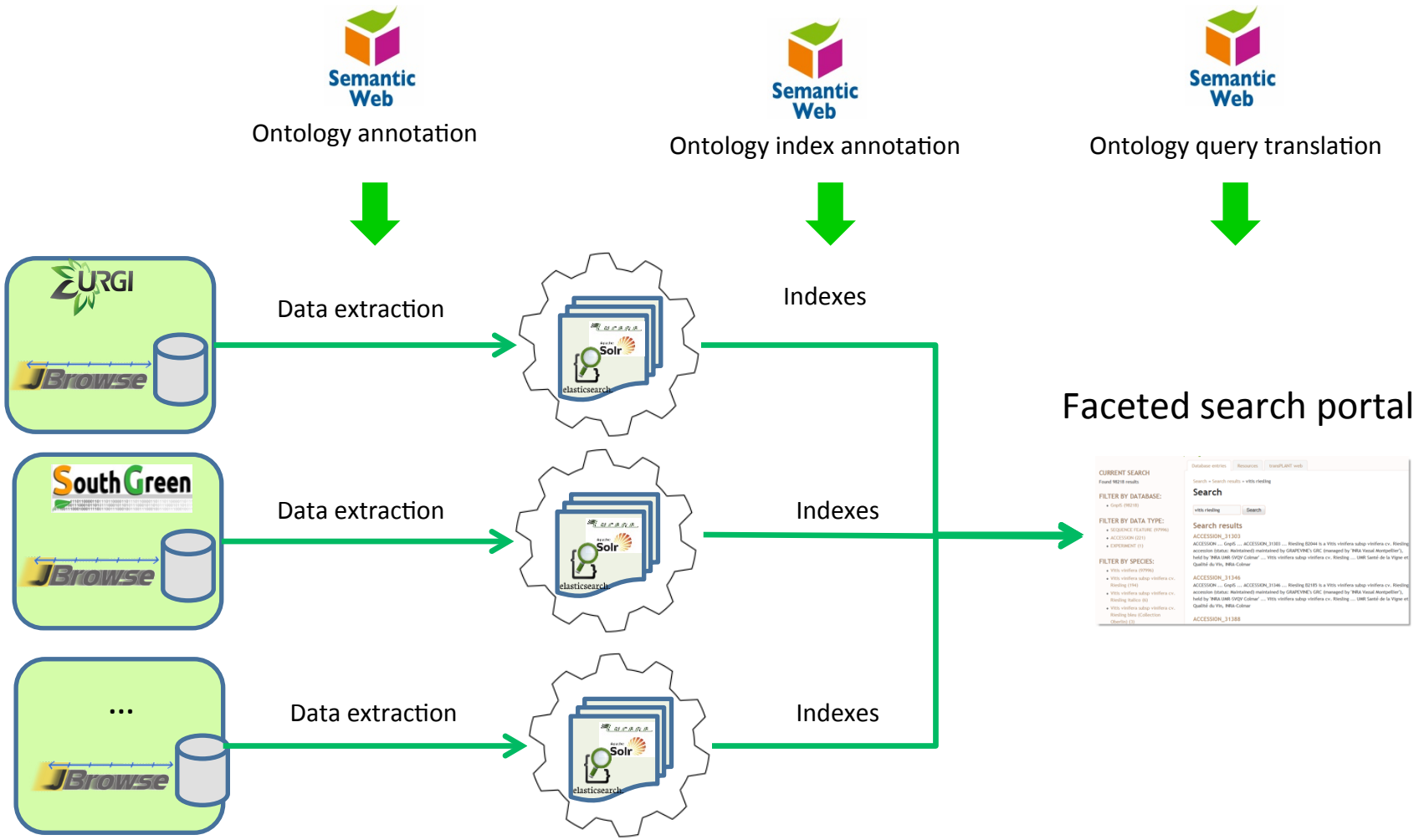


- *Use the very fast distributed indexation system developed in international projects: WheatIS, transplant*
- *Inject Semantic web capabilities in this system*

