Plant Cyber-infrastructures: wheatIS & French elixir node

Hadi Quesneville

▶ **To cite this version:**

Hadi Quesneville. Plant Cyber-infrastructures: wheatIS & French elixir node. BIOTALENT, May 2017, Poznan, Poland. hal-02786556

HAL Id: hal-02786556
https://hal.inrae.fr/hal-02786556

Submitted on 5 Jun 2020

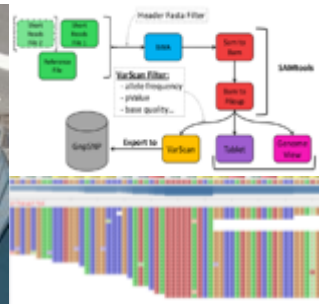
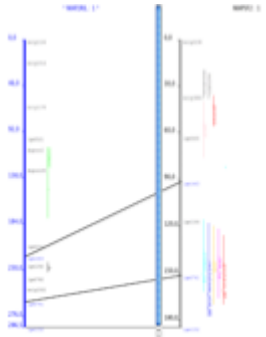
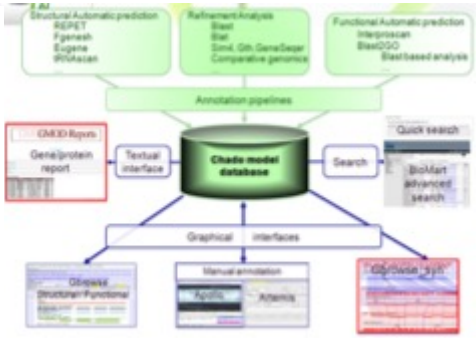
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

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 Cyber-infrastructures: WheatIS & French Elixir node

Hadi Quesneville

Biotalent, May 29, Poznan

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

What is a cyber-infrastructure

A data infrastructure

- Collecte/Store/Manage
- Integrate/Visualize
- Explore/Mine
- Process

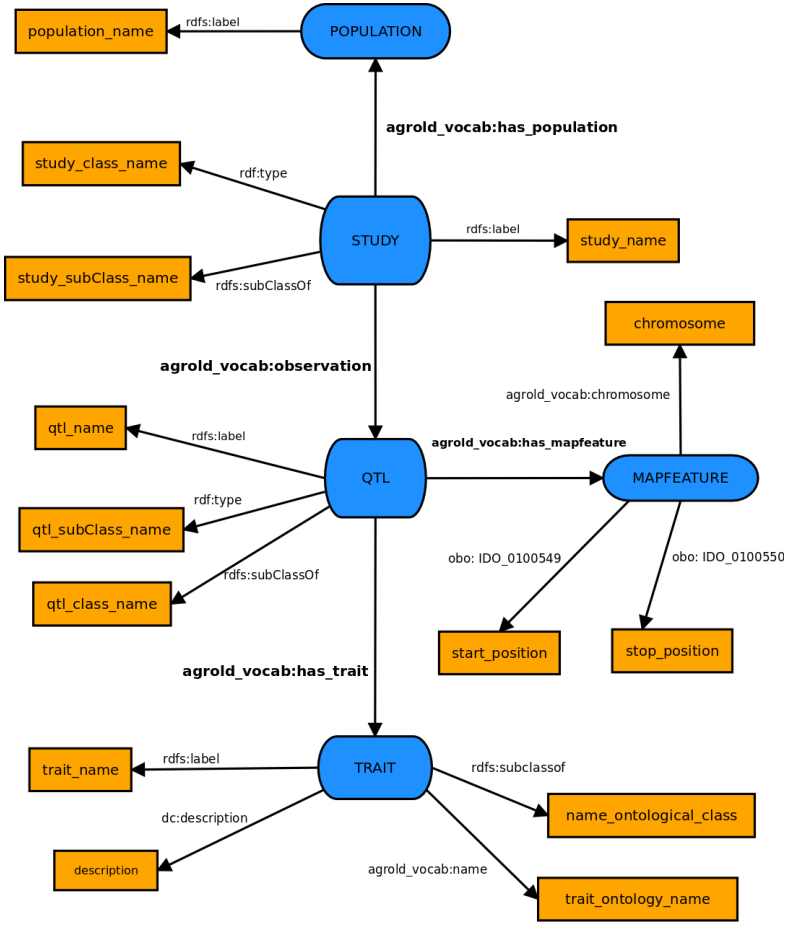
Serve a community

- Plant scientists
- Wheat research
- ...

Large scope

- Several institutes
- Country
- Continent
- Worldwide
- ...

Plant scientific challenge



Scientific challenge

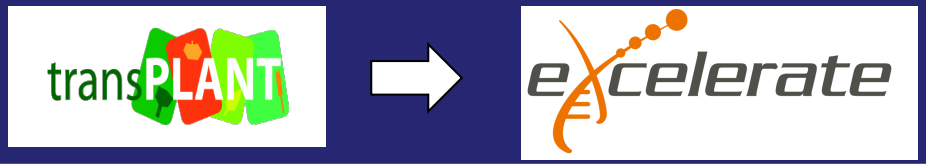
Study of Complex interactions

Phenotype = Genome
x Environment
x Management
practices

Some cyber-infrastructure under construction

WheatIS: the information system of the International Wheat Initiative

Elixir: The European Bioinformatics infrastructure

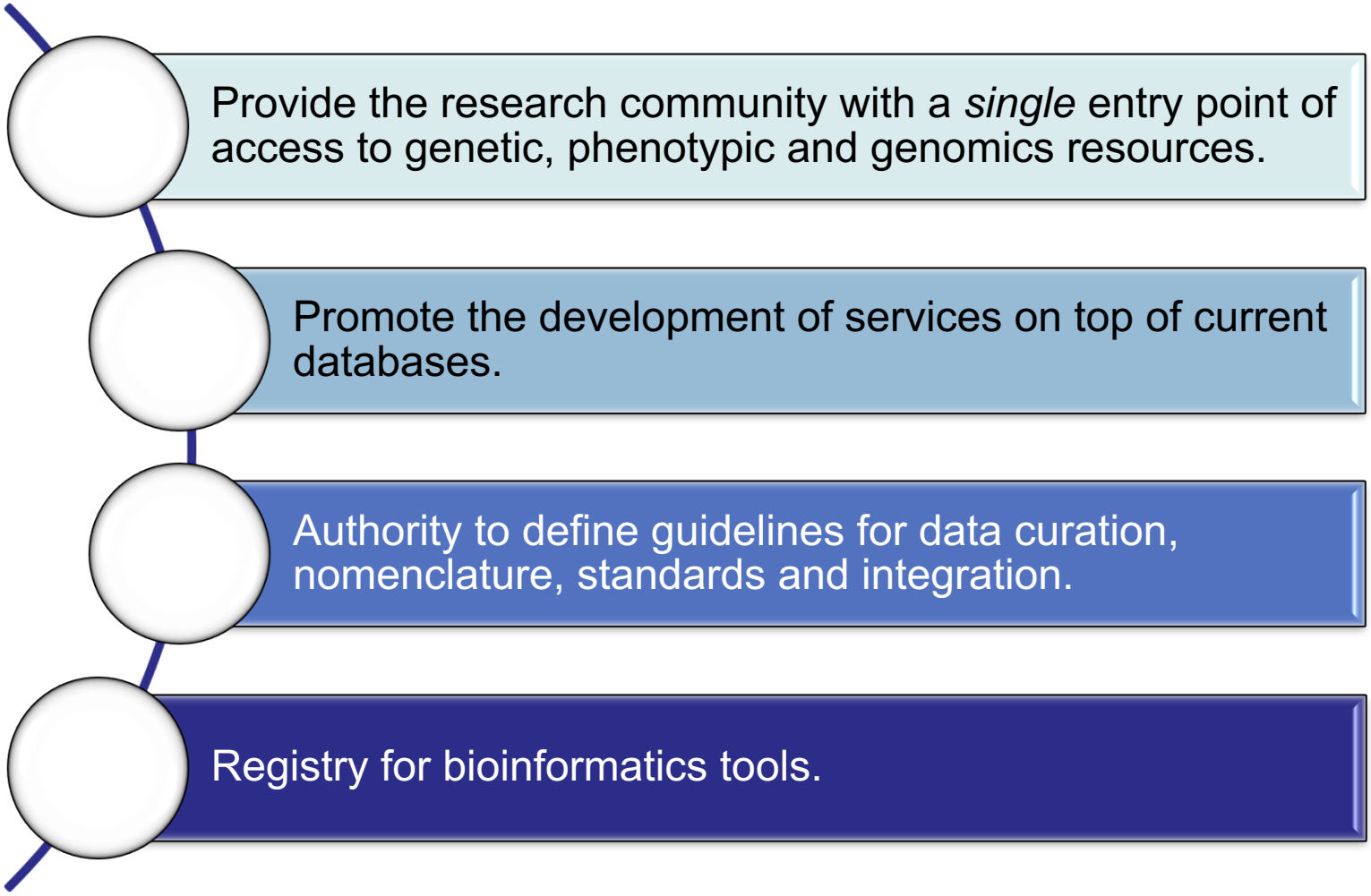


CyVerse: The US cyberinfrastructure (iPlant)

IFB: French Bioinformatics Institute French Elixir node

Phenome/FPPN/EPPN: Information system for French Plant Phenotyping Network

Cyber-infrastructure goals



IFB/Elixir – Plant Bioinformatics

The French Plant bioinformatics node

(33 FTEs)

Genetics and genomics resources for plants and crop parasites

(INRA)

Genomic resource for southern and mediterranean plants.

(CIRAD, INRA, IRD)

Contributors (data, tools, and expertise)

Resources for plants, symbionts and pathogens

(INRA,CNRS)

Marine biology analysis

(CNRS, UPMC)

Arthropods for Agroecosystems

(INRA)

