



**HAL**  
open science

## 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

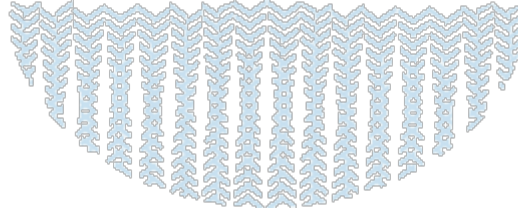
**HAL Id: hal-02786556**

**<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.

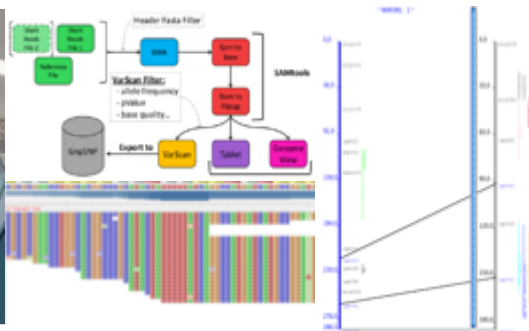
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# Plant Cyber-infrastructures: WheatIS & French Elixir node

Hadi Quesneville

Biotalent, May 29, Poznan



# 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
- ...

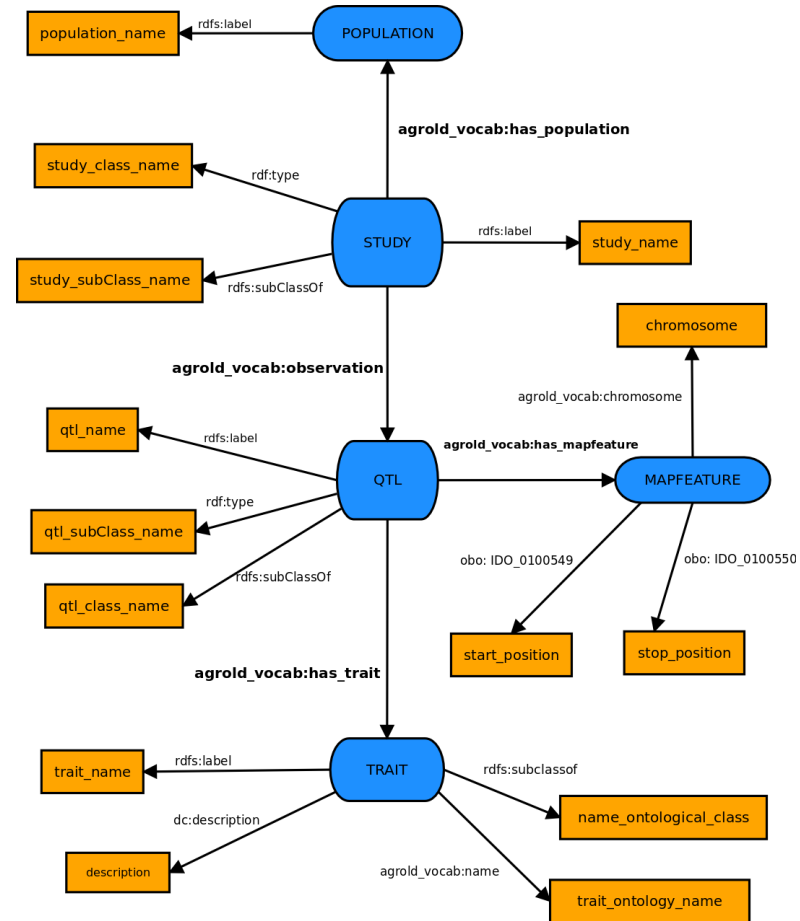
# Plant scientific challenge



## Scientific challenge

Study of Complex interactions

*Phenotype = Genome  
x Environment  
x Management  
practices*







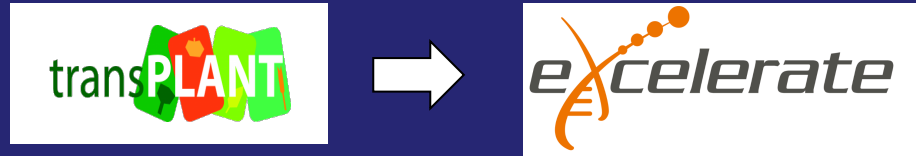
# Some cyber-infrastructure under construction



**WheatIS:** the information system of the International Wheat Initiative



**Elixir:** The European Bioinformatics infrastructure



**CyVerse:** The US cyberinfrastructure (iPlant)



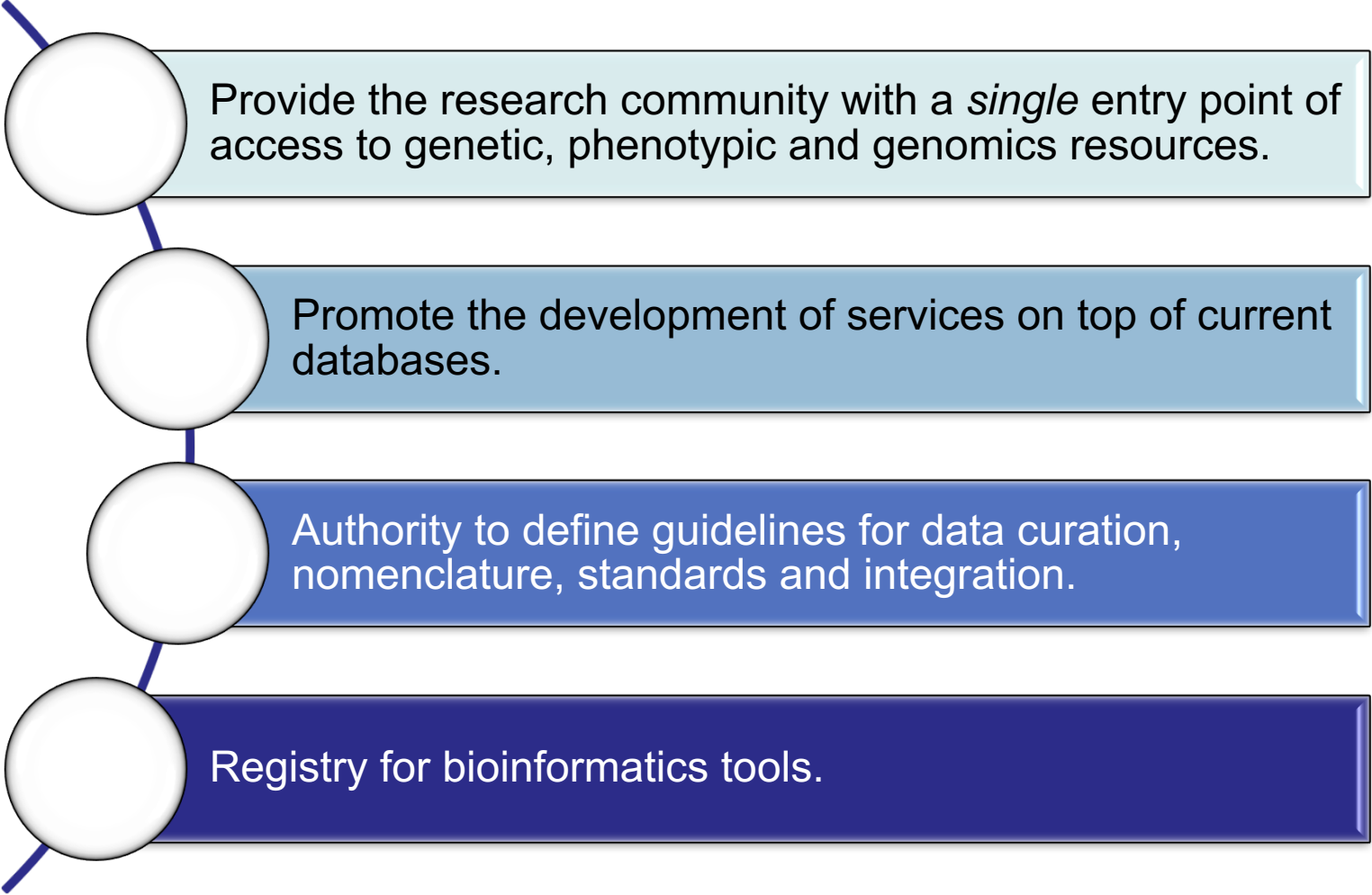
**IFB:** French Bioinformatics Institute French Elixir node



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



# Cyber-infrastructure goals



Provide the research community with a *single* entry point of access to genetic, phenotypic and genomics resources.

Promote the development of services on top of current databases.

