# Plant biology open data interoperability in the big data era

Anne-Francoise Adam-Blondon, Sophie S. Durand, Erik Kimmel, Raphaël-Gauthier R.-G. Flores, Cyril Pommier, Michael M. Alaux, Delphine Steinbach, Hadi Quesneville

# Plant biology: open data interoperability in the big data era

A-F Adam-Blondon, S Durand, E Kimmel, R Flores, C Pommier, M Alaux, D Steinbach, H Quesneville

EPSO-FEBS 2014, le 27/06/14

# A bioinformatic unit for crops and pathogens


Phenotypes

- GxE
- QTL maps
- GWAS, GS

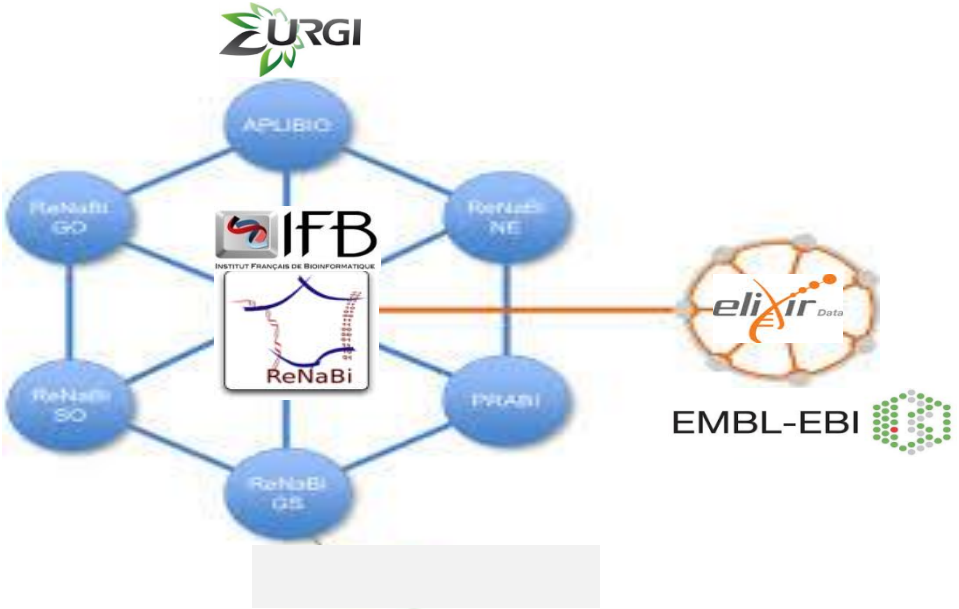
Genetics

- Genetic maps
- Genetic markers
- Genetic resources

Genome

- Annotation
- Transcriptome
- SNP / Structural variant

# URGI is a node of the french network of bio-informatics facilities (IFB-ReNaBi)



# Challenges

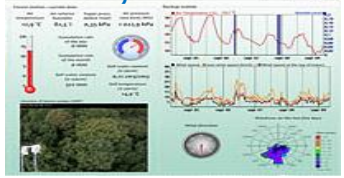
# Necessity to connect data stored into different information systems

- Because the volumes are becoming too big for one information system (Ex: NGS)
- Because it is impossible to store all data in a single data model (Ex: phenotyping)
- Because data relevant for a scientific question may be stored in different databases dedicated to other purposes

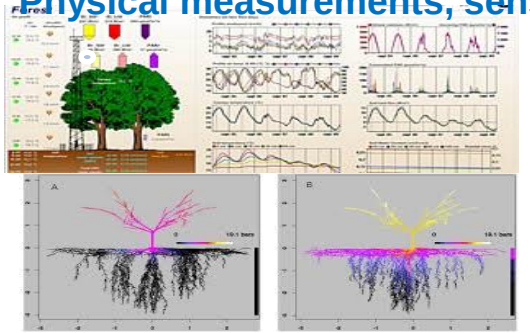
# Necessity to organize and query heterogeneous data collected in different laboratories/context

Different data structures <-> different initial question
Potentially different experimental protocoles