Services

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

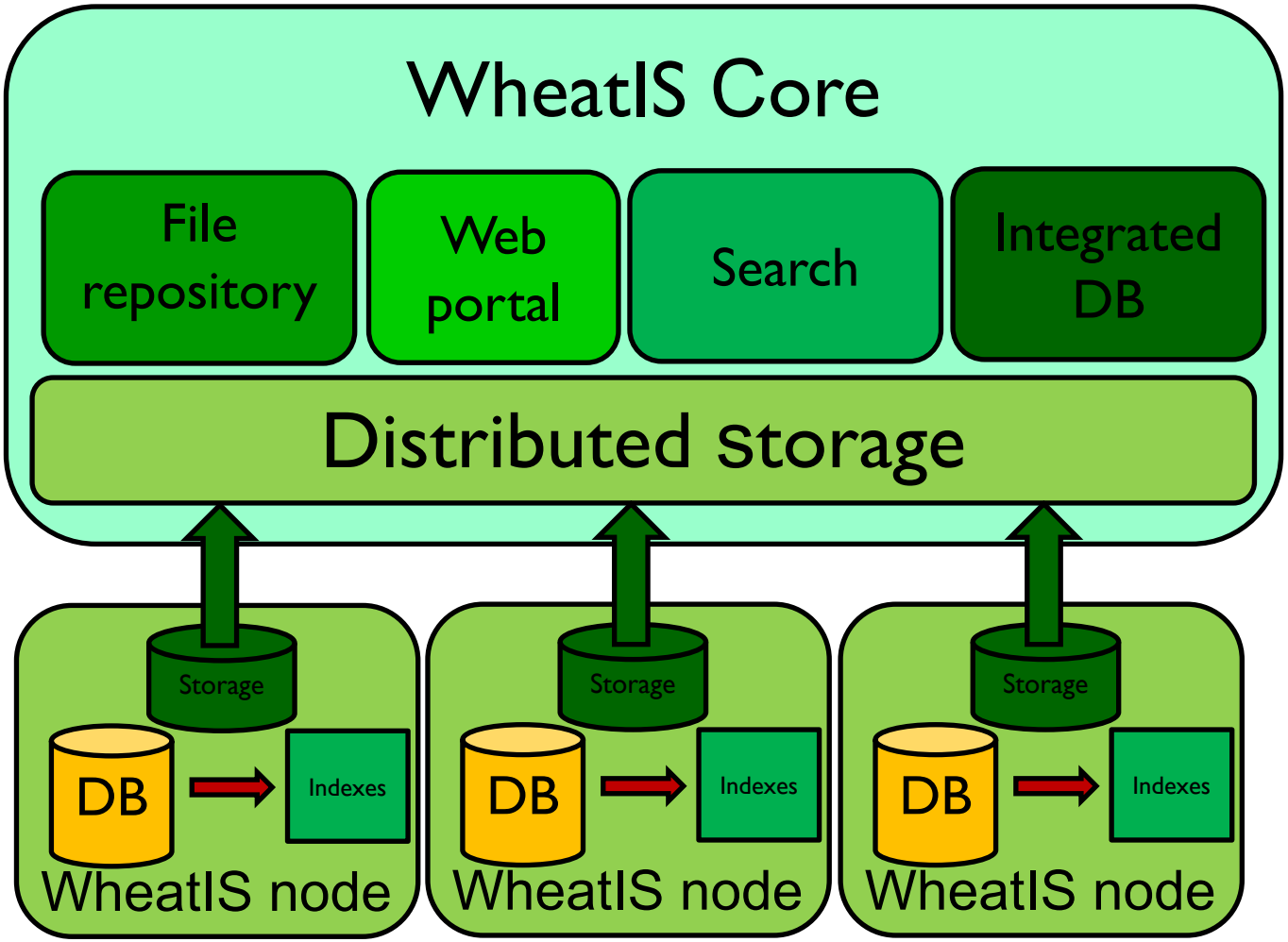
WheatIS – An information system for the wheat international research community

WheatIS Expert Working Group

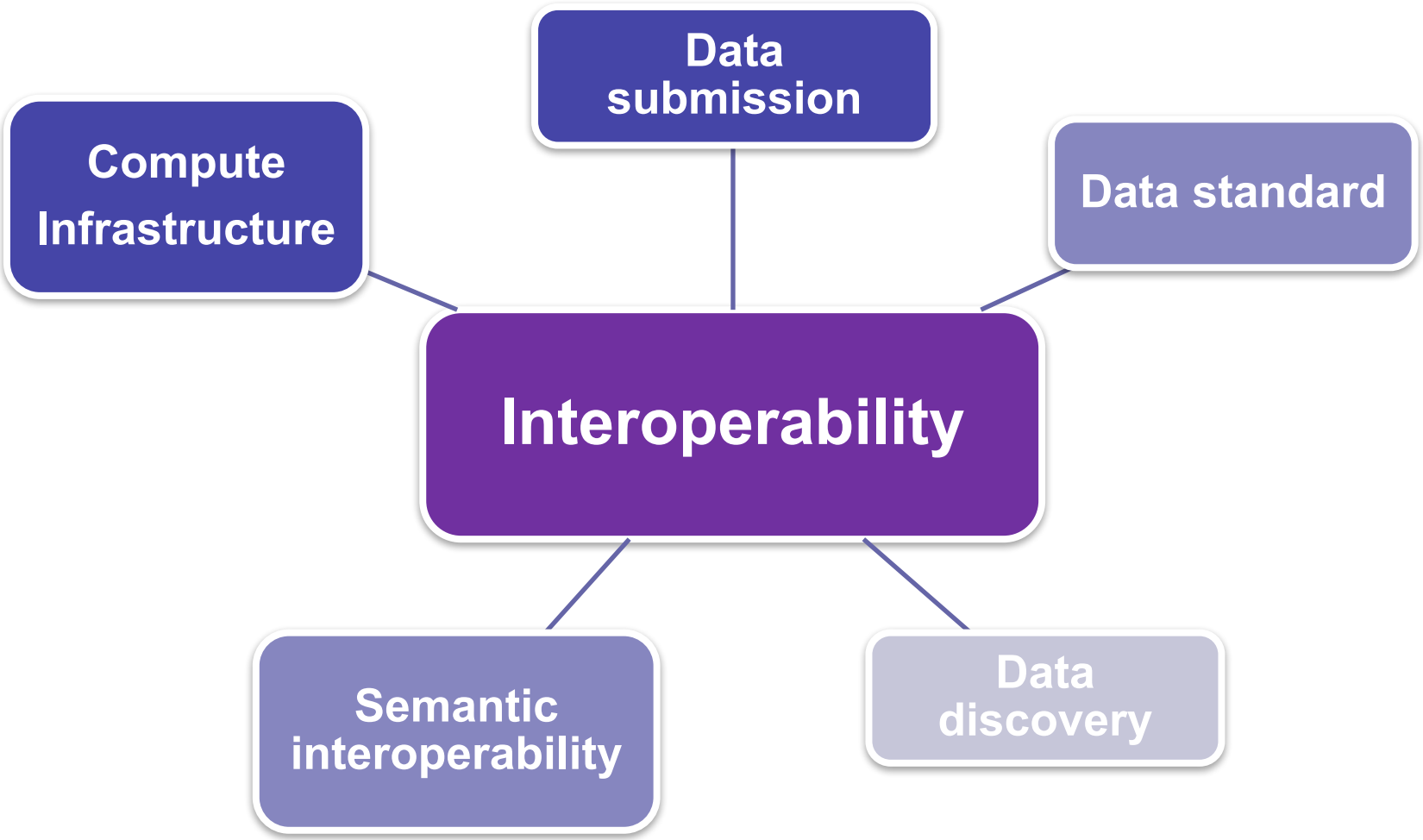
- Build projects
- Build infrastructure
- Report to the Wheat Initiative

WheatIS nodes (#12)

WheatIS Architecture



Cyber-infrastructures key issue



Data must follow FAIR principles

force11.org

To be Findable:

- F1. (meta)data are assigned a globally unique and eternally **persistent identifier**.
- F2. data are described with **rich metadata**.
- F3. (meta)data are registered or indexed in a **searchable** resource.
- F4. metadata specify the data identifier.

To be Accessible:

- A1 (meta)data are **retrievable** by their identifier using a standardized communications protocol.
  - A1.1 the protocol is open, free, and universally implementable.
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- A2 metadata are accessible, even when the data are no longer available.

To be Interoperable:

- I1. (meta)data use a **formal**, accessible, shared, and broadly applicable **language** for knowledge representation.
- I2. (meta)data use **vocabularies** that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

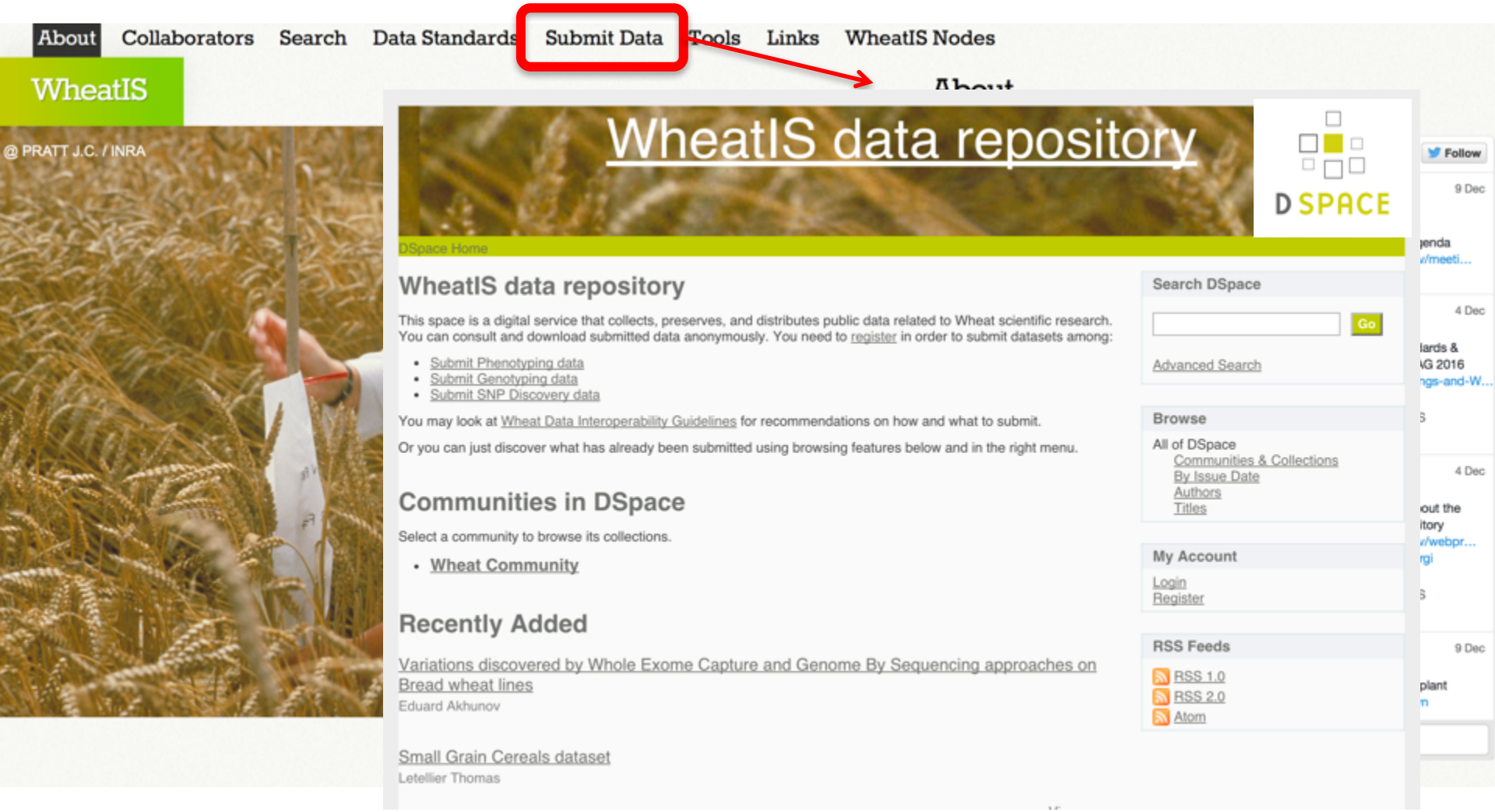
To be Re-usable:

- R1. meta(data) have a plurality of accurate and **relevant attributes**.
  - R1.1. (meta)data are released with a clear and accessible data usage **license**.
  - R1.2. (meta)data are associated with their provenance.
  - R1.3. (meta)data meet domain-relevant community