Authority to define guidelines for data curation, nomenclature, standards and integration.

Registry for bioinformatics tools.

# IFB/Elixir – Plant Bioinformatics





**The French Plant bioinformatics  
node  
(33 FTEs)**



**Genetics and genomics resources  
for plants and crop parasites  
(INRA)**

**South Green<sup>®</sup>  
bioinformatics platform**

**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)**

**BIPAA**

**Arthropods for  
Agroecosystems  
(INRA)**

# Services



analysis

genomics annotation

software hosting

data repository

data integration

database design

software engineering



A person is seen from behind, walking away on a stone-paved path. They are carrying two large, heavy bundles of harvested wheat stalks, one on each side, balanced on a long wooden pole across their shoulders. The person is wearing a green and white striped jacket and dark trousers. The path leads through a lush, green landscape with scattered rocks and dense vegetation. In the background, there are rolling hills and a bright sun shining from the upper right, creating a lens flare effect. The overall scene is peaceful and rural.

**WheatIS – An information system for  
the wheat international research  
community**



# WheatIS Expert Working Group



Wheat initiative  
commissioned a  
WheatIS EWG

*Expert  
Working  
Group*

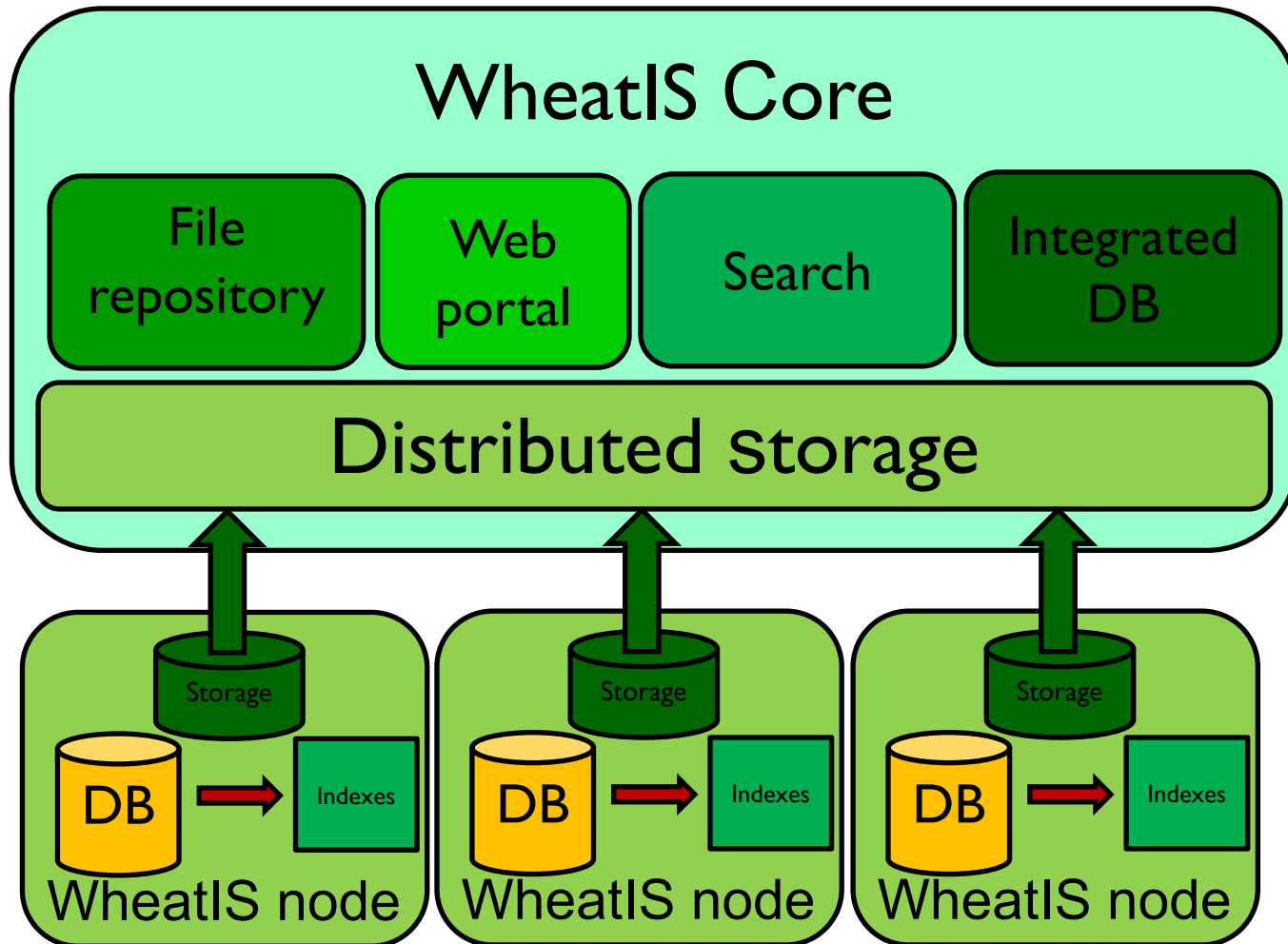


- Build projects
- Build infrastructure
- Report to the Wheat Initiative

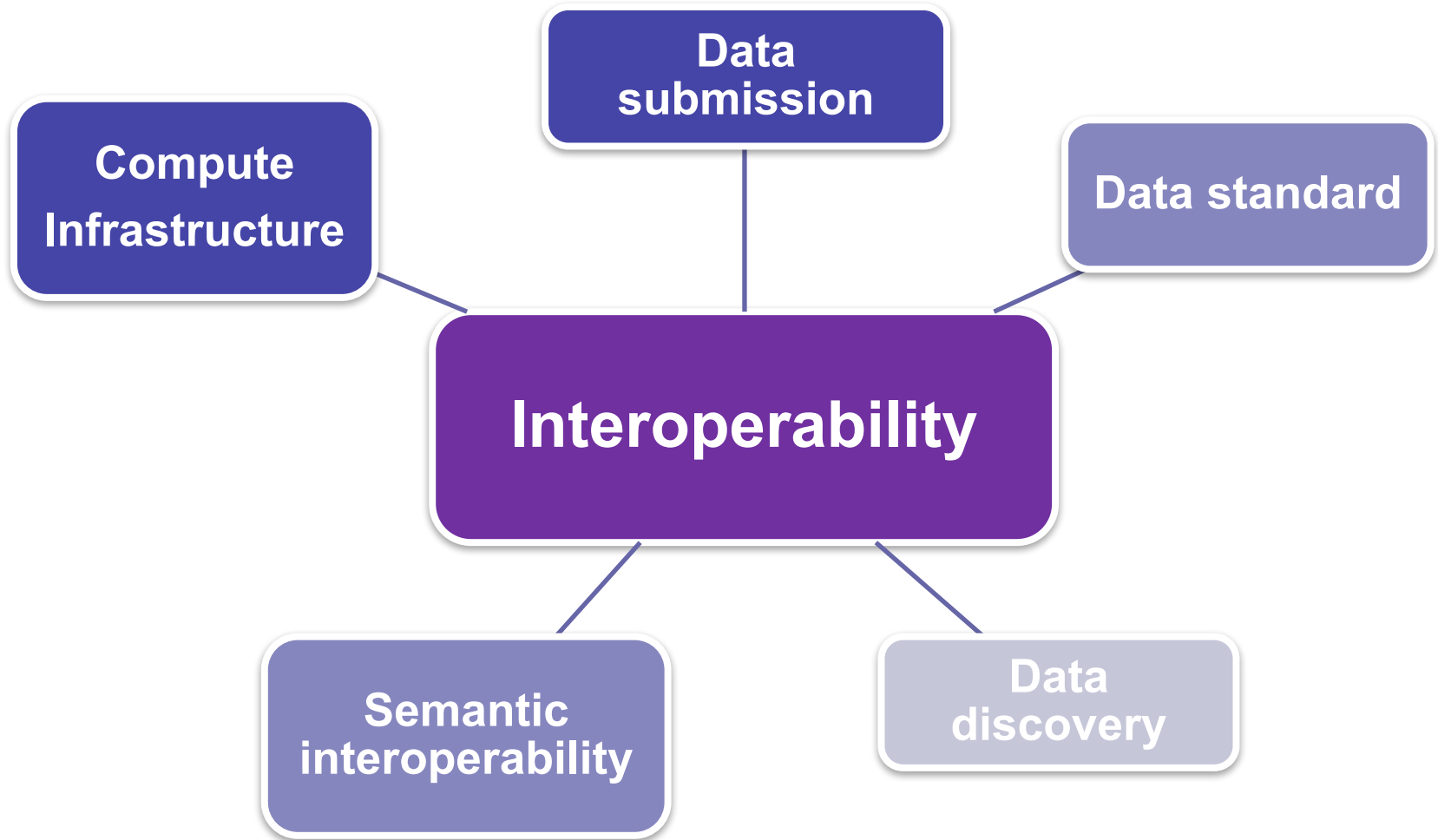
# WheatIS nodes (#12)



