



HAL
open science

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

► To cite this version:

Anne-Francoise Adam-Blondon, Sophie S. Durand, Erik Kimmel, Raphaël-Gauthier R.-G. Flores, Cyril Pommier, et al.. Plant biology open data interoperability in the big data era. Plant Biology Europe EPSO/FEPBS 2014, Jun 2014, Dublin, Germany. hal-02801067

HAL Id: hal-02801067

<https://hal.inrae.fr/hal-02801067v1>

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

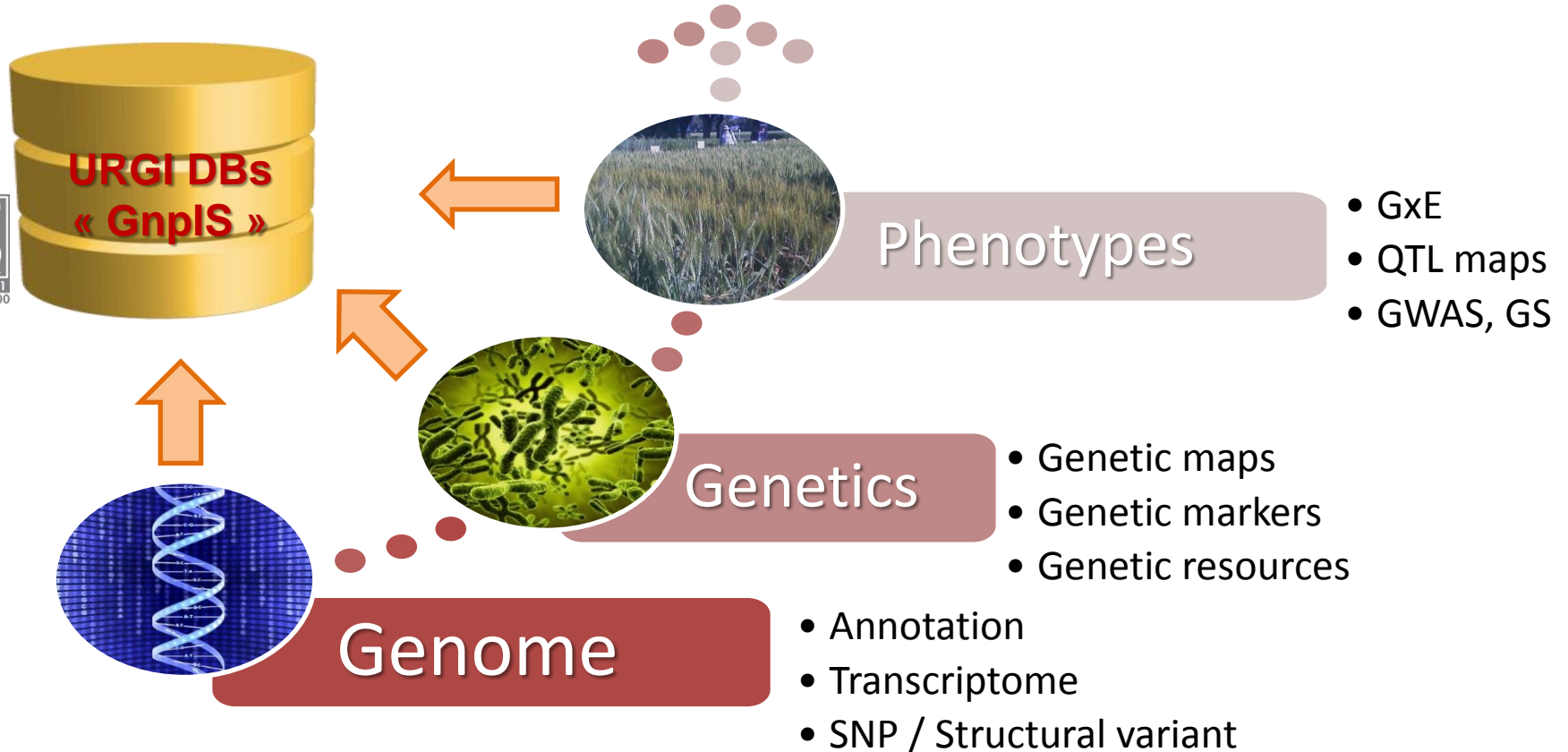


Plant biology: open data interoperability in the big data era

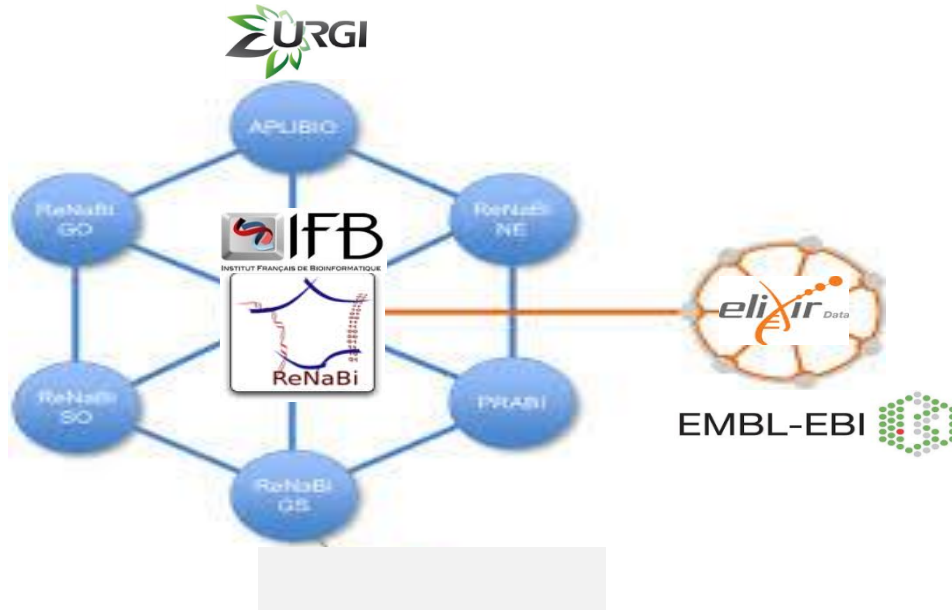
A-F Adam-Blondon, S Durand, E Kimmel, R Flores, C Pommier, M Alaux,
D Steinbach, H Quesneville



A bioinformatic unit for crops and pathogens



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



Challenges

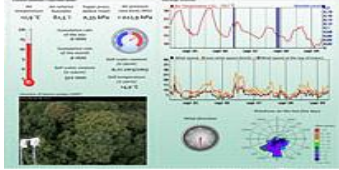
Necessity to connect data stored into different information systems

- 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

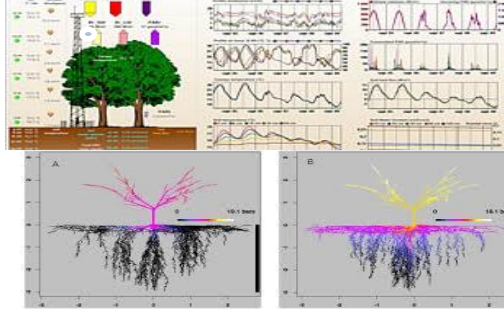
Necessity to organize and query heterogeneous data collected in different laboratories/context

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

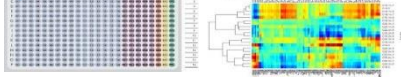
Climate, environment



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)