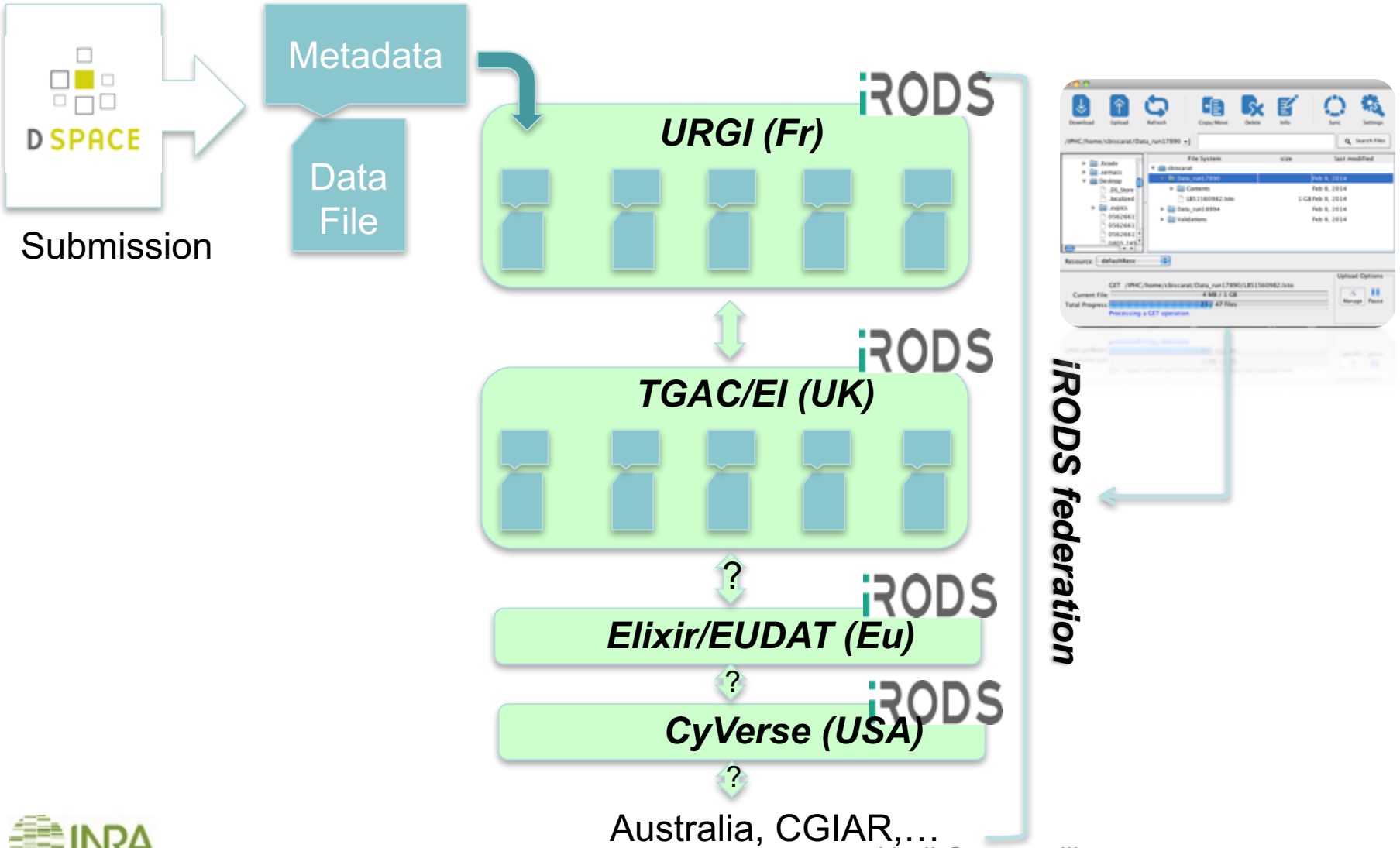
DATA SUBMISSION

Hadi Quesneville

Data file repository

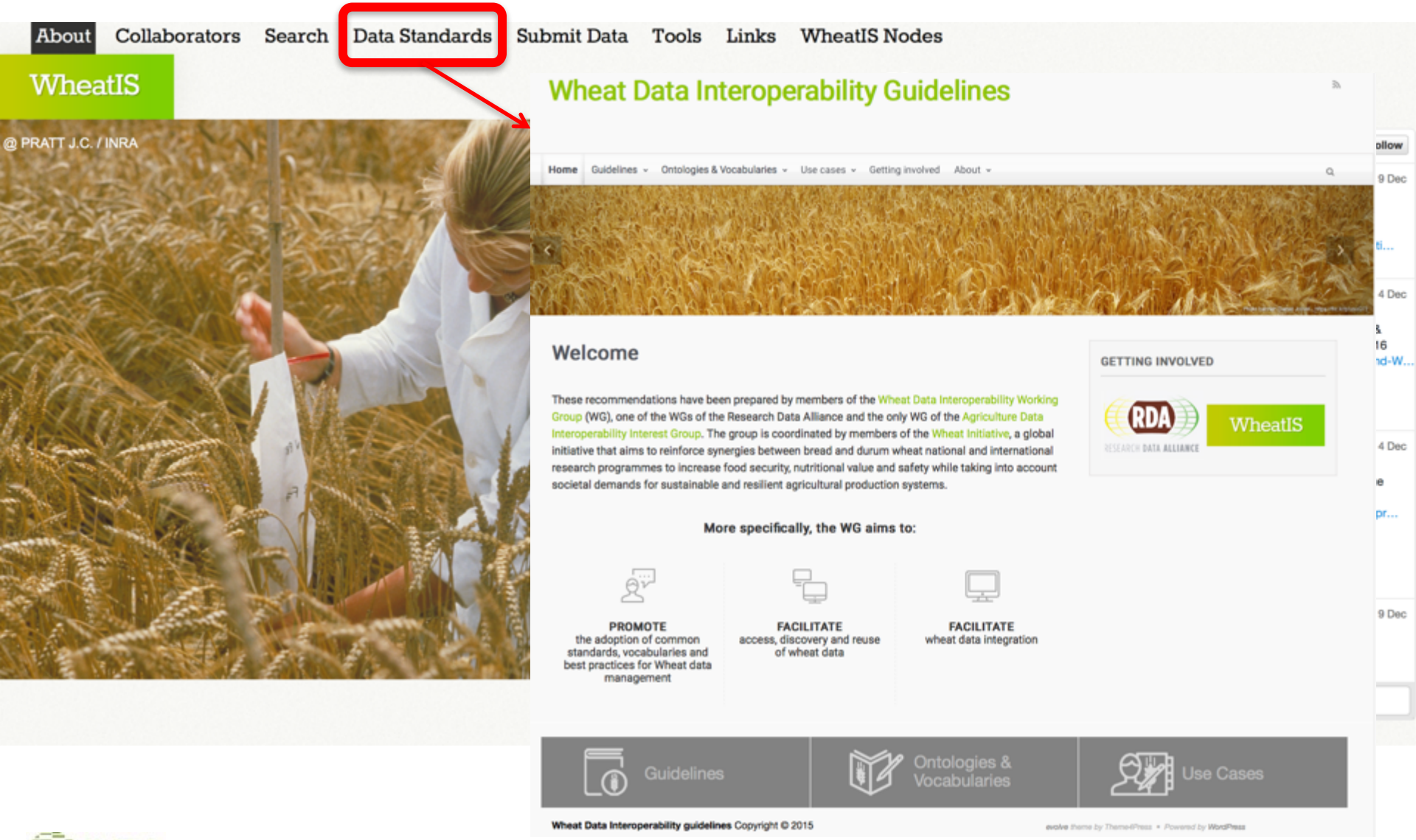


Data workflow



DATA STANDARDS

wheatis.org



INRA SCIENCE & IMPACT

Hadi Quesneville

Wheat Data Interoperability Guidelines

Home | Guidelines | Ontolog

Welcome

These recommendations hav
Group (WG), one of the WGs
Interoperability Interest Grou
initiative that aims to reinfor
research programmes to inc
societal demands for sustai

PROMOTE
the adoption of commo
standards, vocabularies a
best practices for Wheat c
management

Guidel

Wheat Data Interoperability gu

Home / Phenotypes

Phenotypes

Phenotypes are the observable characteristics of an organism
between genes and the environment in which it grows. Plant ph
several fields, such as breeding programs and biological or agr
location environments, germplasm bank characterisation or bio
and associated value) can be used in several fields of expertise
Furthermore, phenotypes are in interaction, like for instance in t
pretreatments have an effect on the glucose yield or in breedin
affected by the number of grains and the grain weight.

This section provides standard formats that are used by the co
and trials in fields and the minimum metadata required for doc
platforms.

Recommendations

Summary

1. Data format: use data matrices in csv, excel
2. Metadata and vocabularies: use complete metadata for at l
   observation variables
3. Keep curated data (checked outliers)

1. Data formats

We recommend following minimum format principles with data
least variables (trait along with method, units and scales or env
germplasms.

ISA-Tab is an implementation of this principle. It consists of or
files and metadata files, the latter being used for data discovery
information can be found on this dedicated page or in this pres
Genome 2015. It is currently well suited for generation by softw
format, phenotype specific configuration and tools are under in

See the germplasm recommendations for data format regardin

2. Metadata and vocabularies

Observation variables

Observation Variables include trait and environment variables.

We recommend using existing variables, listed in the vocabular
To create new observation variables, we recommend using the
Template available at Crop Ontology website. It must include al
description, abbreviation, synonyms, methods, and scales) to d
creation and sharing. The most important field in this template
remain stable and never be modified. Furthermore it must neve
deprecated if needed. This way, it can be used in trials and rem

For Nursery and Trial metadata and description we recommend
Ontology which describes the terms related to nurseries and tr
environments, study design, etc. These metadata are actively a

For biorefinery, we recommend using the Biorefinery ontology v
and terms associated with biomass composition and characte
area, particle size, porosity, etc.), physico-chemical pretreatmer
experimental processes descriptions.

Recommended Variable ontologies and vocabularie

- Wheat crop ontology
- INRA Wheat Ontology (soon publicly available)
- Wheat Phenotype
- Biorefinery ontology
- XEO, XEML Environment Ontology

For the difference between metadata, ontologies and vocabula

3. Raw data

We recommend sharing at least clean documented raw data, like plant

Phenotype data lifecycle begins with acquisition, then cleaning, elabor
elaboration combines several variables, like phenological stages and t
elaborated/computed variables used as input for analyses softwares.
and phenology can be combined to get height at flowering. Different e
produced for different purposes, it is therefore important to be able to
ones from raw data.

Some popular Tools

1. Repositories, information systems and dat tools

The Breeding Managment System, BMS generates standard format fo
trial data in fields and uses for variables the Crop Research Ontology f
experiment related metadata and trait related ontologies of the Crop O
makes it possible to analyze data directly using statistical tools such a
R.

GnpIS is an INRA information system designed for plant and pest gen
scientists to mine genomic, phenomic and genetic data. For phenomic
allows data discovery through a keyword based, google like, search en
The latter allows dataset building for genetic or phenomic analysis. Du
is based on a strict identification of germplasms on variables through
the Crop Ontology.

The Breeding API specifies a standard interface for plant phenotype/g
serve their data to crop breeding applications. It is a *shared, open* API,
providers and data consumers who wish to participate.

For biorefinery applications, the best match pretreatment-biomass acti
yields can be found through the @Web platform. The *Documents* tab c
information by a kind of pretreatment (topics Bioref-XX). Data availabl
pretreatments used, biomass types and characterization, etc. In the fu
possible to find the best match pretreatment-phenotype.

iPlant collaborative offers many services that allow the analysis of ger
phenotypic data.

2. Data acquisition

Field Book is a simple app for taking phenotypic notes on field researc
in the field has traditionally been a laborious process requiring writing
by transcription. We have created Field Book to replace paper field boo
collection speed with greater data integrity.

Things to follow in the future

Candidate formats

- ped
- Bagit
- Hadoop File System : HDF5

Written by: WDI working group
Published on: 02 October 2014
Updated on: 27 April 2015

2 Comments

Bettina Berger 27 November 2015

Sincere apologies for not taking sufficient time for a thorough assessment of this site. Just some minor points I noticed while browsing through.

I could not find a link to PATO, which I would assume is a useful tool for annotating phenotypes
Also, there's no reference to iPlant in the US. This may be intentional, but the first phenotype dataset we were requested to make available for a publication was deposited on iPlant and more may follow in future.