# WheatIS Architecture



# Cyber-infrastructures key issue



# Data must follow FAIR principles

*force11.org*

## To be **F**indable:

- 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 **A**ccessible:

- 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 **I**nteroperable:

- 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 **R**e-usable:

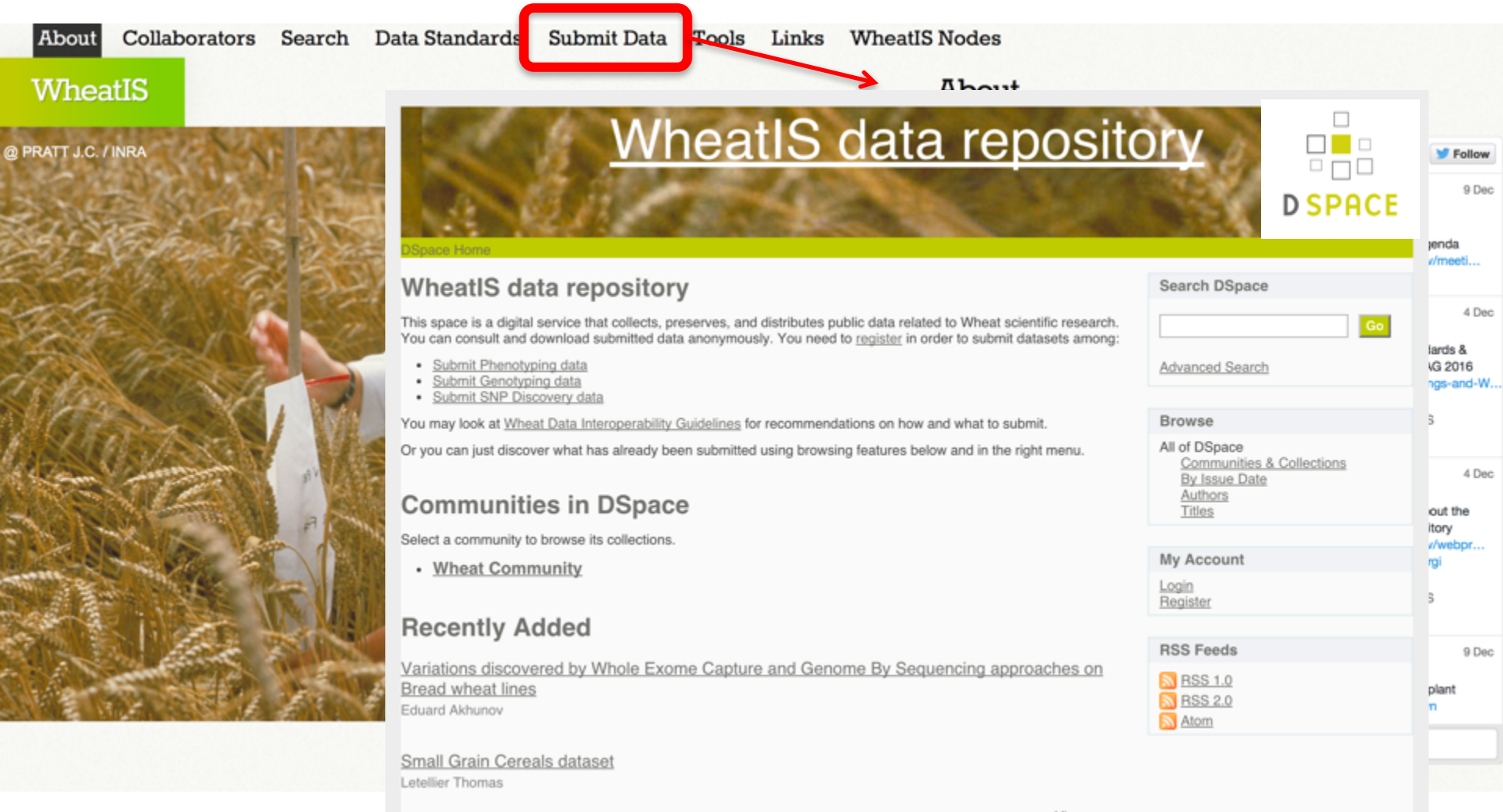
- R1. meta(data) have a plurality of accurate and **relevant attributes**.
  - R1.1. (meta)data are released with a clear and accessible data usage **license**.
  - R1.2. (meta)data are associated with their provenance.
  - R1.3. (meta)data meet domain-relevant community



# DATA SUBMISSION



# Data file repository



The screenshot shows the WheatIS data repository website. At the top, a navigation bar includes links for 'About', 'Collaborators', 'Search', 'Data Standards', 'Submit Data', 'Tools', 'Links', and 'WheatIS Nodes'. The 'Submit Data' link is highlighted with a red box and a red arrow pointing to the main content area. The main content area features a large banner with the text 'WheatIS data repository' and the DSpace logo. Below the banner, there is a section titled 'WheatIS data repository' with a description: 'This space is a digital service that collects, preserves, and distributes public data related to Wheat scientific research. You can consult and download submitted data anonymously. You need to register in order to submit datasets among:'. A list of links follows: 'Submit Phenotyping data', 'Submit Genotyping data', and 'Submit SNP Discovery data'. Below this, there is a section titled 'Communities in DSpace' with a link to 'Wheat Community'. Another section titled 'Recently Added' lists 'Variations discovered by Whole Exome Capture and Genome By Sequencing approaches on Bread wheat lines' by Eduard Akhunov and 'Small Grain Cereals dataset' by Letellier Thomas. On the right side, there is a sidebar with a search box, a 'Browse' section, a 'My Account' section, and an 'RSS Feeds' section.

Navigation: About Collaborators Search Data Standards **Submit Data** Tools Links WheatIS Nodes

WheatIS

@ PRATT J.C. / INRA

## WheatIS data repository

DSpace Home

**WheatIS data repository**

This space is a digital service that collects, preserves, and distributes public data related to Wheat scientific research. You can consult and download submitted data anonymously. You need to [register](#) in order to submit datasets among:

- [Submit Phenotyping data](#)
- [Submit Genotyping data](#)
- [Submit SNP Discovery data](#)

You may look at [Wheat Data Interoperability Guidelines](#) for recommendations on how and what to submit.

Or you can just discover what has already been submitted using browsing features below and in the right menu.

### Communities in DSpace

Select a community to browse its collections.

- [Wheat Community](#)

### Recently Added

[Variations discovered by Whole Exome Capture and Genome By Sequencing approaches on Bread wheat lines](#)  
Eduard Akhunov