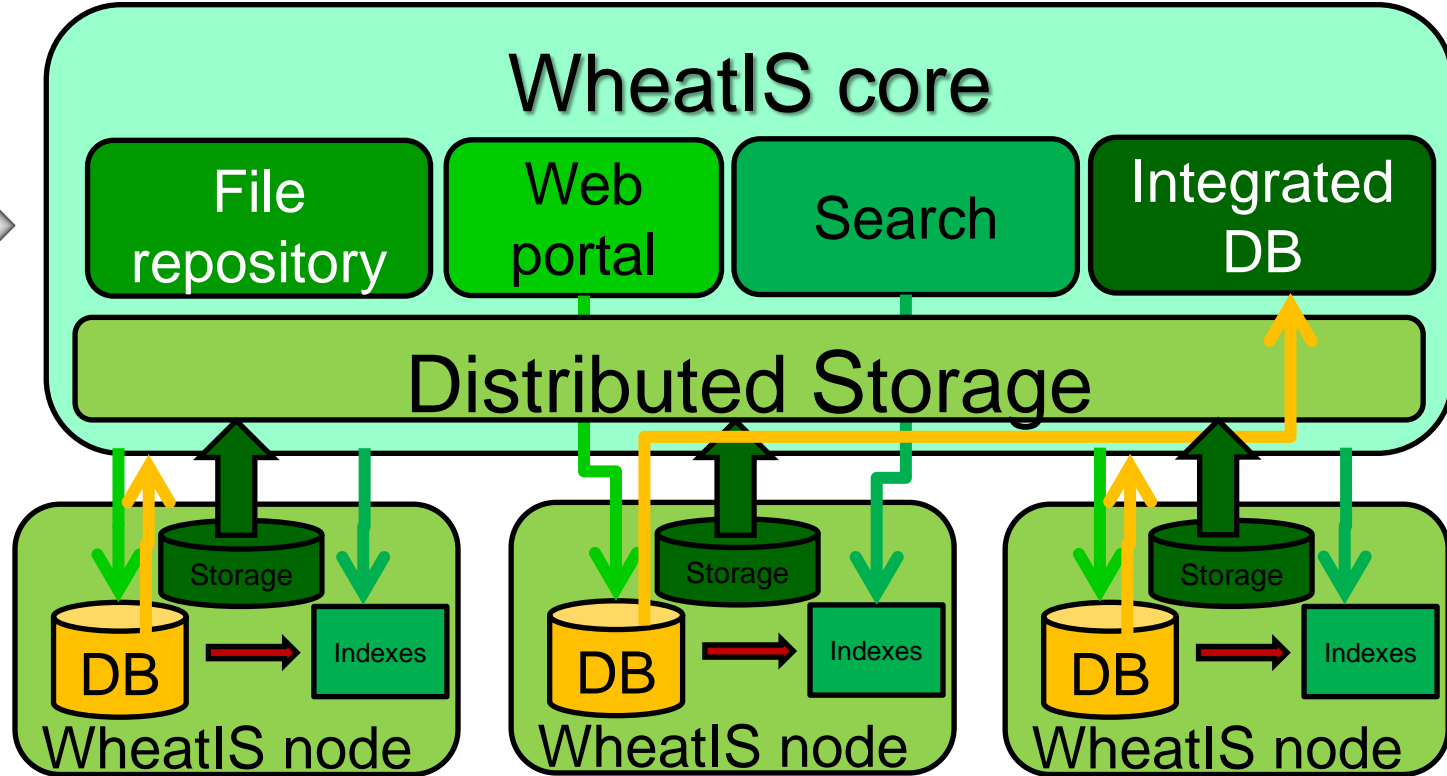
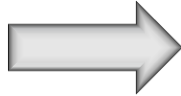


Information system for French Plant Phenotyping Network (**Phenome**, coord F. Tardieu)



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 specificity
- alignment with other international initiatives
- tools to help users

Ontologies / Thesaurus

References Ontologies

PATO

- Area
- ...

Plant Ontology

- Leaf
- ...

Unit Ontology

- cm2
- ...