Reply

Cyril Pommier 22 December 2015

Thank you for your feedback.
We have added the link to PATO on the ontology page. There is now also a quick reference to iPlant, but not very detailled since we are not direct users of this system. Feel free to send us a more detailled description, we will work on its integration.

Reply

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment

Name *

Email *

Website

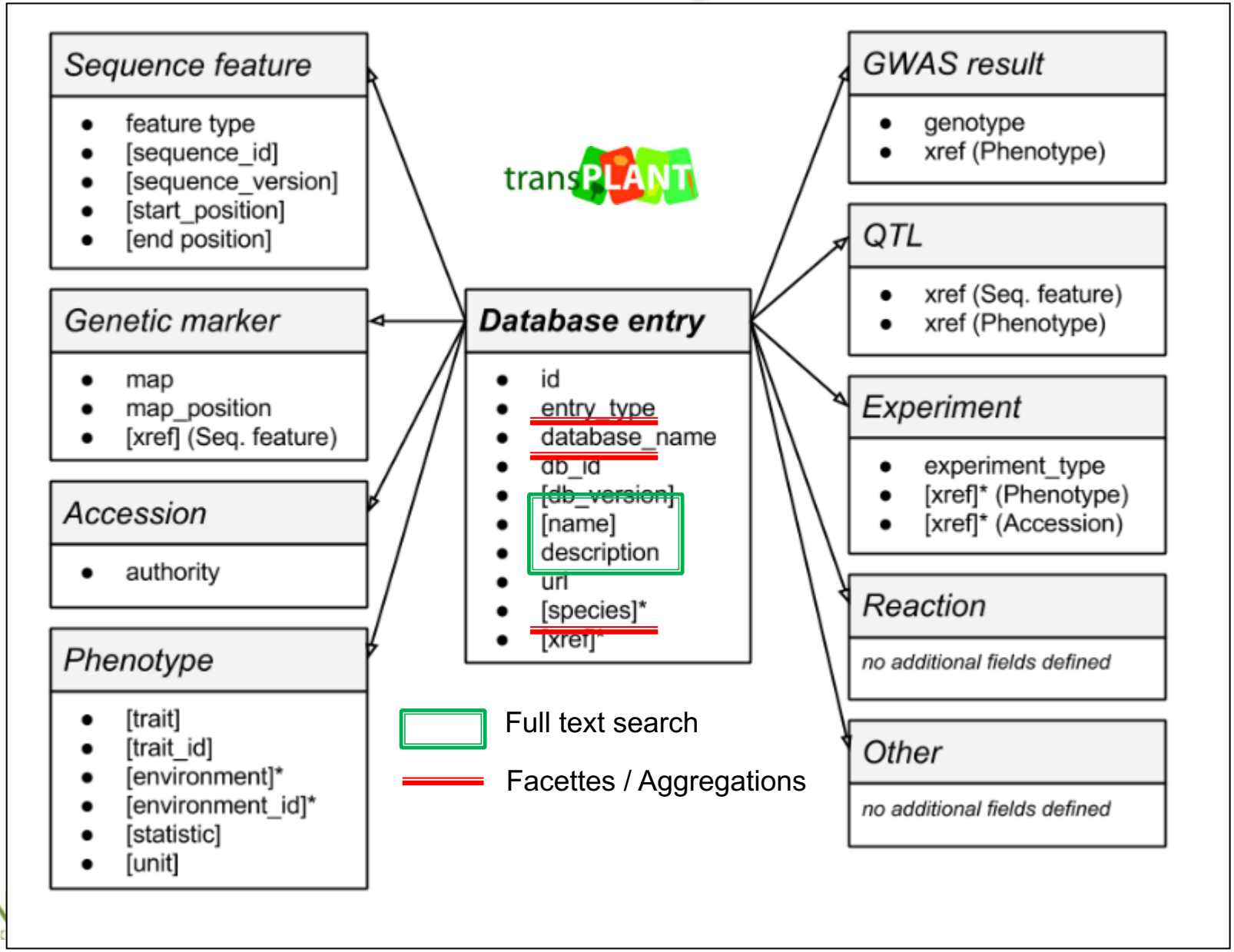
Post Comment

DATA DISCOVERY

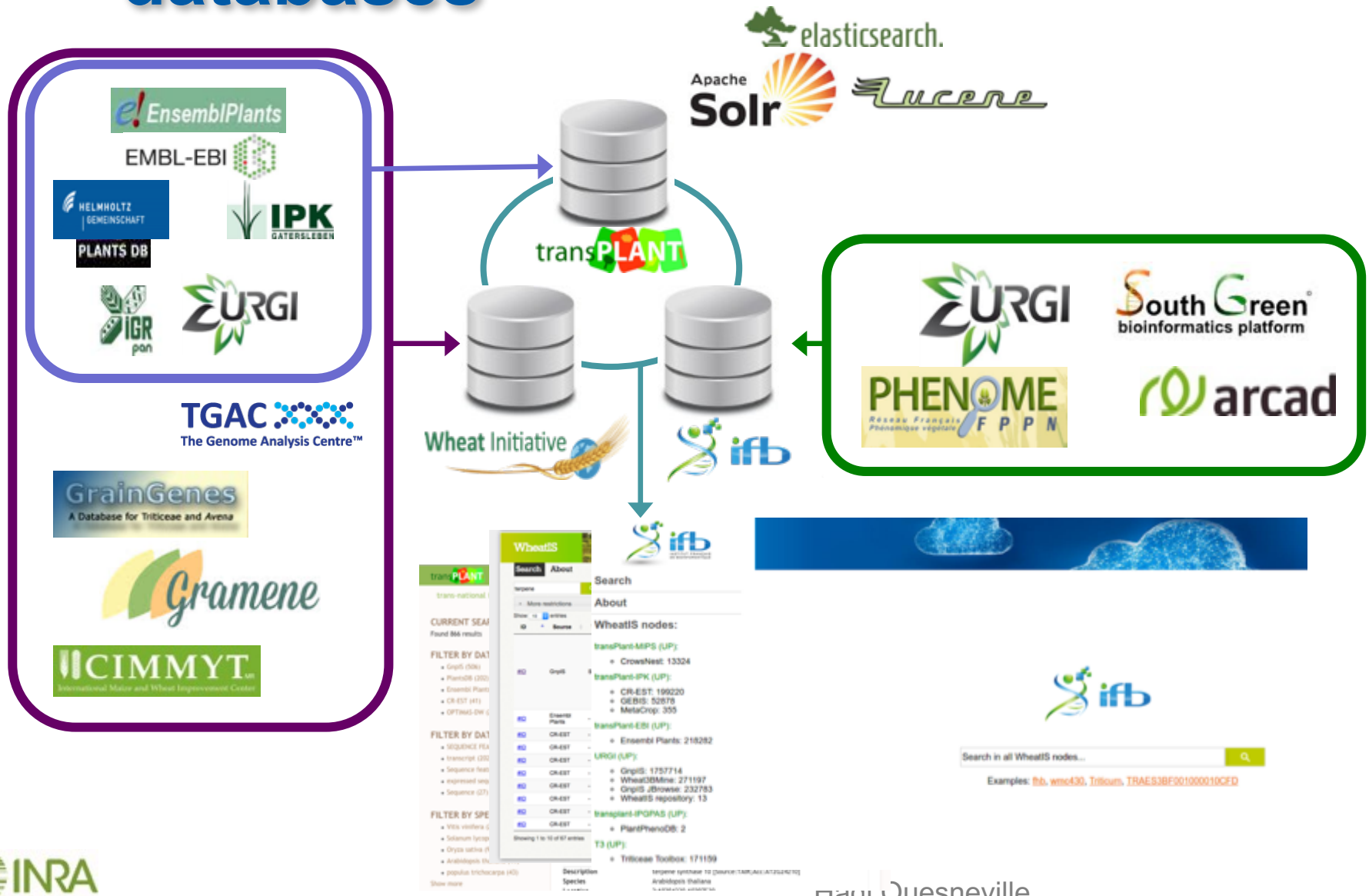
INRA
SCIENCE & IMPACT

Hadi Quesneville

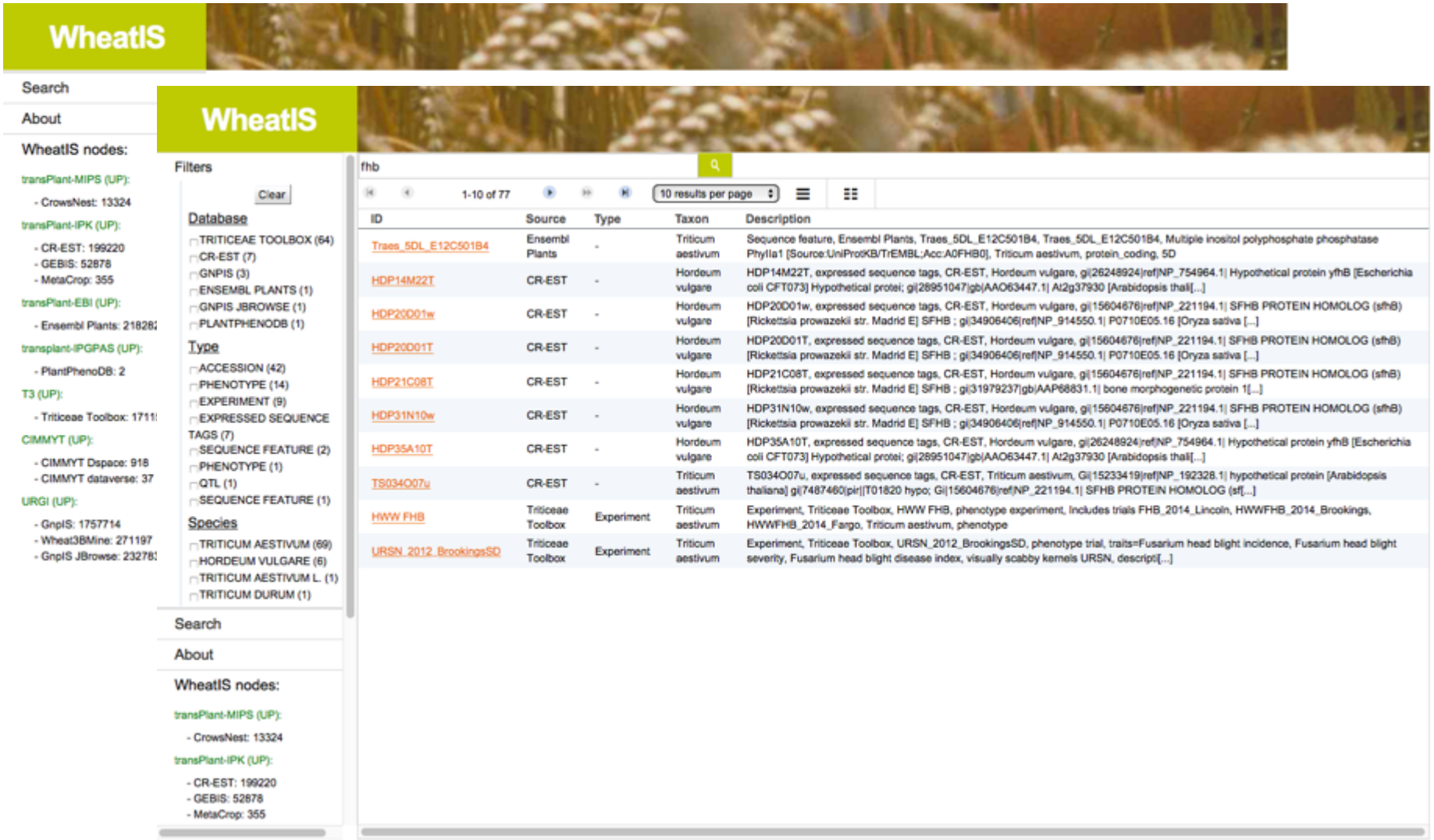
Data Discovery data model



Full text search of distributed databases

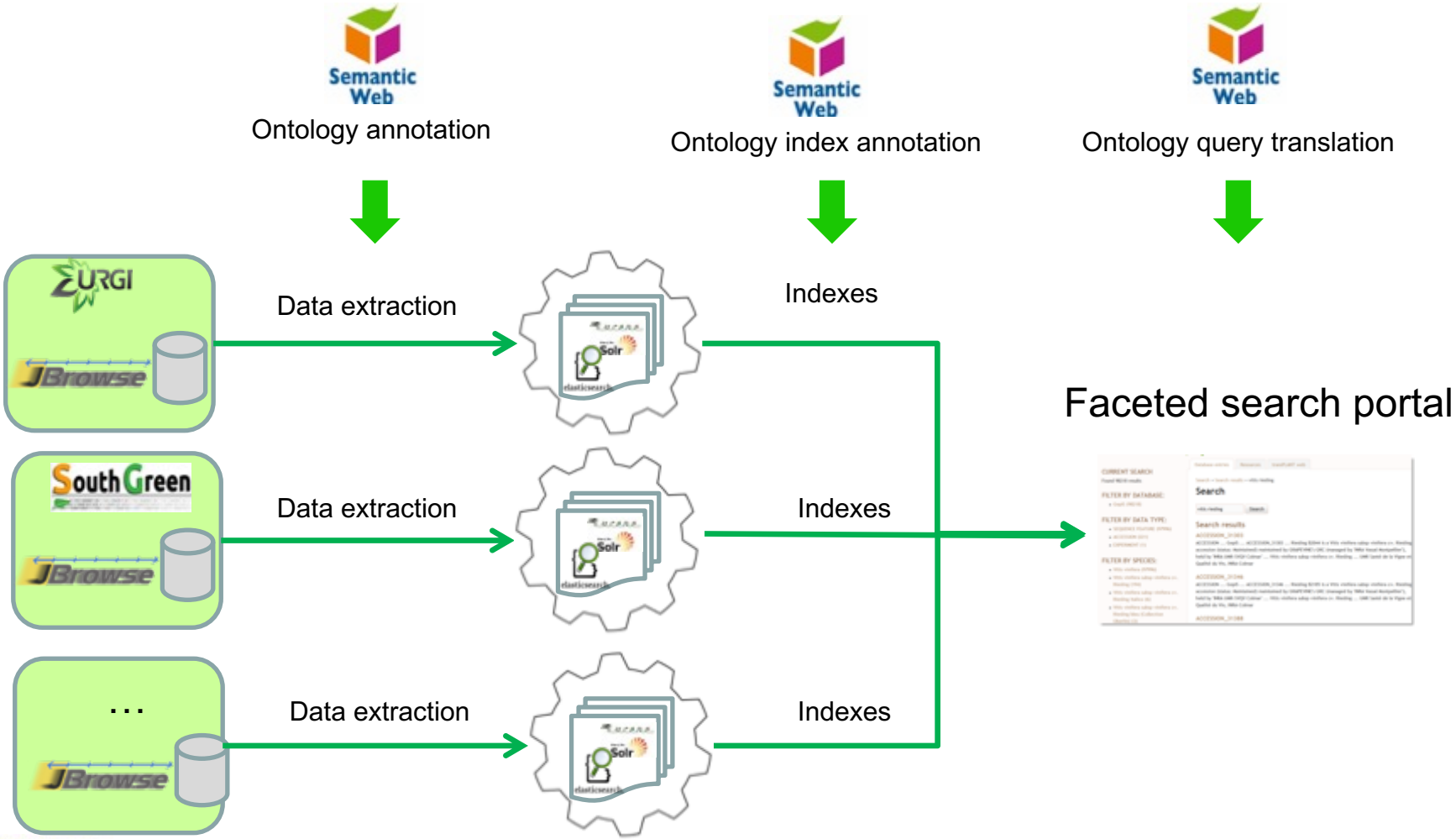


wheatis.org

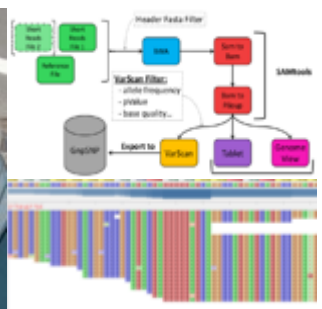
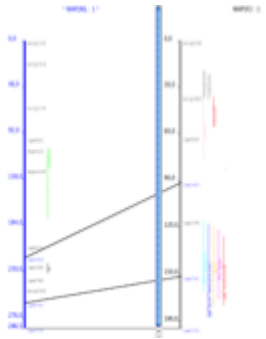
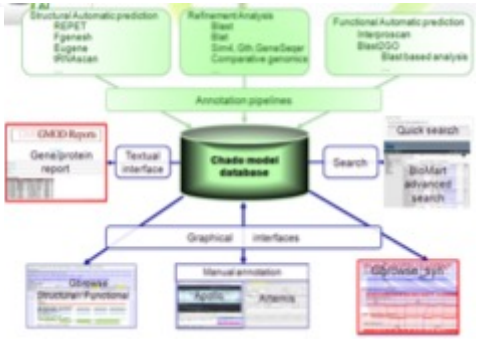