[Small Grain Cereals dataset](#)  
Letellier Thomas

Search DSpace

Advanced Search

My Account

RSS Feeds

Follow

9 Dec

9 Dec

4 Dec

9 Dec

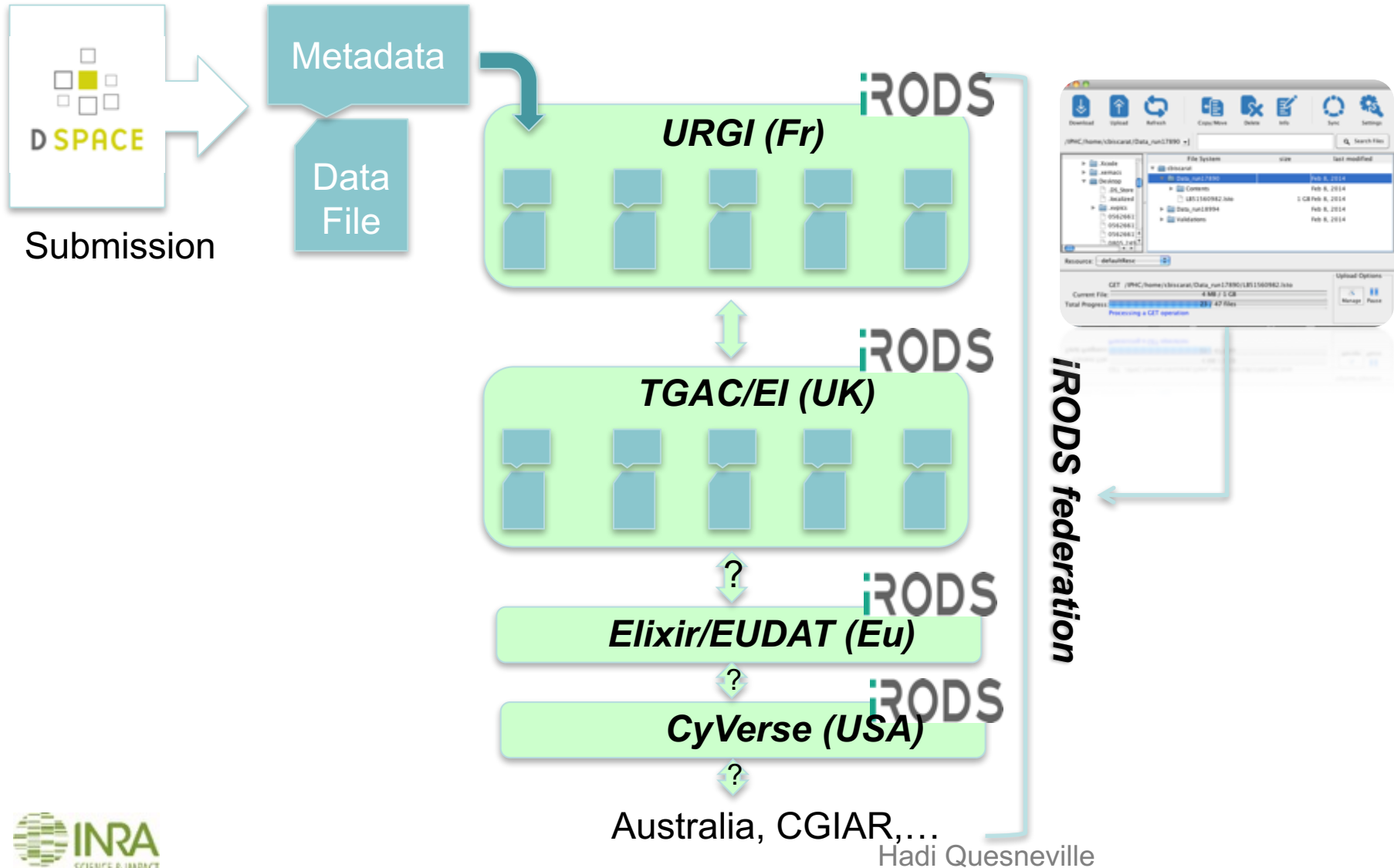
4 Dec

9 Dec

9 Dec

9 Dec

# Data workflow





# DATA STANDARDS



@ PRATT J.C. / INRA



## Wheat Data Interoperability Guidelines

[Home](#) [Guidelines](#) [Ontologies & Vocabularies](#) [Use cases](#) [Getting involved](#) [About](#)



### Welcome

These recommendations have been prepared by members of the [Wheat Data Interoperability Working Group \(WG\)](#), one of the WGs of the [Research Data Alliance](#) and the only WG of the [Agriculture Data Interoperability Interest Group](#). The group is coordinated by members of the [Wheat Initiative](#), a global initiative that aims to reinforce synergies between bread and durum wheat national and international research programmes to increase food security, nutritional value and safety while taking into account societal demands for sustainable and resilient agricultural production systems.

#### GETTING INVOLVED



WheatIS

#### More specifically, the WG aims to:



**PROMOTE**  
the adoption of common standards, vocabularies and best practices for Wheat data management



**FACILITATE**  
access, discovery and reuse of wheat data



**FACILITATE**  
wheat data integration



Guidelines



Ontologies & Vocabularies



Use Cases



# Wheat Data Interoperability Guidelines

Home **Guidelines** Ontology



## Welcome

These recommendations have been developed by the **Wheat Group (WG)**, one of the **Wheat Data Interoperability Interest Groups**, an initiative that aims to reinforce research programmes to increase societal demands for sustainable wheat production.



**PROMOTE**  
the adoption of common standards, vocabularies and best practices for Wheat Data management



Guidelines

Wheat Data Interoperability Group

Home / Phenotypes

## Phenotypes

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

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

### Recommendations

#### Summary

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

#### 1. Data formats

We recommend following minimum format principles with data matrices (trait along with method, units and scales or environment and germplasm).

**ISA-Tab** is an implementation of this principle. It consists of one or more files and metadata files, the latter being used for data discovery; information can be found on this dedicated [page](#) or in [this presentation](#) Genome 2015. It is currently well suited for generation by software, phenotype specific configuration and tools are under development. See the [germplasm recommendations](#) for data format regarding

#### 2. Metadata and vocabularies

##### Observation variables

Observation Variables include trait and environment variables.

We recommend using existing variables, listed in the vocabulary. To create new observation variables, we recommend using the [Template](#) available at [Crop Ontology](#) website. It must include a description, abbreviation, synonyms, methods, and scales) to be created and shared. The most important field in this template remain stable and never be modified. Furthermore it must never be deprecated if needed. This way, it can be used in trials and remain

For Nursery and Trial metadata and description we recommend using [Crop Ontology](#), which describes the terms related to nurseries and trial environments, study design, etc. These metadata are actively maintained. For biorefinery, we recommend using the [Biorefinery ontology](#) and terms associated with biomass composition and characteristics (area, particle size, porosity, etc.), physico-chemical pretreatment experimental processes descriptions.

##### Recommended Variable ontologies and vocabularies

- Wheat crop ontology
- INRA Wheat Ontology (soon publicly available)
- Wheat Phenotype
- Biorefinery ontology
- XEQ, XEM, Environment Ontology

For the difference between metadata, ontologies and vocabularies

## 3. Raw data

We recommend sharing at least clean documented raw data, like plant phenotype data lifecycle begins with acquisition, then cleaning, elaboration combines several variables, like phenological stages and trait elaborated/computed variables used as input for analyses softwares. and phenology can be combined to get height at flowering. Different data produced for different purposes, it is therefore important to be able to access ones from raw data.

### Some popular Tools

#### 1. Repositories, information systems and data tools

The Breeding Management System, **BMS** generates standard format for trial data in fields and uses for variables the [Crop Research Ontology](#) for experiment related metadata and trait related ontologies of the [Crop Ontology](#) makes it possible to analyze data directly using statistical tools such as [R](#).

**GrpIS** is an INRA information system designed for plant and pest genetic scientists to mine genomic, phenomic and genetic data. For phenomic data discovery through a keyword based, google like, [search engine](#). The latter allows [dataset building](#) for genetic or phenomic analysis. It is based on a strict identification of germplasm on variables through the [Crop Ontology](#).

The [Breeding API](#) specifies a standard interface for plant phenotype/germplasm to 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 acid yields can be found through the [@Web platform](#). The Documents tab offers information by a kind of pretreatment (topics Bioref-XX). Data available pretreatments used, biomass types and characterization, etc. In the future possible to find the best match pretreatment-phenotype.

[iPlant collaborative](#) offers many services that allow the analysis of genomic and phenotypic data.

#### 2. Data acquisition

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

### Things to follow in the future

#### Candidate formats

- ped
- Bagit
- Hadoop File System : HDFS

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 detailed since we are not direct users of this system. Feel free to send us a more detailed description, we will work on its integration.