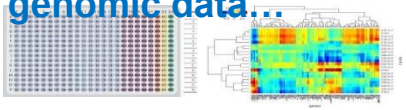
**Climate, environment**



**Physical measurements, sensors…**



**Metabolites, proteins, genomic data…**



**Post-harvest**

**Bibliography, human sciences…**

Development of guidelines, ontologies and standards by the **community of data producers/researchers**

# Consequences on information systems

Towards distributed systems

# Work in progress

# Towards distributed information systems

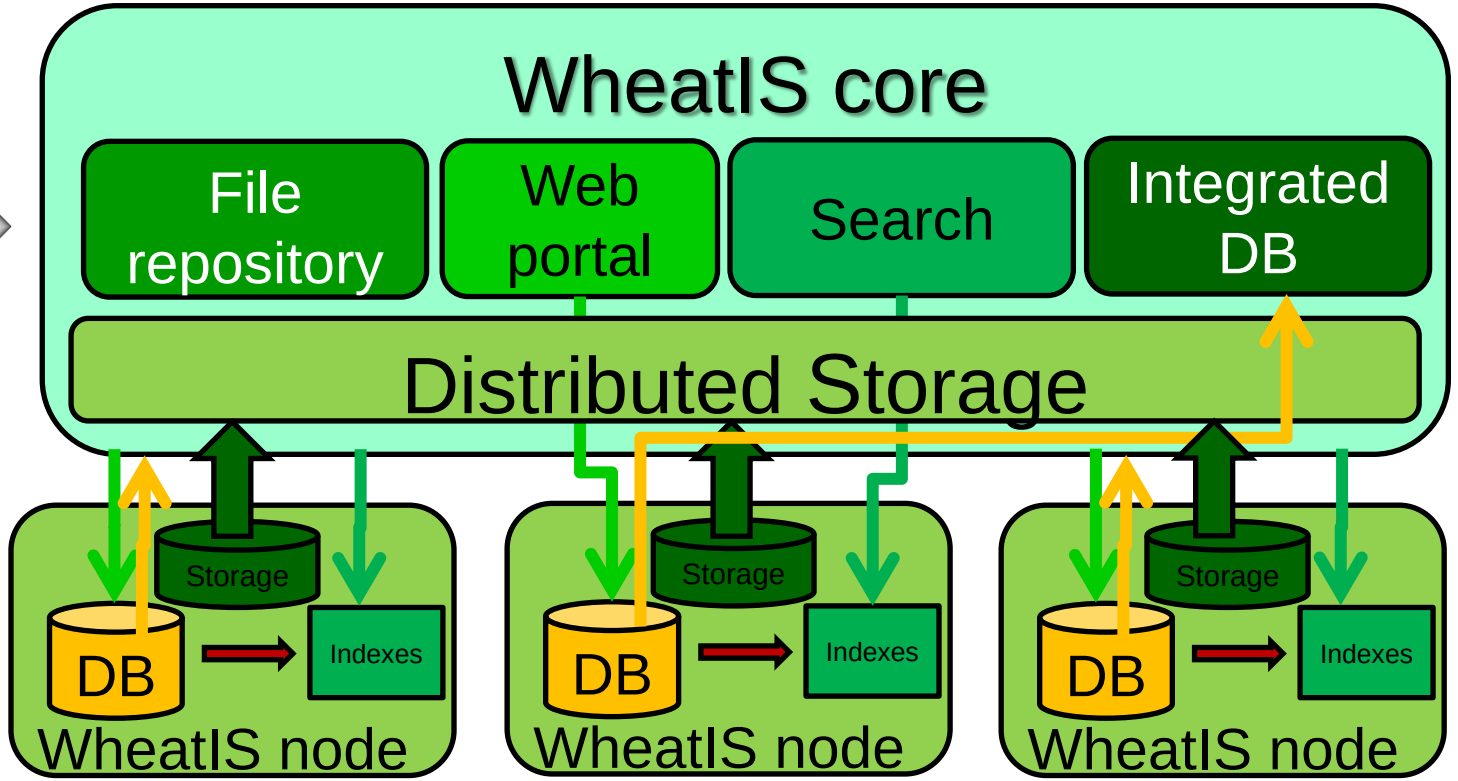
**WheatIS:** the information system of the International Wheat Initiative (coord. H. Lucas): (chair: H. Quesneville)