Hadi Quesneville

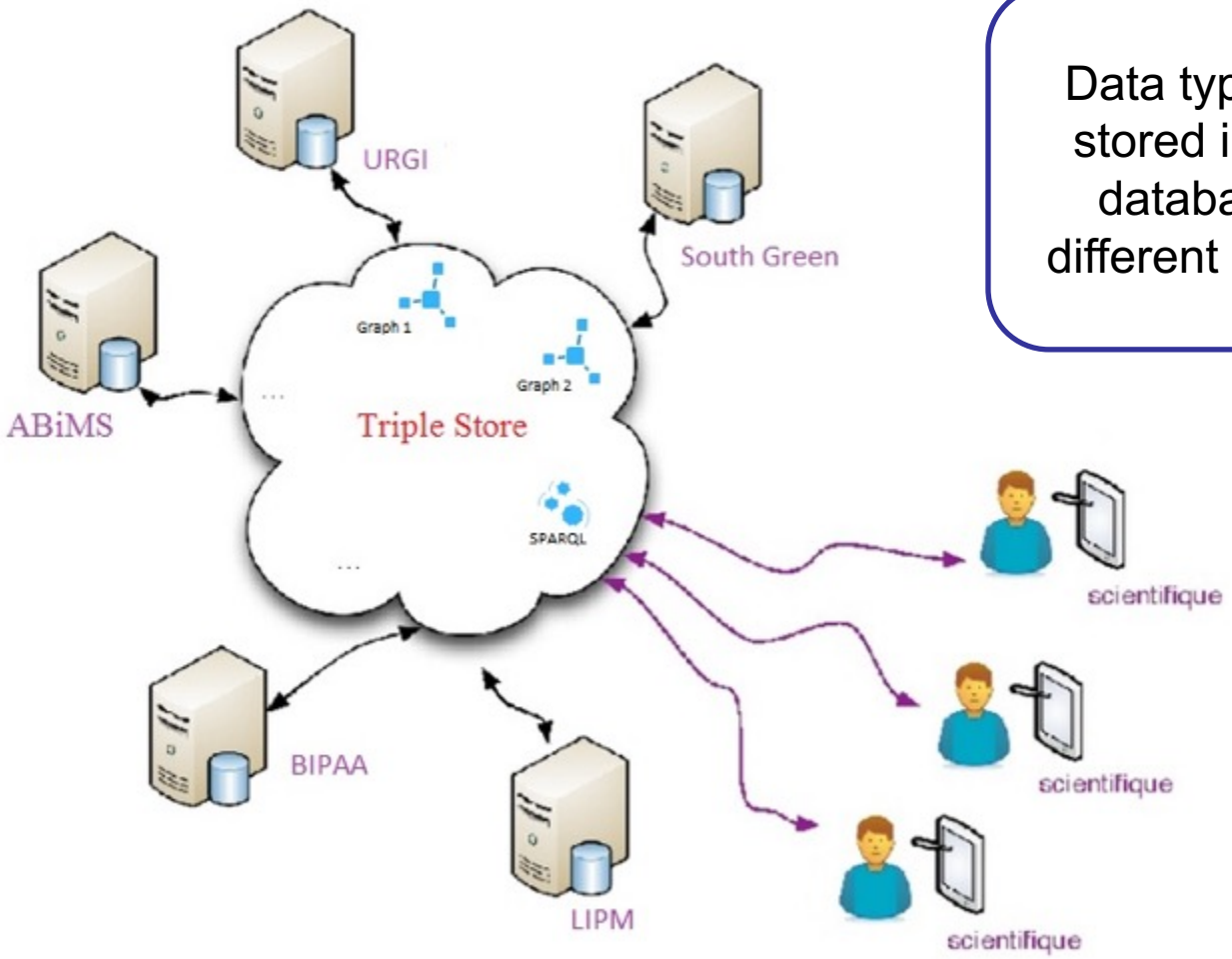
Next for data discovery



Semantic Interoperability

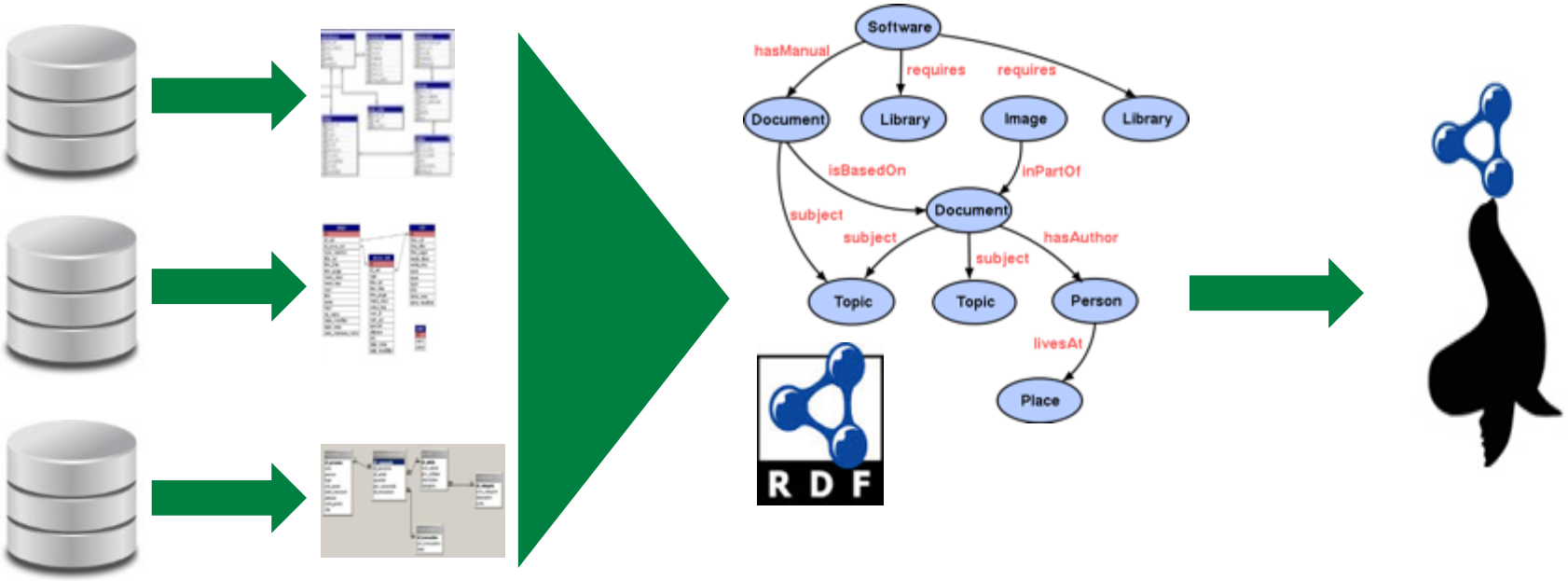




Why a semantic interoperability



Data types can be stored in different databases with different data model

Develop a web semantic interoperability



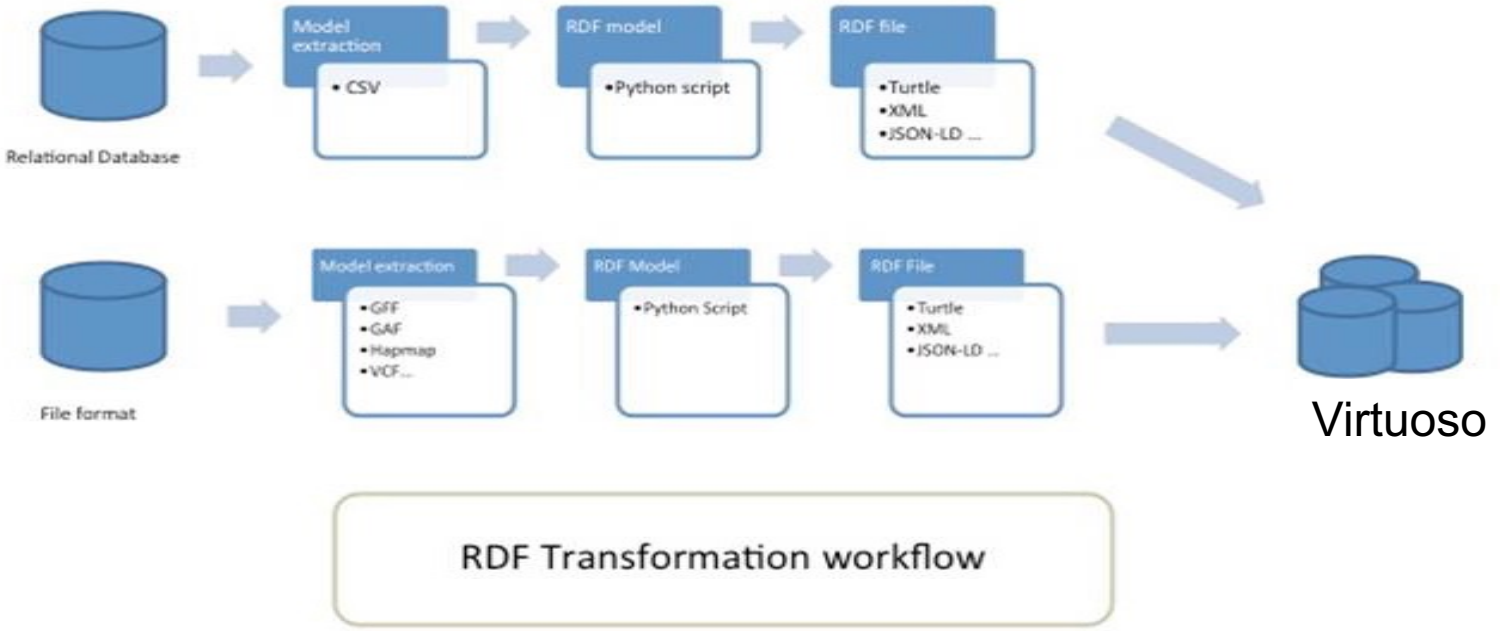
Ontology based annotation of database schemas (GO, PO, TO, CO, …)

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)

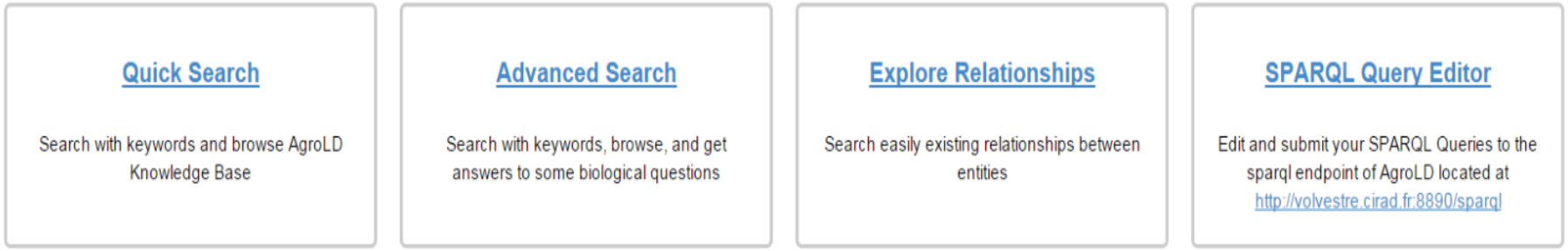
Workflow ETL



- **Some workflows are generic enough to be reused (i.e. GFF3 or HapMap)**
- **Codes available on GitHub, https://github.com/SouthGreenPlatform/AgroLD.**

AgroLD portal
(www.agrold.org)

- **Different user profiles**
- **Provides different entry points**



Integrates information from:

- **Ontologies:** Gene Ontology (GO), Plant Ontology (PO), Plant Trait Ontology (TO), Plant Environment Ontology (EO)…

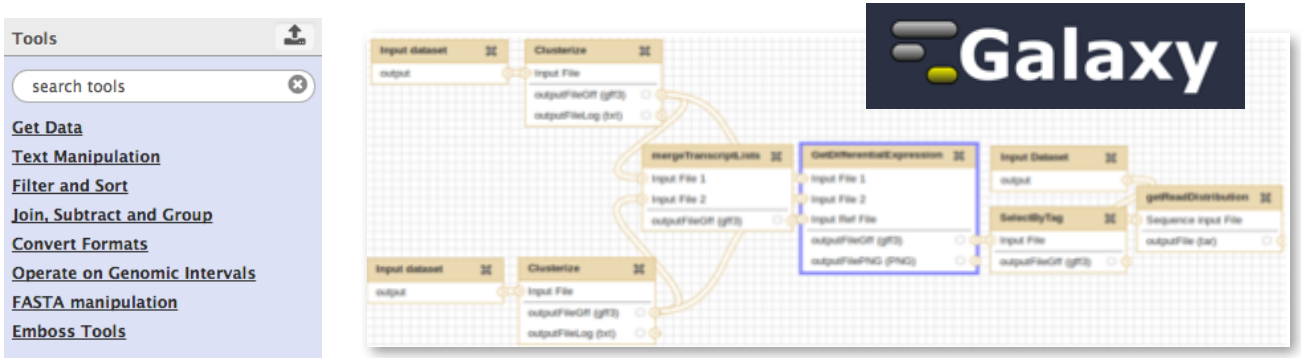
- **Information sources:**
  - Ontology association: GOA, Gramene (TO, PO and EO)
  - Gene/Protein information: OryGenesDB, Gramene, UniProt
  - QTL information: TropGeneDB, Gramene
  - Polymorphisms: SniPlay
  - Pathway information: Gramene - Cyc
  - Phenotype information: Oryza Tag Line
  - Phylogenomic information: GreenPhylDB

Phenotypic data integration

- **RDF conversion from the Breeding API standard**
- **Benefit from the work of the BRAPI international consortium**
  - REST API and JSON format
  - Web Service Swagger

https://urgi.versailles.inra.fr/ws/swagger-ui/index.html#/Phenotyping

Compute infrastructure



Annotation analysis

- Gene structural and functional annotation
- Repeat detection and annotation (TE, SSR, SD)