Reply

## Leave a Reply

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

Comment

Text input area for comment

Name \*

Text input field for name

Email \*

Text input field for email

Website

Text input field for website

Post Comment



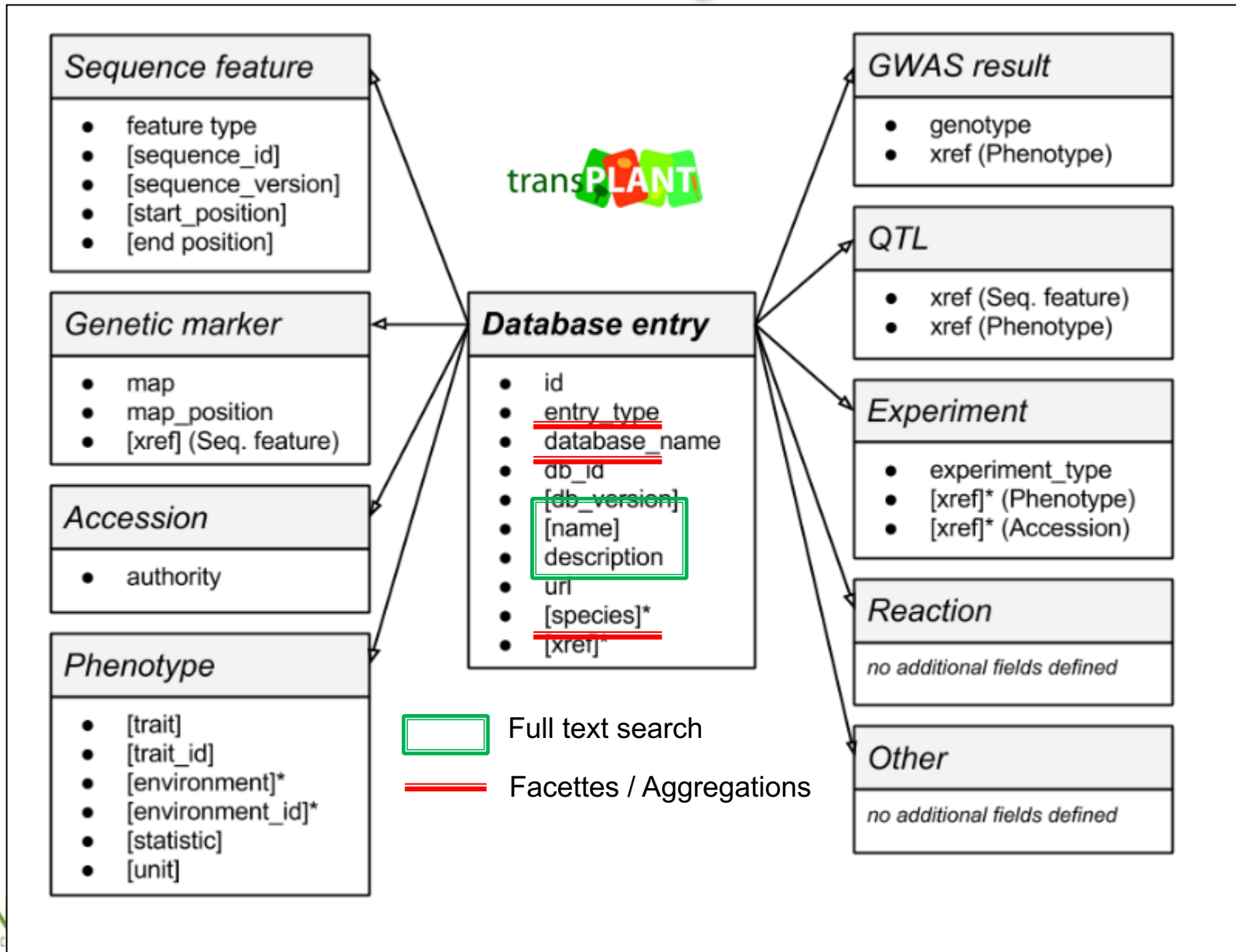




# DATA DISCOVERY



# Data Discovery data model





# Full text search of distributed databases

EnsemblPlants  
 EMBL-EBI  
 IPK GATERSLEBEN  
 HELMHOLTZ GEMEINSCHAFT  
 PLANTS DB  
 IGR pan  
 ZURGI  
 TGAC The Genome Analysis Centre™  
 GrainGenes A Database for Triticeae and Avena  
 Gramene  
 CIMMYT International Maize and Wheat Improvement Center



ZURGI  
 South Green bioinformatics platform  
 PHENOME Réseau Français Phénomique végétale FPPN  
 arcad



transPLANT  
 WheatIS  
 ifb

Search results for 'WheatIS nodes':  
 transPlant-MPS (UP):  
 + CrowsNest: 13324  
 transPlant-IPK (UP):  
 + CR-EST: 199220  
 + GEBIS: 52976  
 + MetaCrop: 355  
 transPlant-EBI (UP):  
 + Ensembl Plants: 215282  
 URGH (UP):  
 + GpGIS: 1757714  
 + WheatDBMine: 271197  
 + GpGIS JBrowse: 232783  
 + WheatIS repository: 13  
 transplant-IPGPAS (UP):  
 + PlantPhenoDB: 2  
 T3 (UP):  
 + Triticeae Toolbox: 171159

Search in all WheatIS nodes...  
 Examples: [fbh](#), [wmc430](#), [Triboum](#), [TRAE53BF001000010CFD](#)



## WheatIS

Search

About

WheatIS nodes:

transPlant-MIPS (UP):

- CrowsNest: 13324

transPlant-IPK (UP):

- CR-EST: 199220

- GEBIS: 52678

- MetaCrop: 355

transPlant-EBI (UP):

- Ensembl Plants: 21828;

transplant-iPGPAS (UP):

- PlantPhenoDB: 2

T3 (UP):

- Triticeae Toolbox: 1711;

CIMMYT (UP):

- CIMMYT Dspace: 918

- CIMMYT dataverse: 37

URGI (UP):

- GnpIS: 175714

- Wheat3BMine: 271197

- GnpIS JBrowse: 23278;

## WheatIS

Filters

Clear

### Database

TRITICEAE TOOLBOX (64)

CR-EST (7)

GNPIS (3)

ENSEMBL PLANTS (1)

GNPIS JBrowse (1)

PLANTPHENODB (1)

### Type

ACCESSION (42)

PHENOTYPE (14)

EXPERIMENT (9)

EXPRESSED SEQUENCE

TAGS (7)

SEQUENCE FEATURE (2)

PHENOTYPE (1)

QTL (1)

SEQUENCE FEATURE (1)

### Species

TRITICUM AESTIVUM (69)

HORDEUM VULGARE (6)

TRITICUM AESTIVUM L. (1)

TRITICUM DURUM (1)

Search

About

WheatIS nodes:

transPlant-MIPS (UP):

- CrowsNest: 13324

transPlant-IPK (UP):