Google-like query tool allowing to retrieve information in the databases of the transnational **TransPLANT** infrastructure (coord P. Kersey, EBI)

Information system for French Plant Phenotyping Network (**Phenome**, coord F. Tardieu)

Building a portal for the french crop **germplasm collections** (ARCAD-FEDER, J-L Pham coord)

# WheatIS architecture



# Definition of standards

**Survey** of existing standards: **(1) data, (2) ontologies, (3) meta-data**

"**Cookbook**": how to produce easily shareable, reusable and interoperable "wheat data"


Identification of **end-users** categories and **WheatIS nodes**

**Challenge : adoption of the recommendations by the community**

- simple ontologies
- good balance between genericity and necessary specifity
- alignment with other international initiatives
- tools to help users

# Ontologies / Thesaurus

## References Ontologies

**PATO**
- Area
- …

**Plant Ontology**
- Leaf
- …

**Unit Ontology**
- cm2
- …

Plant Ontology PO Consortium

## Applied Ontologies

### Crop Ontology

#### Wheat

**Leaf Area**
- Unit : cm2
- Method

**Yield**
- Unit : t/ha
- Method

**Rust**
- Scale
- Method
- Stage

#### Vitis

**Budbreak date**

**Young Shoot: aperture of tip**
- OIV:001
- Scale:1-3-5
- UPOV:3

Generation Challenge Programme — Cultivating Plant Diversity for the Resource Poor

Bioversity International

## Thesaurus

**BreedWheat**
- Leaf Area
  - **BreedWheat Method**
- Yield

Breed Wheat

**Phenome**
- Leaf Area
- Budbreak date

PHENOME Réseau Français Phénomique végétale FPPN

AnaEE Analysis and Experimentation on Ecosystems France

# Full text queries of distributed databases

GnpIS

EnsemblPlants

EMBL-EBI

MetaCrop

IPK GATERSLEBEN

CR-EST

IPK GATERSLEBEN

GEBIS

IPK GATERSLEBEN

HELMHOLTZ GEMEINSCHAFT

PLANTS DB

PolapgenDB

IGR pan

Lucene indexes

Apache Solr

Lucene

elasticsearch.

Wheat Initiative

Google like query

Google like list of results

http://www.transplantdb.eu/

User web interface


EMBL-EBI

transPLANT trans-National Infrastructure for Plant Genomic Science

Home | Resources | Events | Variation Archive | About

Current search

Search found 971 items

rubisco

Filter by database:

CR-EST (864)

Ensembl Plants (62)

PlantsDB (45)

Filter by data type:

expressed sequence tags (864)

protein_coding (62)

transcript (45)

Filter by species:

Hordeum vulgare (835)

Arabidopsis thaliana (25)

Oryza sativa (21)

Solanum tuberosum (17)

Brassica rapa (13)

Pisum sativum (9)

Musa acuminata (8)

Nicotiana tabacum (8)

Triticum urartu (7)

Chlamydomonas reinhardtii (6)

Aegilops tauschii (5)

Medicago truncatula (5)

Sorghum bicolor (5)

Solanum lycopersicum (3)

Search across plant genomics resources

Enter terms rubisco Search

Search results

Results from PlantsDB:

1. Sb01g000380.1

similar to RuBisCO large subunit-binding protein subunit alpha, chloroplast precursor - ID=Sb01g000380;Description="similar to RuBisCO large subunit-binding protein subunit ... chloroplast precursor" ...

2. Sb09g014430.1

similar to RuBisCO large subunit-binding protein subunit alpha, chloroplast precursor - ID=Sb09g014430;Description="similar to RuBisCO large subunit-binding protein subunit ... chloroplast precursor" ...

More results in PlantsDB...