Structural analysis

- SNP calling, InDel detection
- Transcript assembly

Functional analysis

- Differential expression
- ChIP-seq, BS-seq

Genetic diversity analysis

- Population genetic statistics
- genome wide association mapping

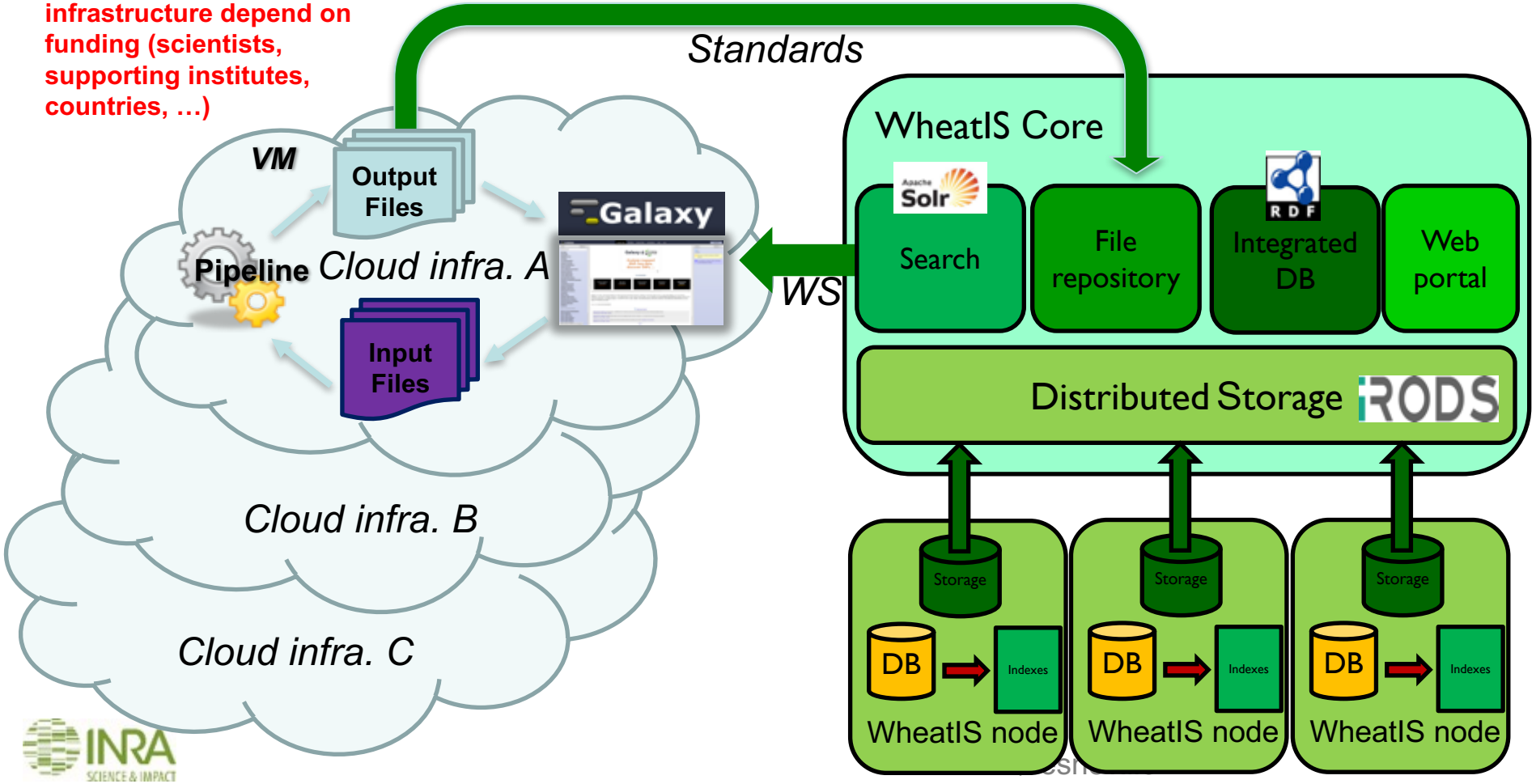
Compute infrastructures

Several Compute infrastructure

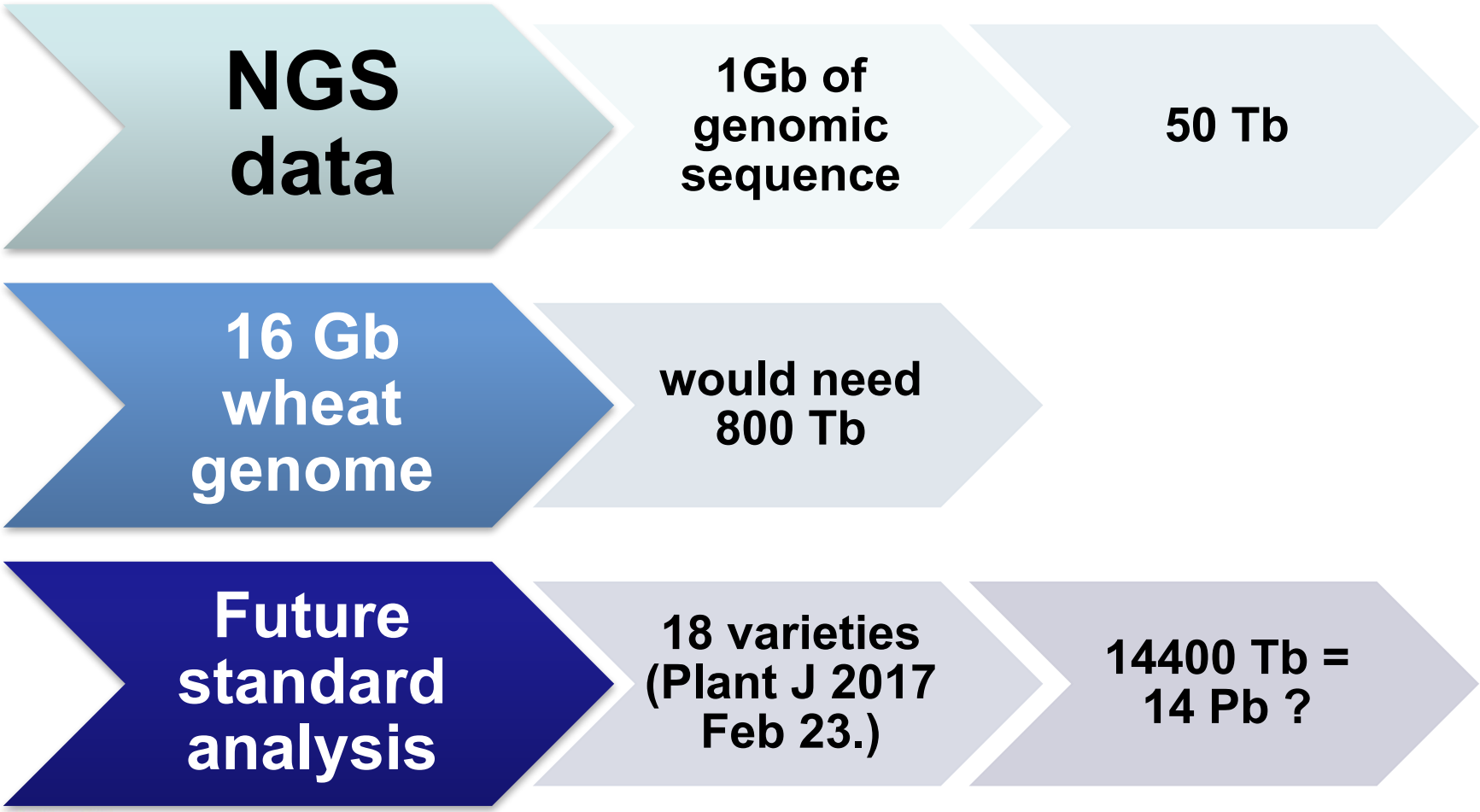
iRODS

One data infrastructure

Availability of infrastructure depend on funding (scientists, supporting institutes, countries, …)