- CR-EST: 199220

- GEBIS: 52678

- MetaCrop: 355

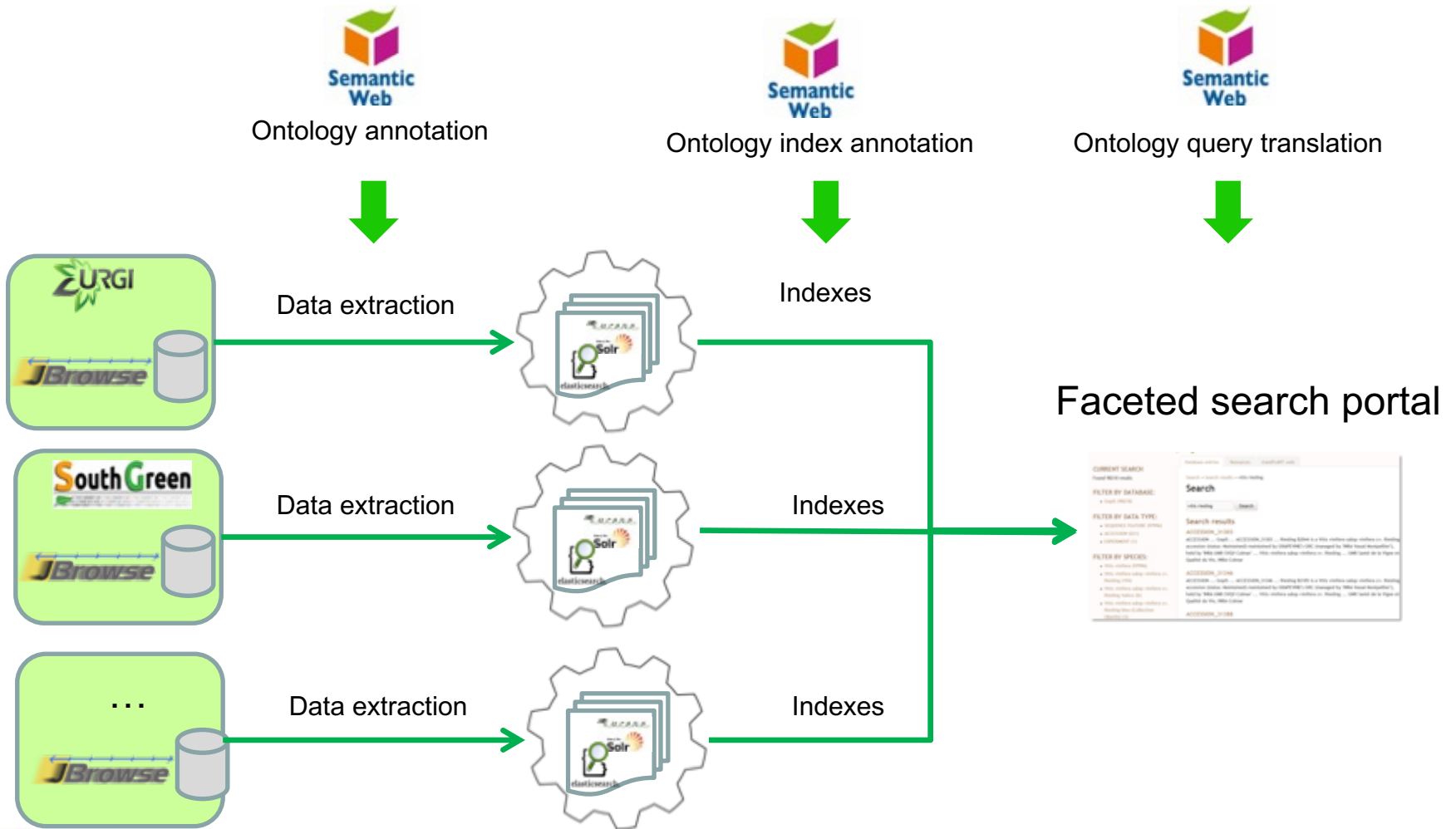
fhb

1-10 of 77

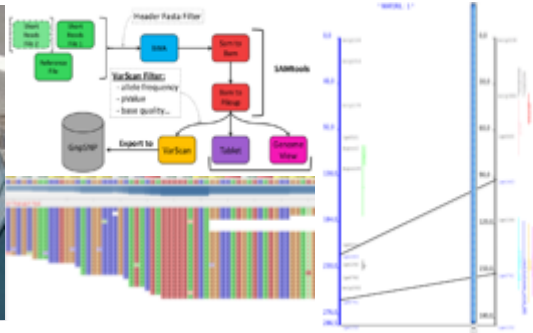
10 results per page

ID	Source	Type	Taxon	Description
<a href="#">Traes_5DL_E12C501B4</a>	Ensembl Plants	-	Triticum aestivum	Sequence feature, Ensembl Plants, Traes_5DL_E12C501B4, Traes_5DL_E12C501B4, Multiple inositol polyphosphate phosphatase Phylla1 [Source:UniProtKB/TrEMBL;Acc:A0FHB0], Triticum aestivum, protein_coding, 5D
<a href="#">HDP14M22T</a>	CR-EST	-	Hordeum vulgare	HDP14M22T, expressed sequence tags, CR-EST, Hordeum vulgare, gi 26248924 ref NP_754964.1  Hypothetical protein ythB [Escherichia coli CFT073] Hypothetical protel; gi 28951047 gb AAO63447.1  A2g37930 [Arabidopsis thal[...]
<a href="#">HDP20D01w</a>	CR-EST	-	Hordeum vulgare	HDP20D01w, expressed sequence tags, CR-EST, Hordeum vulgare, gi 15604676 ref NP_221194.1  SFHB PROTEIN HOMOLOG (sfhB) [Rickettsia prowazekii str. Madrid E] SFHB ; gi 34906406 ref NP_914550.1  P0710E05.16 [Oryza sativa [...]
<a href="#">HDP20D01T</a>	CR-EST	-	Hordeum vulgare	HDP20D01T, expressed sequence tags, CR-EST, Hordeum vulgare, gi 15604676 ref NP_221194.1  SFHB PROTEIN HOMOLOG (sfhB) [Rickettsia prowazekii str. Madrid E] SFHB ; gi 34906406 ref NP_914550.1  P0710E05.16 [Oryza sativa [...]
<a href="#">HDP21C08T</a>	CR-EST	-	Hordeum vulgare	HDP21C08T, expressed sequence tags, CR-EST, Hordeum vulgare, gi 15604676 ref NP_221194.1  SFHB PROTEIN HOMOLOG (sfhB) [Rickettsia prowazekii str. Madrid E] SFHB ; gi 31979237 gb AAP68831.1  bone morphogenetic protein 1[...]
<a href="#">HDP31N10w</a>	CR-EST	-	Hordeum vulgare	HDP31N10w, expressed sequence tags, CR-EST, Hordeum vulgare, gi 15604676 ref NP_221194.1  SFHB PROTEIN HOMOLOG (sfhB) [Rickettsia prowazekii str. Madrid E] SFHB ; gi 34906406 ref NP_914550.1  P0710E05.16 [Oryza sativa [...]
<a href="#">HDP35A10T</a>	CR-EST	-	Hordeum vulgare	HDP35A10T, expressed sequence tags, CR-EST, Hordeum vulgare, gi 26248924 ref NP_754964.1  Hypothetical protein ythB [Escherichia coli CFT073] Hypothetical protel; gi 28951047 gb AAO63447.1  A2g37930 [Arabidopsis thal[...]
<a href="#">TS034O07u</a>	CR-EST	-	Triticum aestivum	TS034O07u, expressed sequence tags, CR-EST, Triticum aestivum, Gi 15233419 ref NP_192328.1  hypothetical protein [Arabidopsis thaliana] gi 7487460 pir T01820 hypo; Gi 15604676 ref NP_221194.1  SFHB PROTEIN HOMOLOG (sfhB) [...]
<a href="#">HWW FHB</a>	Triticeae Toolbox	Experiment	Triticum aestivum	Experiment, Triticeae Toolbox, HWW FHB, phenotype experiment, Includes trials FHB_2014_Lincoln, HWWFHB_2014_Brookings, HWWFHB_2014_Fargo, Triticum aestivum, phenotype
<a href="#">URSN_2012_BrookingsSD</a>	Triticeae Toolbox	Experiment	Triticum aestivum	Experiment, Triticeae Toolbox, URSN_2012_BrookingsSD, phenotype trial, traits=Fusarium head blight incidence, Fusarium head blight severity, Fusarium head blight disease index, visually scabby kernels URSN, descrip[...]