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

Thesaurus

BreedWheat

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

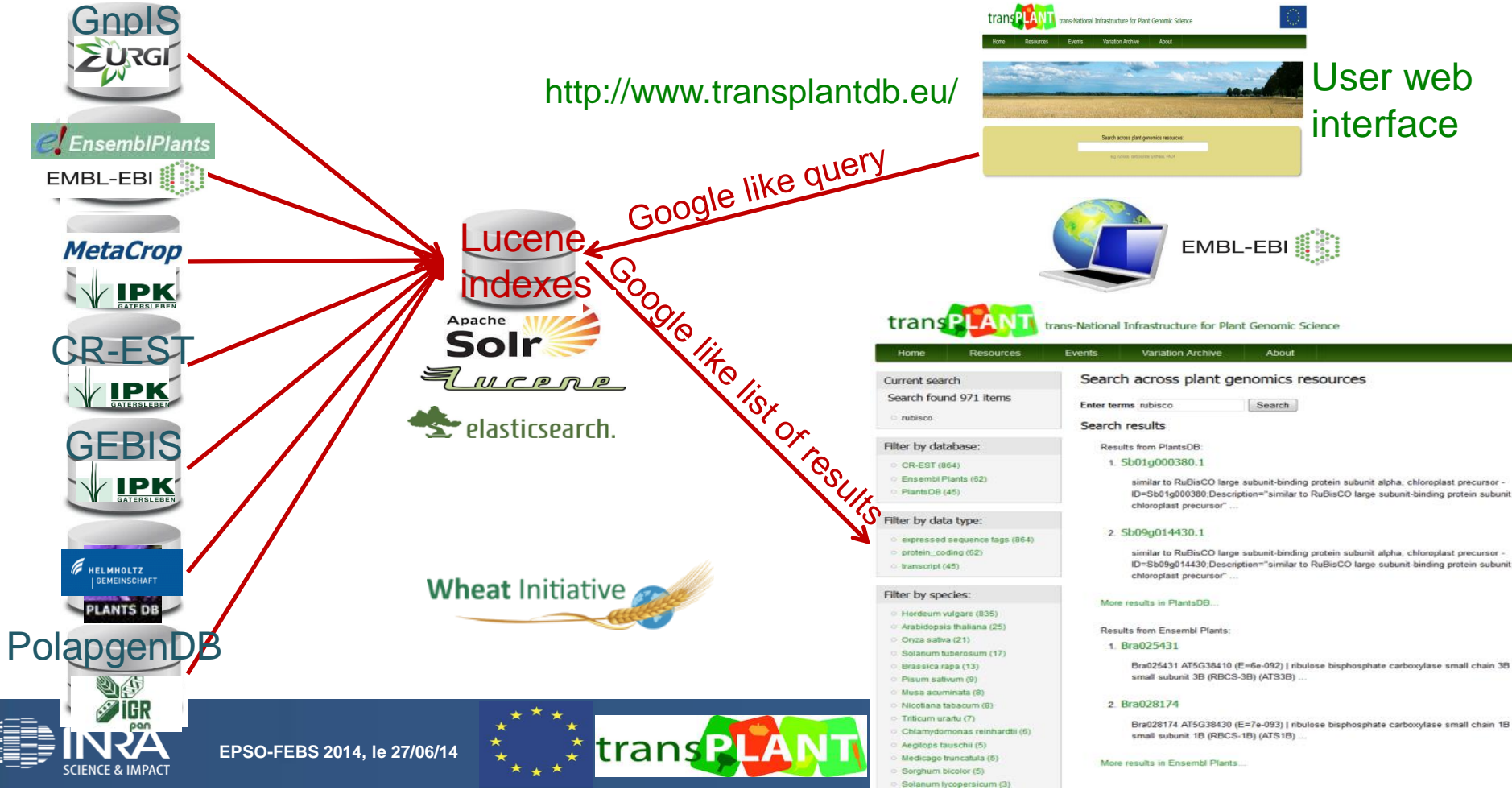


Phenome

- Leaf Area
- Budbreak date

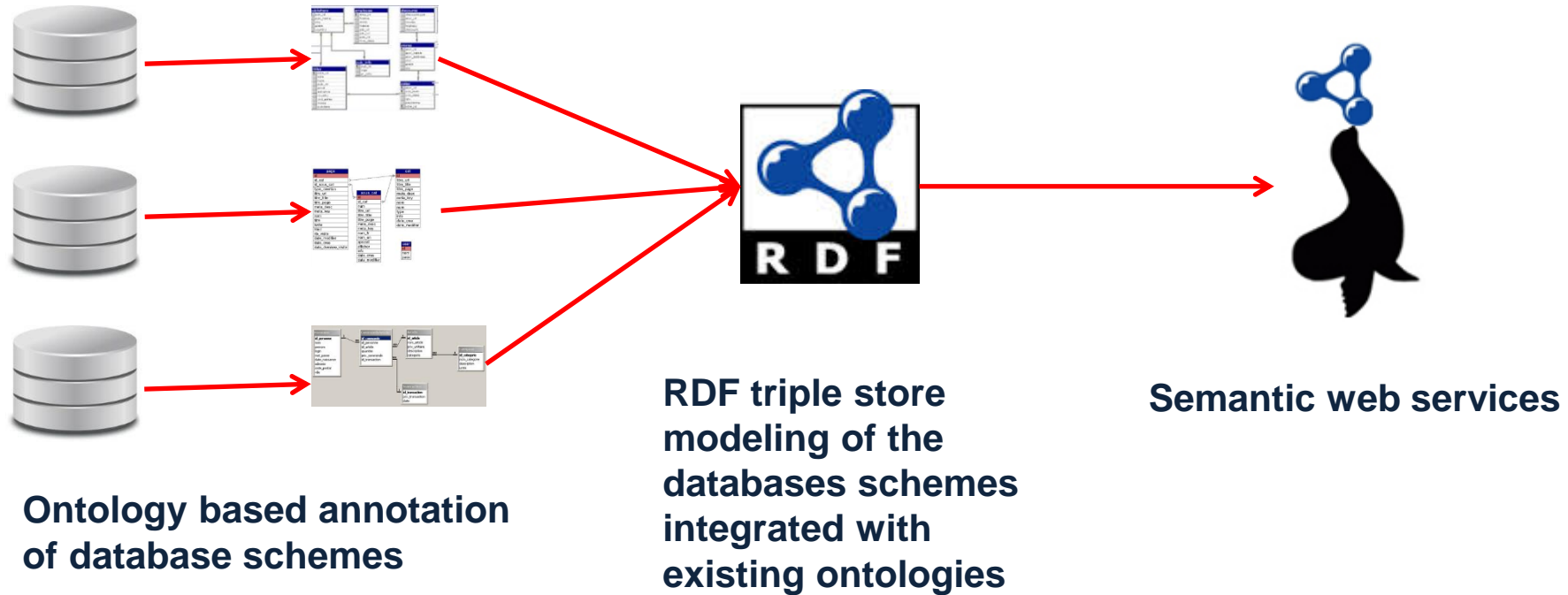


Full text queries of distributed databases



Perspectives

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



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

Financial supports



Wheat Initiative



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



Esther Dzalé Kaboré
Sophie Aubin



Paul Kersey
Dan Bolser...

WheatIS timeline

Step 1: Network building

Definition of standards

- Define standards, nomenclature, formats.
- Meta-data exchange

WheatIS
=
A web platform to
exchange data

Step 2: Integrated portal

Search of data

- DBs federation
- Google-like search

WheatIS
=
A portal to access a
network of DBs

Step 3: Integrated DB

Integration of data

- In one place
- Focused on relevant data sets
- Consolidated and consistent

WheatIS
=
A integrated DB