A wheat use case



Big data issues

Difficult to imagine having such storage capacity

Improve current approach of data analysis

Compute infrastructure availability

Metadata

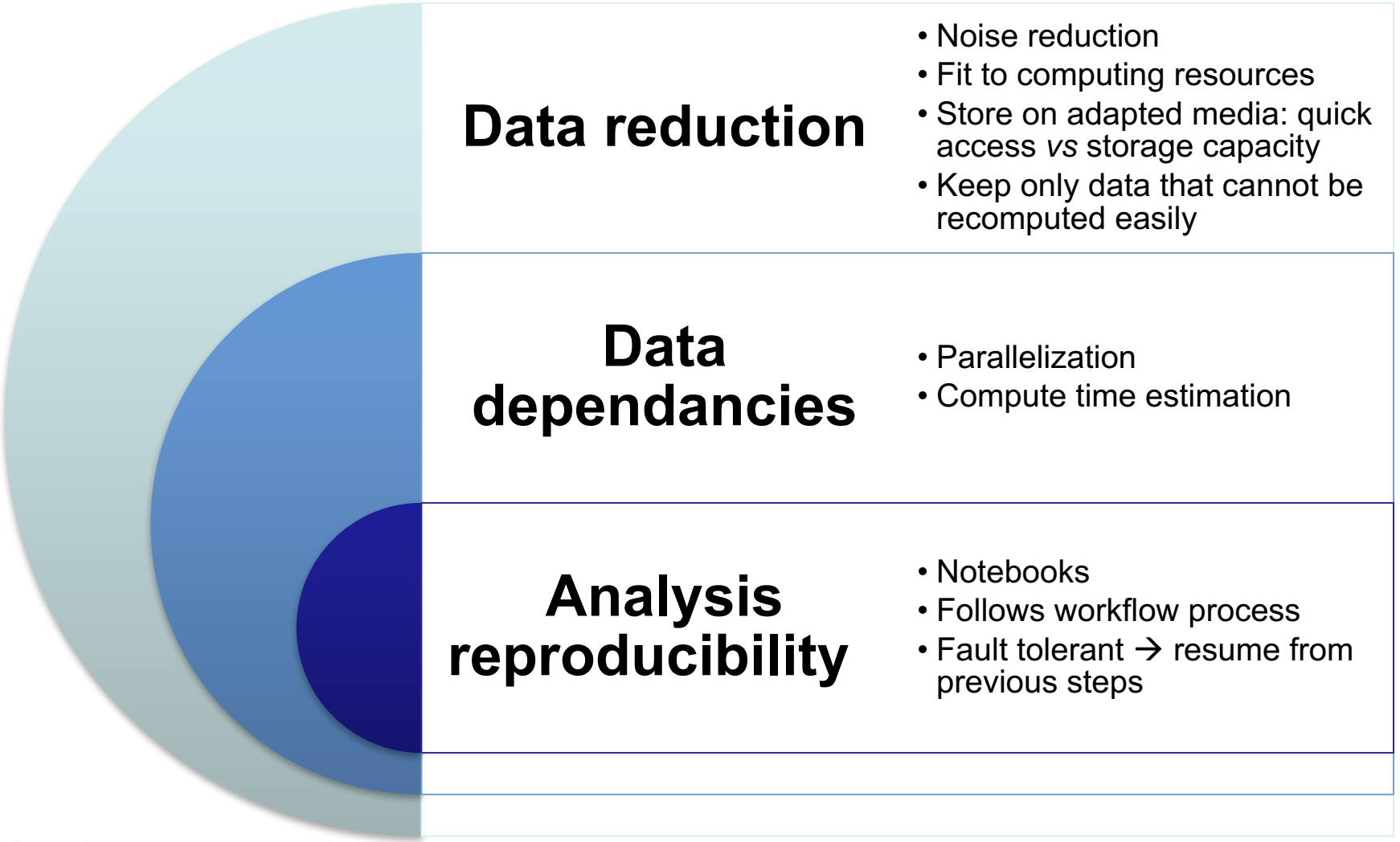
« FAIR » repositories

Efficient data storage

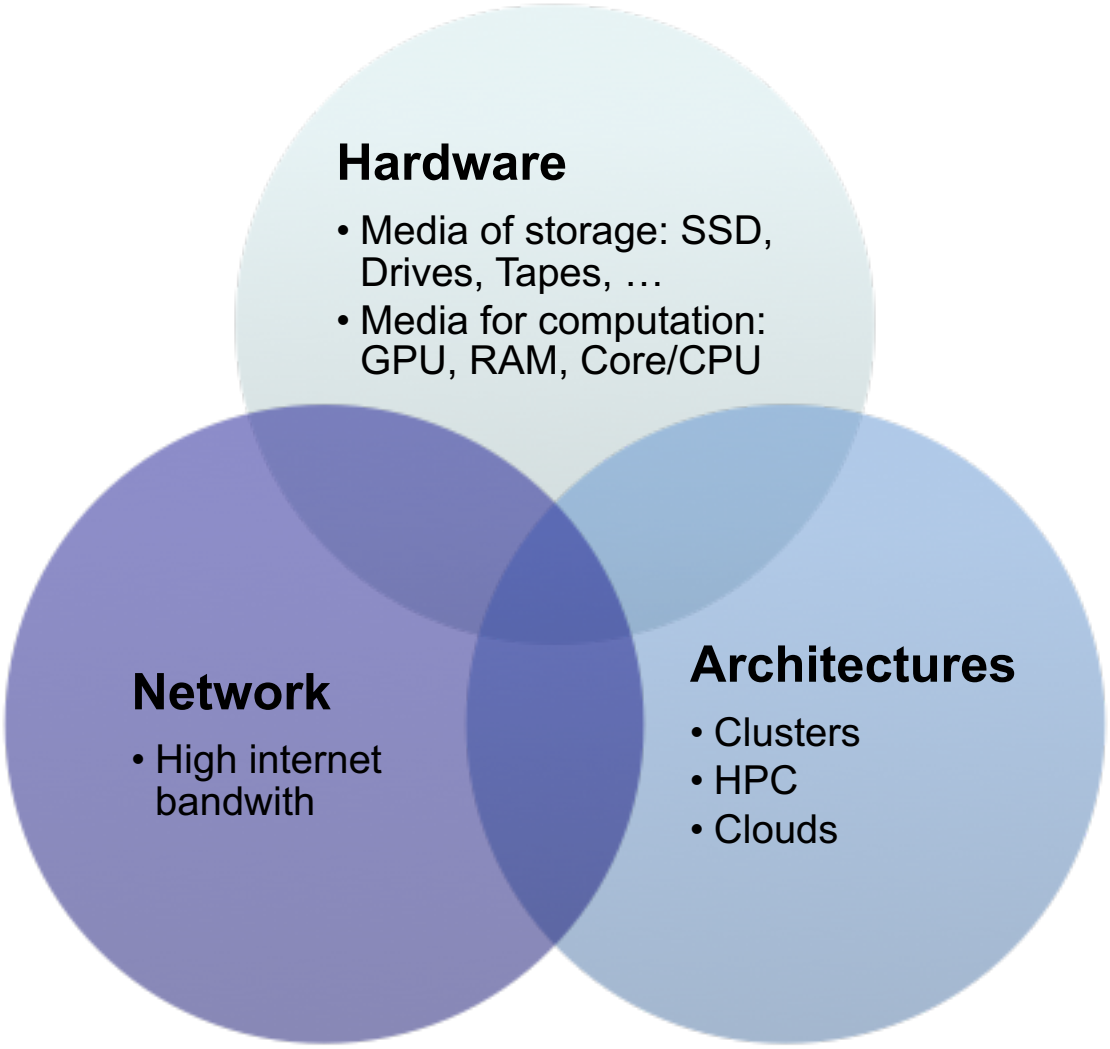
Safe data transfert protocoles

File completion validation : *MD5 checksum*, …
Fault tolerance: *Rsync*, …

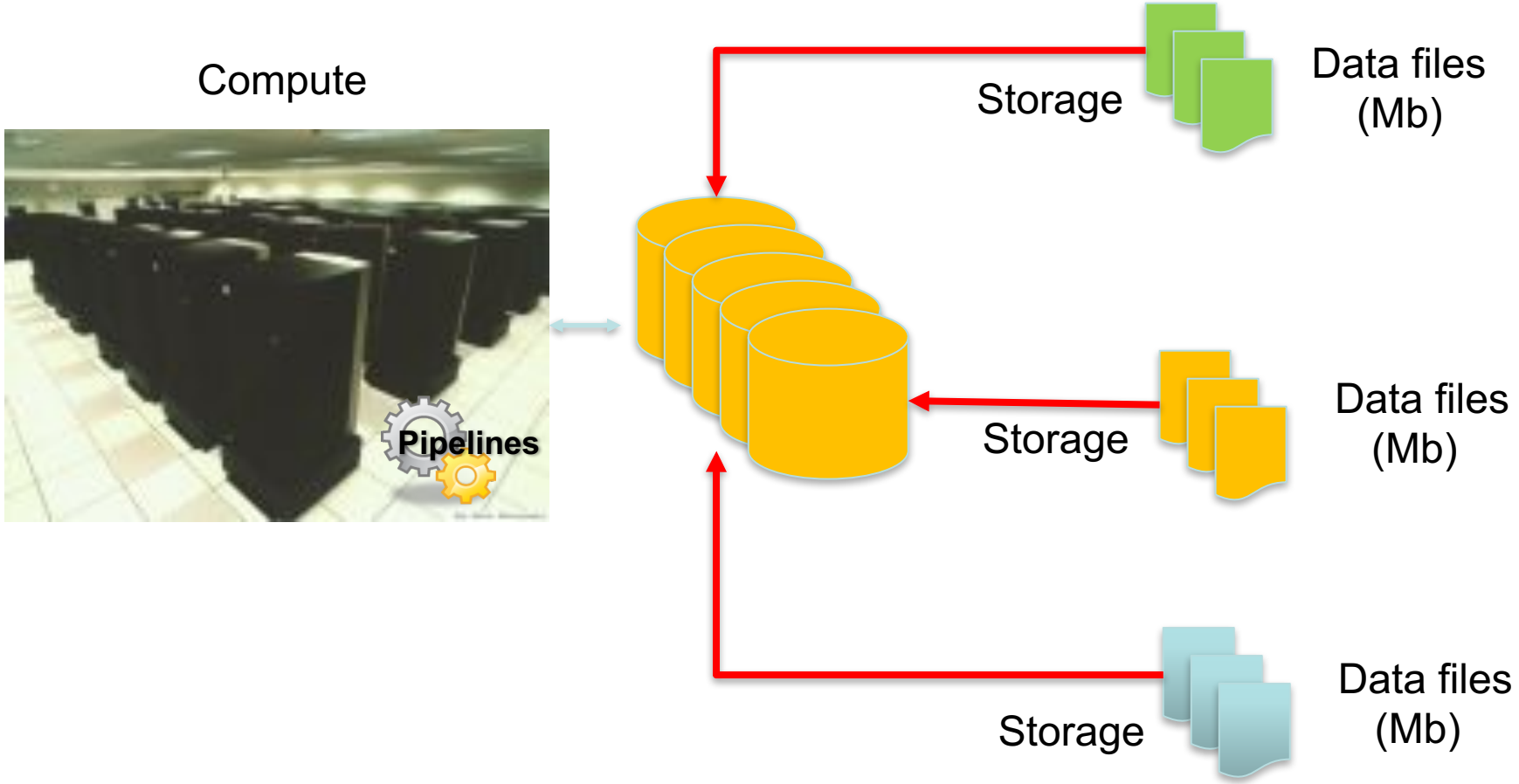
Efficient workflows



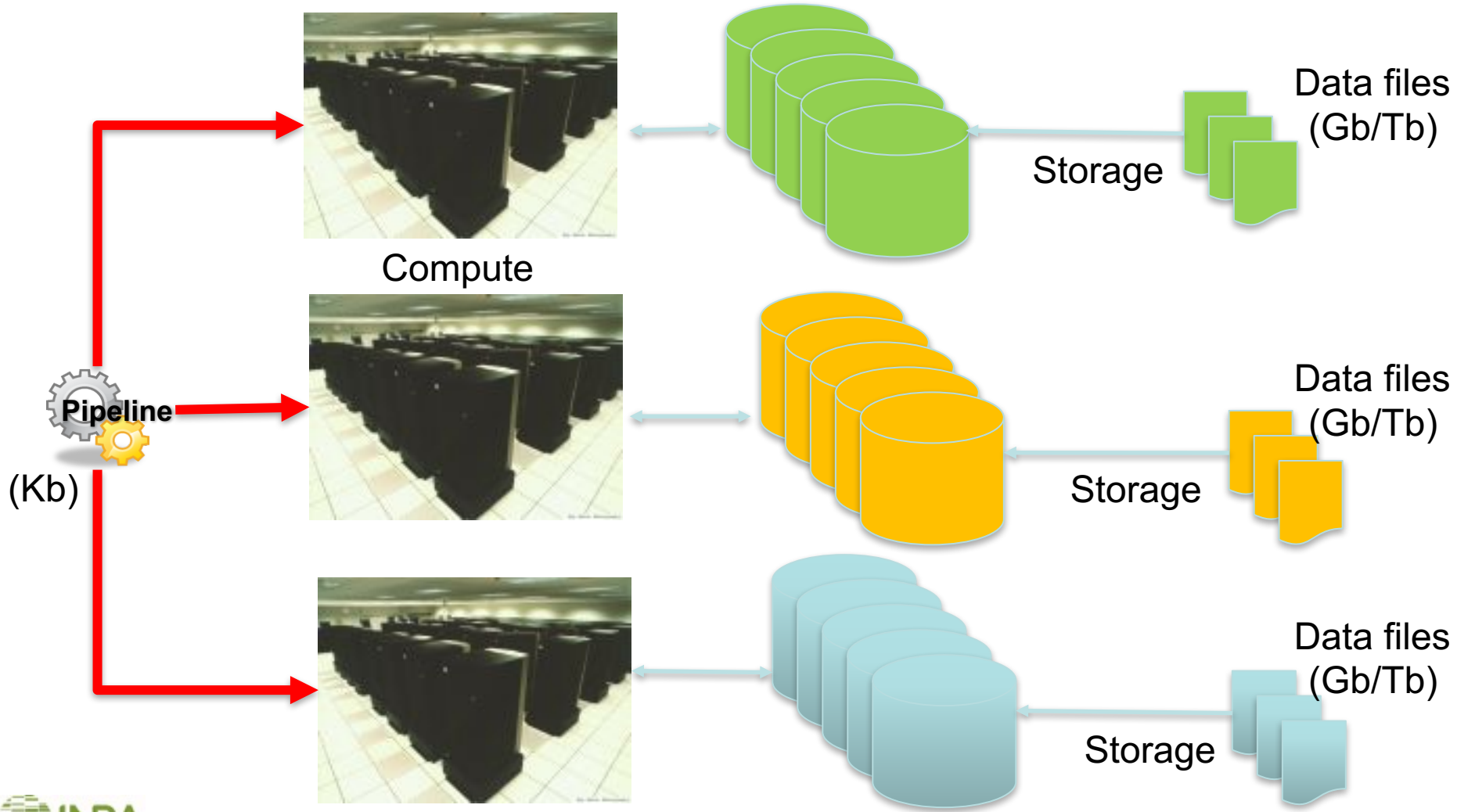
Efficient computer resources



Today's paradigm: moving data to tools



Paradigm shift: moving tools to data



Moving workflows

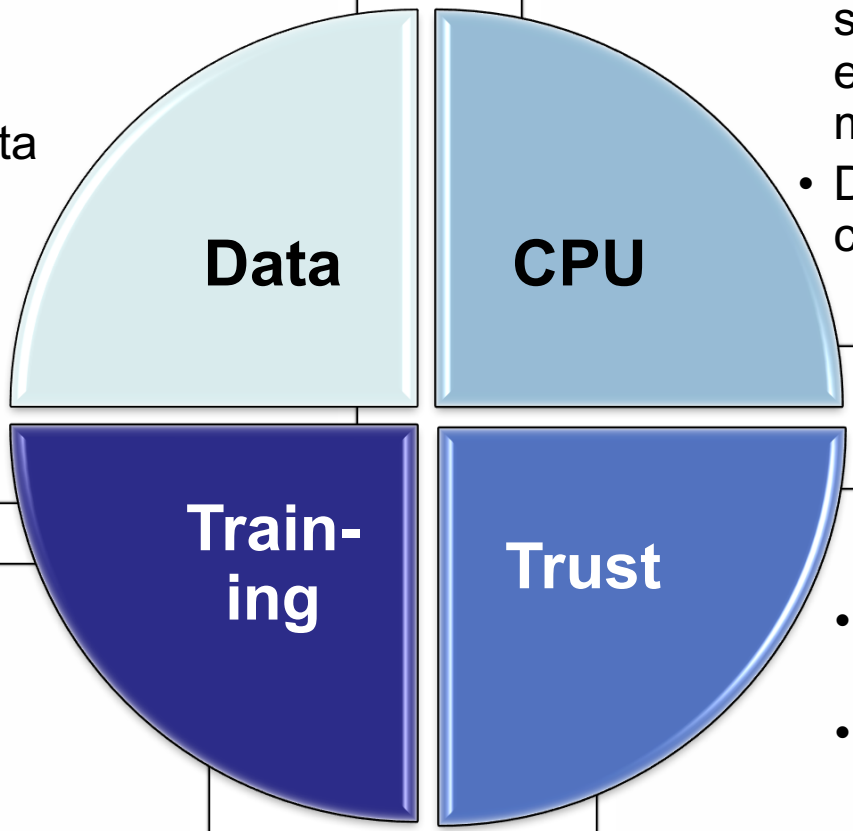
- **Moving complex workflows is challenging**
  - Hardware dependancies (Cluster)
  - OS dependancies (Unix/Windows/…, Scheduler)
  - Software dependancies

- **Technical solutions exist**
  - Virtualisation: moving an image as a file
  - Installation package: rpm, conda, ansible, puppet,…

➔ **Cloud infrastructure**

Conclusions: Cyber-infrastructure challenges

- Dispersed data.
- Curation and annotation
- Heretogeneous data type
- Establishment and adoption of standards
- Easy submission



- Big data issues (DNA sequences, images, environment measures, …)
- Data access from compute resources

- Bioinformaticians
- Researchers

- Accessibility, easy exploitation, etc.
- Ethics (credit for data)
- Institutional support

Acknowledgments

R. Flores,
O. Inizan,
M. Loaec,
C. Pommier,

P. Larmande,
M. Ruiz,
et al.

Christophe Caron

Esther Dzalé Yeumo Kadoré,
Richard Fulss
et al.

Dave Edwards,
Gerard Lazo,
Mario Caccamo,
et al.

P. Kersey,
P. Krajewski,
H. Cwiek,
et al.

Collaborators
Fabrice Legeai (BIPAA)
Jérome Gouzy (LIPM)