# Next for data discovery

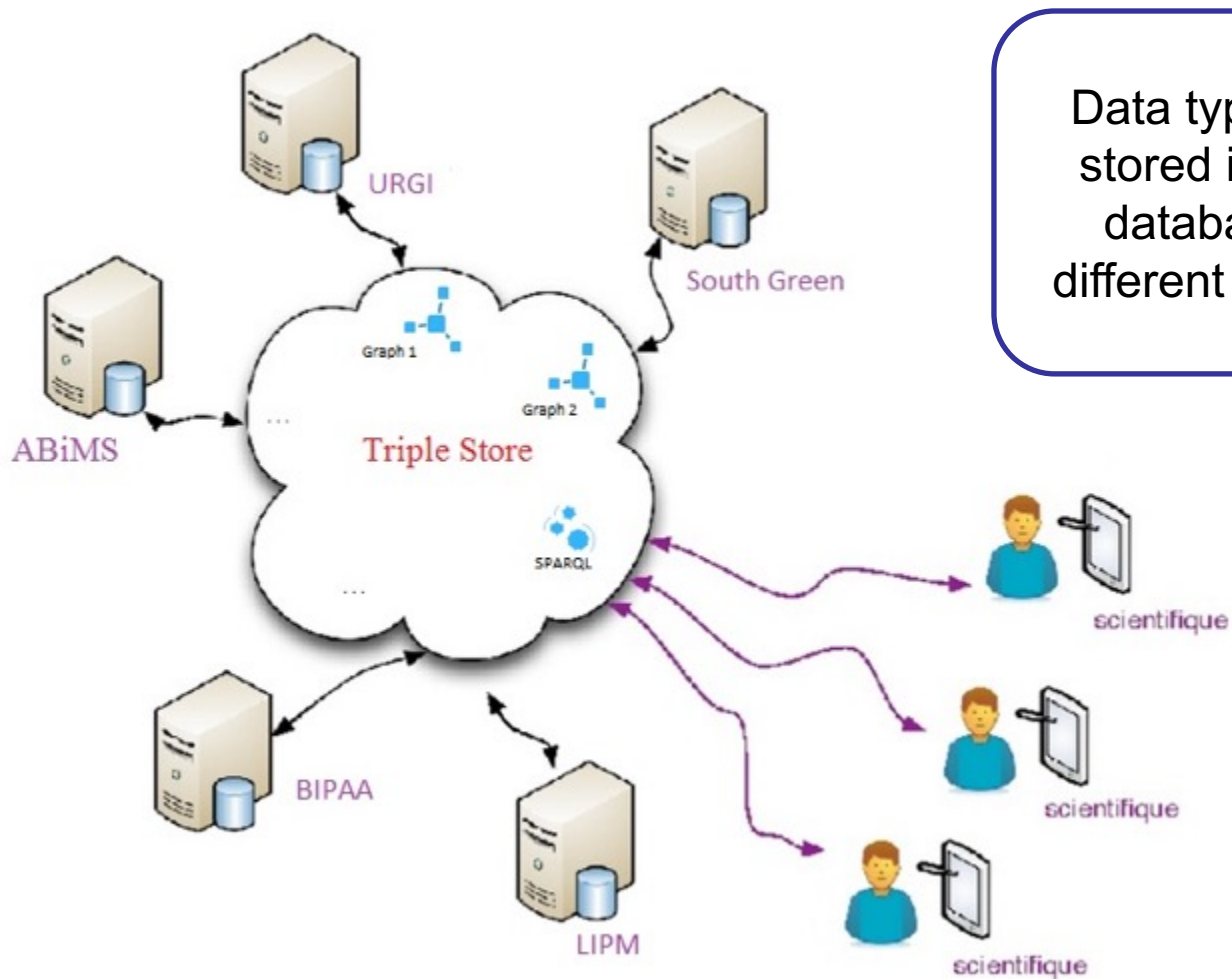


# Semantic Interoperability





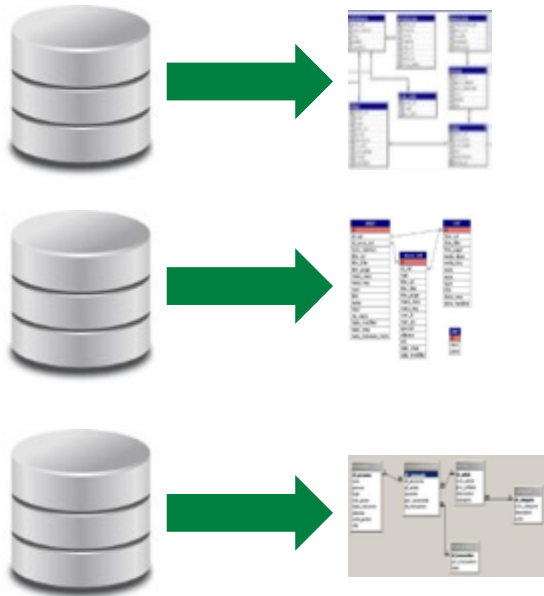
# Why a semantic interoperability



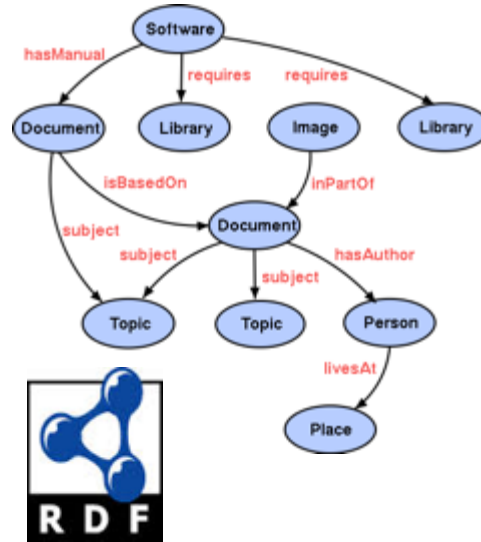
Data types can be stored in different databases with different data model



# Develop a web semantic interoperability



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



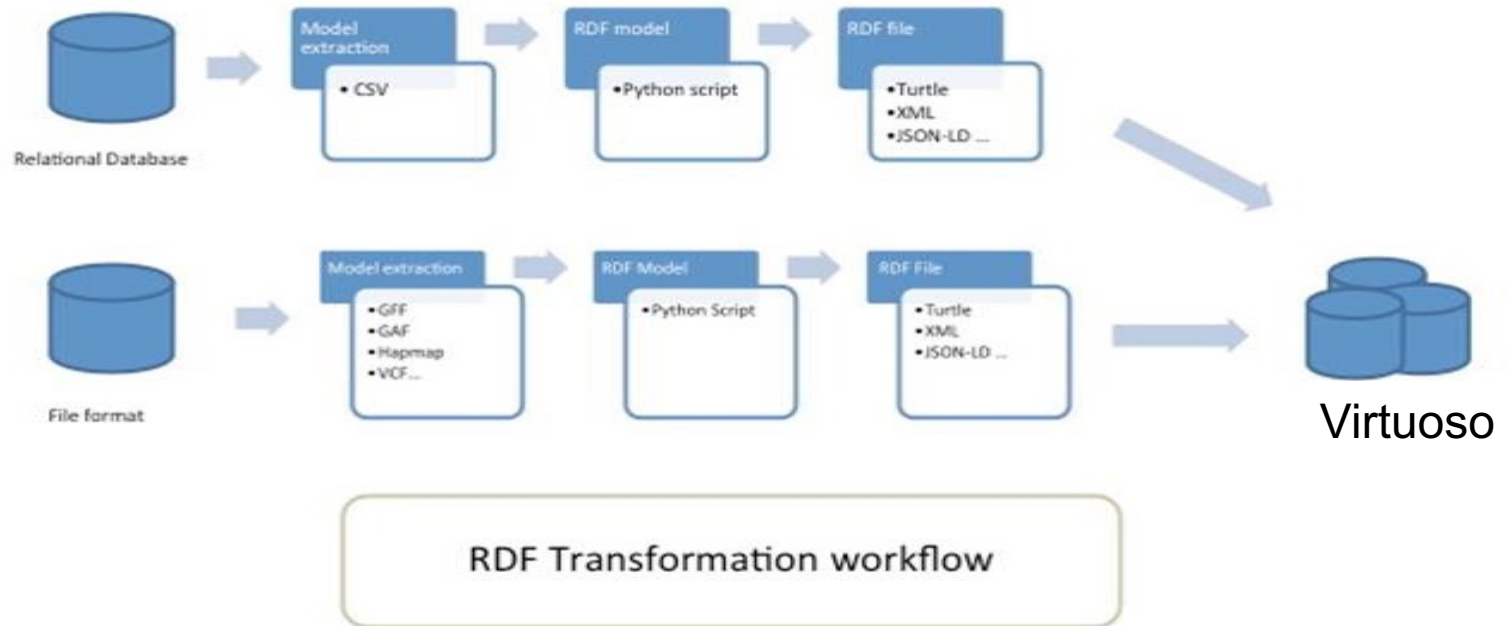
RDF modeling of the databases schemas

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](http://www.agrold.org))

- Different user profiles
- Provides different entry points

## 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>

# 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

Tools 

search tools 

**Get Data**

**Text Manipulation**

**Filter and Sort**

**Join, Subtract and Group**

**Convert Formats**

**Operate on Genomic Intervals**

**FASTA manipulation**

**Emboss Tools**

**NGS TOOLS**

[NGS: QC and manipulation](#)

[NGS: Mapping](#)

[NGS: Assembly](#)

[NGS: RNA Analysis](#)

[NGS: SAM Tools](#)

[NGS: Peak Calling](#)

[Alternative splicing tools](#)

[Picard Tools](#)

[BEDTools](#)

**VARIANT/SNP TOOLS**

[URGI: MAPHiTS 2.0](#)

[SnpEff tools](#)

[Vcf utils](#)

[VCF Tools](#)

**ANNOTATION TOOLS**

[Gene Ontology Annotation tools](#)

[Repeats annotation](#)

**SEQUENCE SIMILARITY TOOLS**

[NCBI BLAST+](#)

**URGI TOOLS**

[URGI: Get Data for grapevine](#)

[URGI: BAC analysis](#)

[URGI: MAPHiTS - PreProcess Tools](#)

[URGI: MAPHiTS - Tools](#)

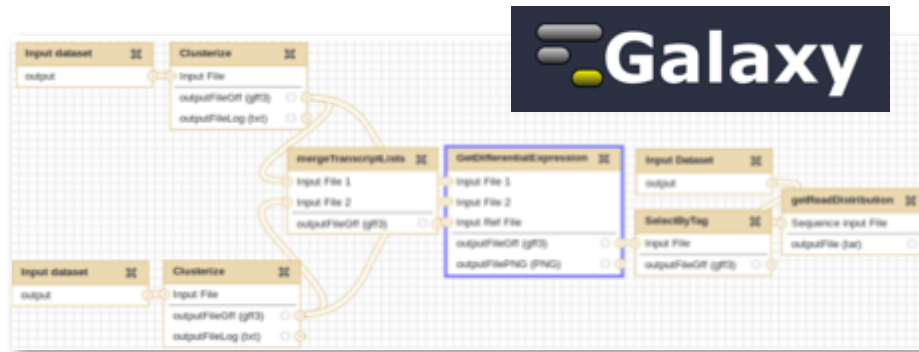
[URGI: MAPHiTS - PostProcess Tools](#)