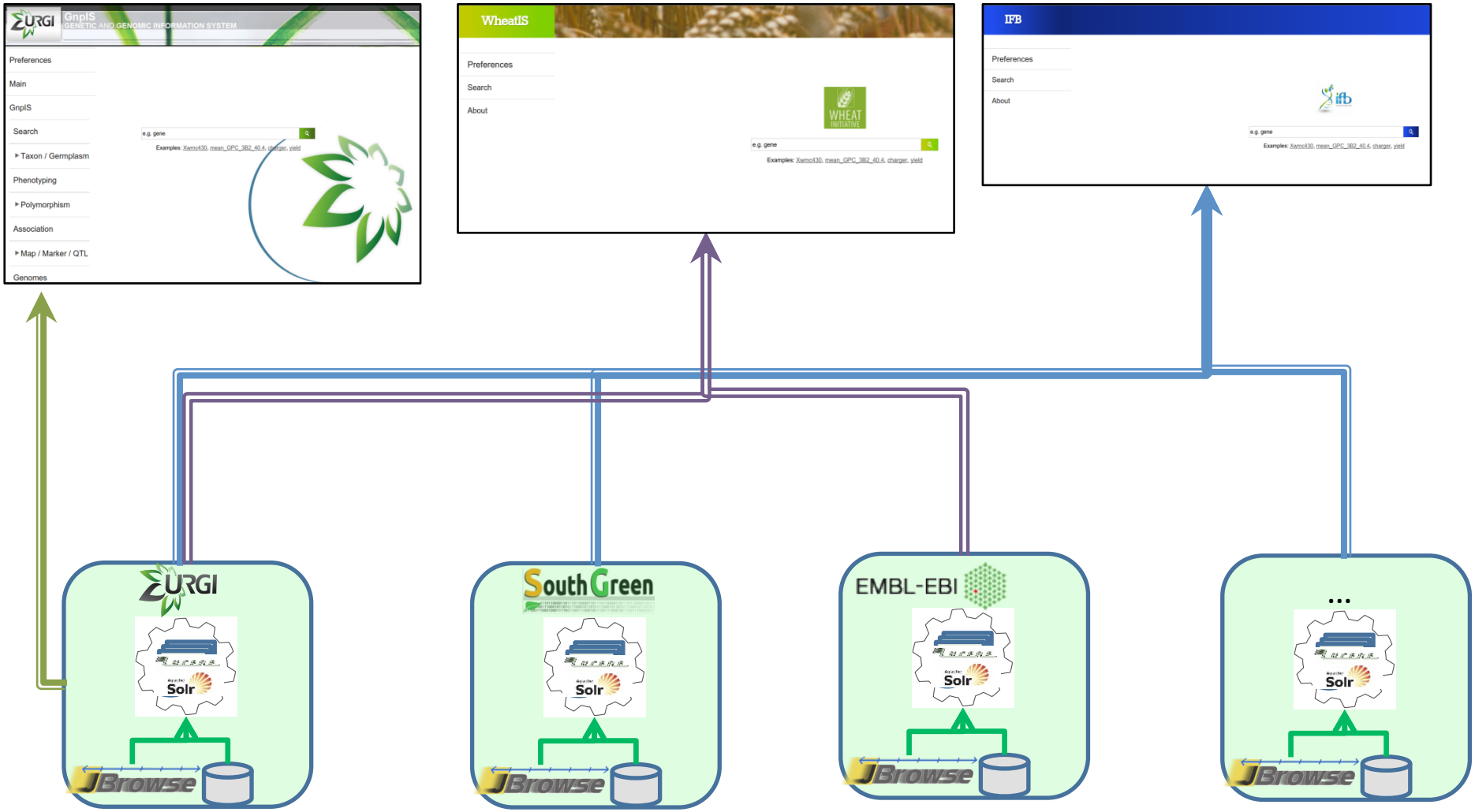
Full text search of distributed databases



Data discovery



WP2 Search portal



Future directions

- Phase II: having both wider and deeper coverage to promote comparative analysis
  - Include varied data types – gene expression data, protein-protein interaction, Transcription factor- target gene

- Developing methods to aid the process of hypotheses generation - e.g. inference rules.

- Query translation – natural language query translation.

- Engage with biologists to mobilise ‘user-pull’:
  - Develop real world use cases – studying the molecular mechanism of panicle differentiation in rice

Acknowledgements

Bioversity International

Elizabeth Arnaud,
Leo Valette,
Marie-Angelique Laporte,
Julian Pietragalla

INRA SCIENCE & IMPACT

Hadi Quesneville
Esther Dzalé-Yeumo,
Cyril Pommier
Florian Philippe

IRD Institut de recherche pour le développement

Pierre Larmande
Alexis Dereeper

cirad

Manuel Ruiz,

Guilhem Sempere,

ifb

Nordine El Hassouni

IBC Institut de biologie computationnelle

Aravind Venkatesan,
Gildas Tagny
Luyen Le Ngoc
Imene Chently

Clement Jonquet
Konstantin Todorov

Université de Montpellier

Contact: pierre.larmande@ird.fr