Results from Ensembl Plants:

1. Bra025431

Bra025431 AT5G38410 (E=6e-092) | ribulose bisphosphate carboxylase small chain 3B ... small subunit 3B (RBCS-3B) (ATS3B) ...

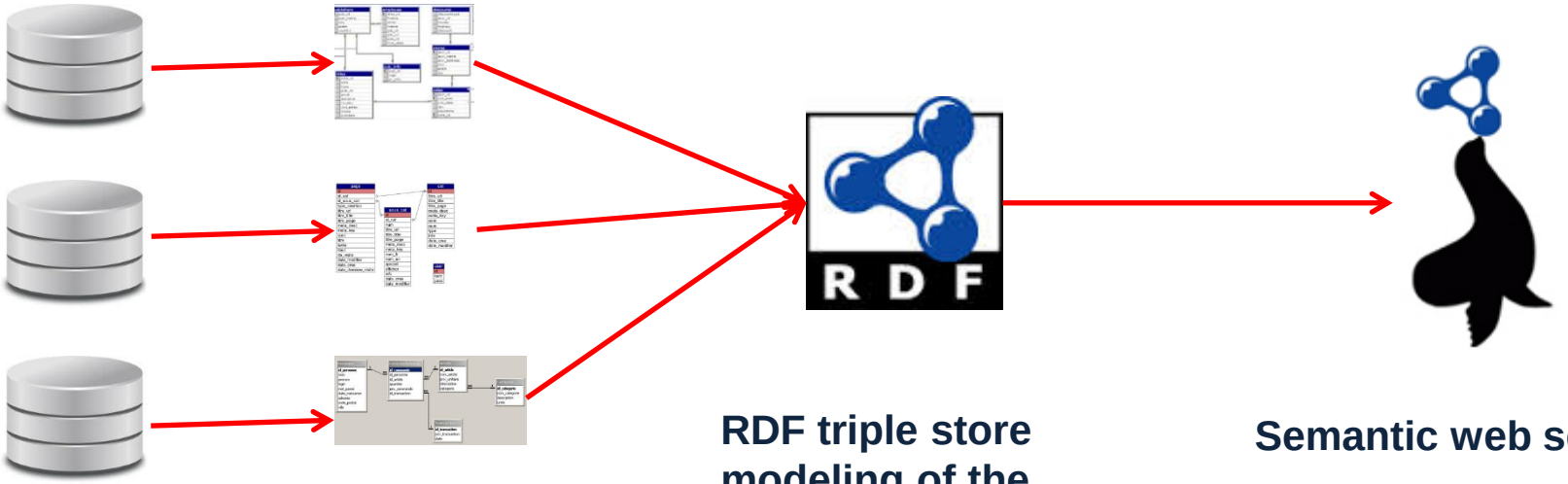
2. Bra028174

Bra028174 AT5G38430 (E=7e-093) | ribulose bisphosphate carboxylase small chain 1B ... small subunit 1B (RBCS-1B) (ATS1B) ...

More results in Ensembl Plants...

# Perspectives

# Develop a web semantic interoperability between the plant databases of the French Elixir node




Ontology based annotation of database schemes

RDF triple store modeling of the databases schemes integrated with existing ontologies

Semantic web services

# Summary

Challenge: the query of high volumes of heterogeneous and distributed data

⇒ Federation of information systems

- ❖ At the national and european level
- ❖ through noSQL technologies (SolR, ElasticSearch…)
- ❖ Web semantic layer

# Aknowledgements

## URGI team

P. Bento
L. Couderc
C. Gageat
A. Keliet
T. Letellier
M. Loaec
A. Ménard
C. Michotey
N. Mohellibi
C. Viseux
S. Meilo

J-F Gibrat
Manuel Ruiz

Jean-Louis Pham
Christophe Jenny
Felix Homa

François Tardieu
Pascal Neveu
Jacques LeGouis
Eric Duchêne

Hélène Lucas
WheatIS Expert Working Group
Mario Caccamo
Dave Edwards
Gerard Lazo…

Esther Dzalé Kaboré
Sophie Aubin

Paul Kersey
Dan Bolser…

Financial supports

# WheatIS timeline