[URGI: MAPHiTS - SNPs Chip Tools](#)

[URGI: S-MART](#)

[URGI: Differential expression Analysis Tools](#)

[Tools for APLIBIO pipelines](#)



## 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

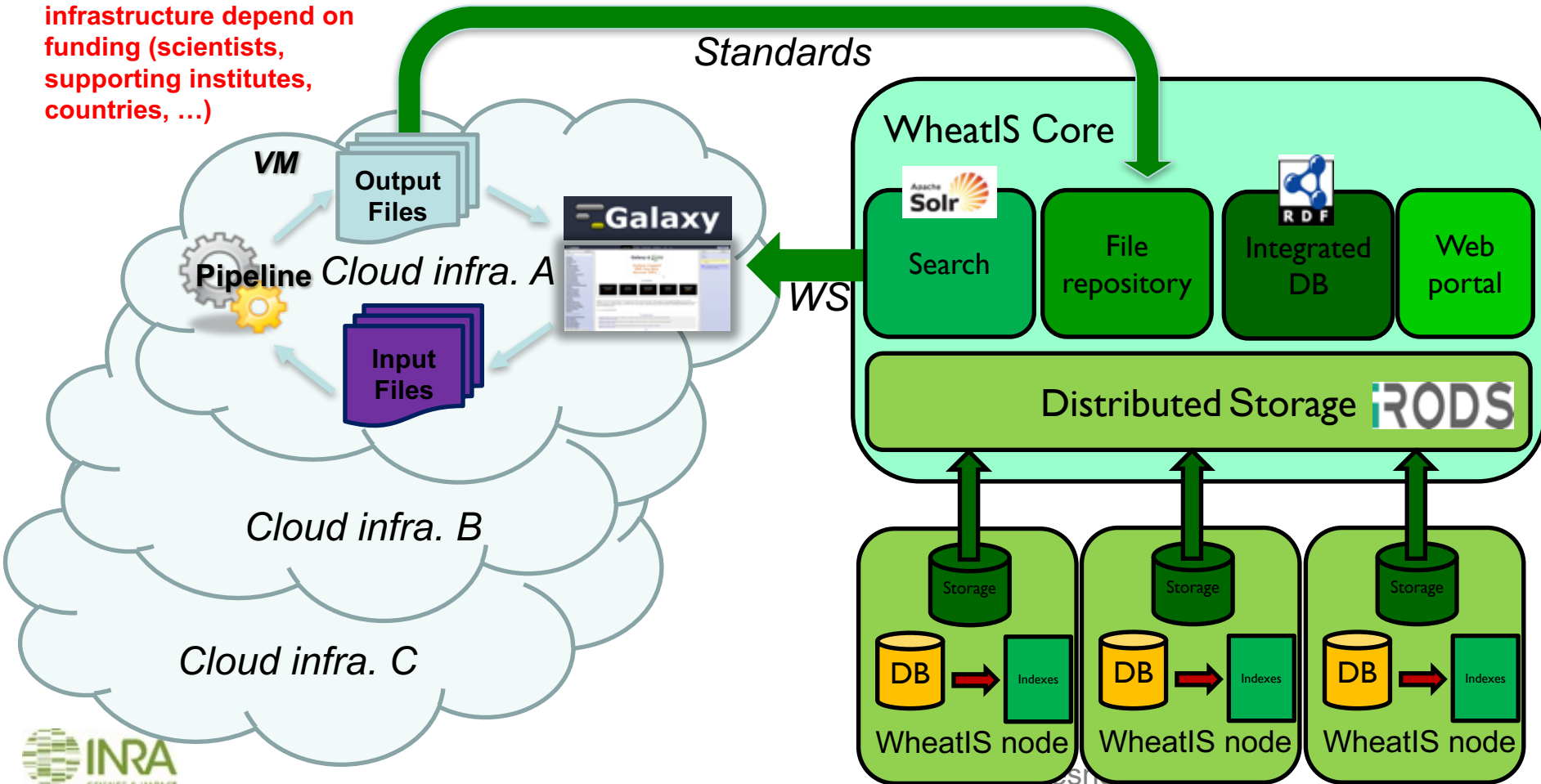
Several Compute infrastructures

One data infrastructure



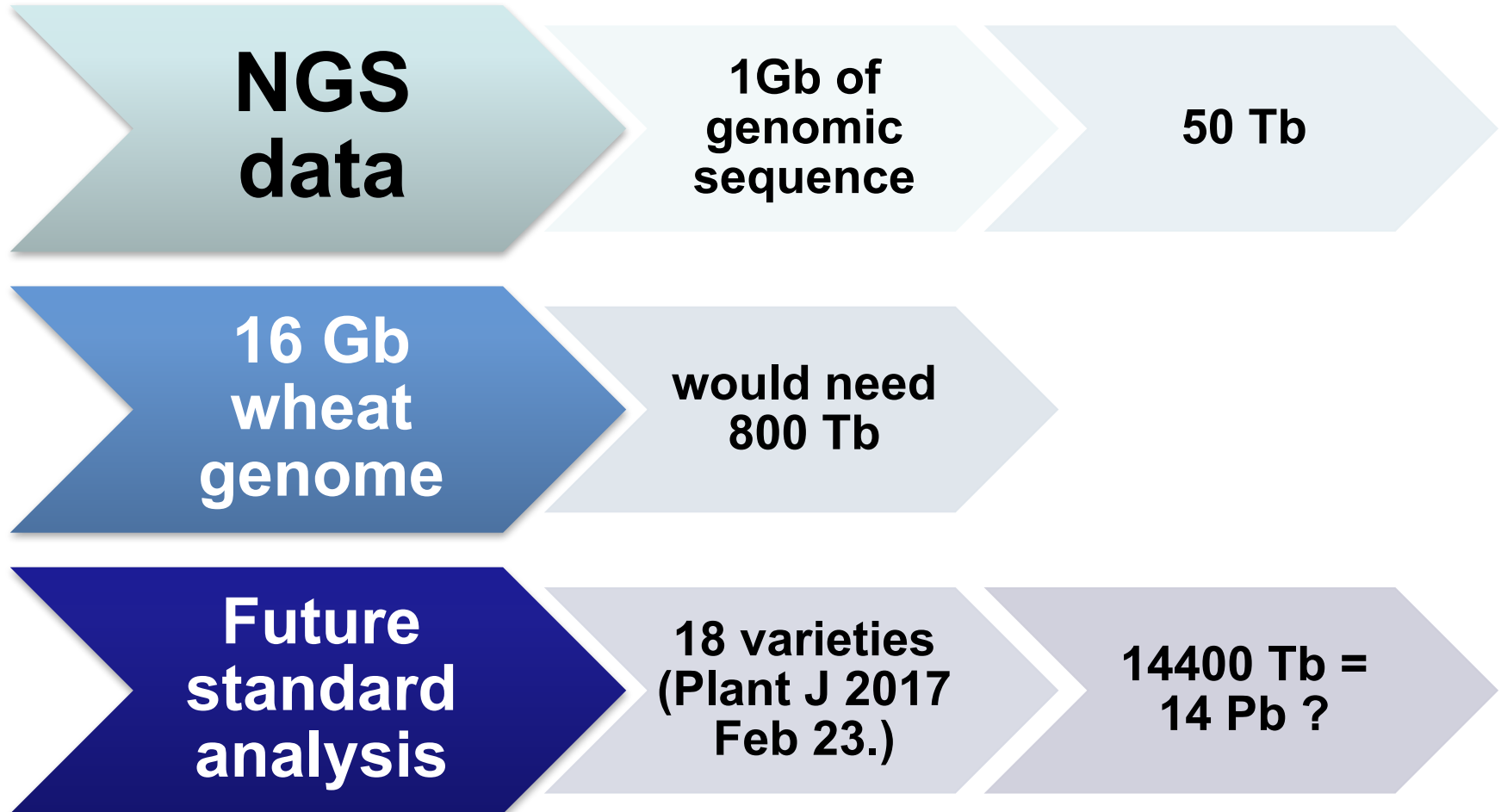
Availability of infrastructure depend on funding (scientists, supporting institutes, countries, ...)

Standards





# 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



Safe data  
transfert  
protocoles

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

# Efficient workflows



## Data reduction

- Noise reduction
- Fit to computing resources
- Store on adapted media: quick access vs storage capacity
- Keep only data that cannot be recomputed easily

## Data dependancies

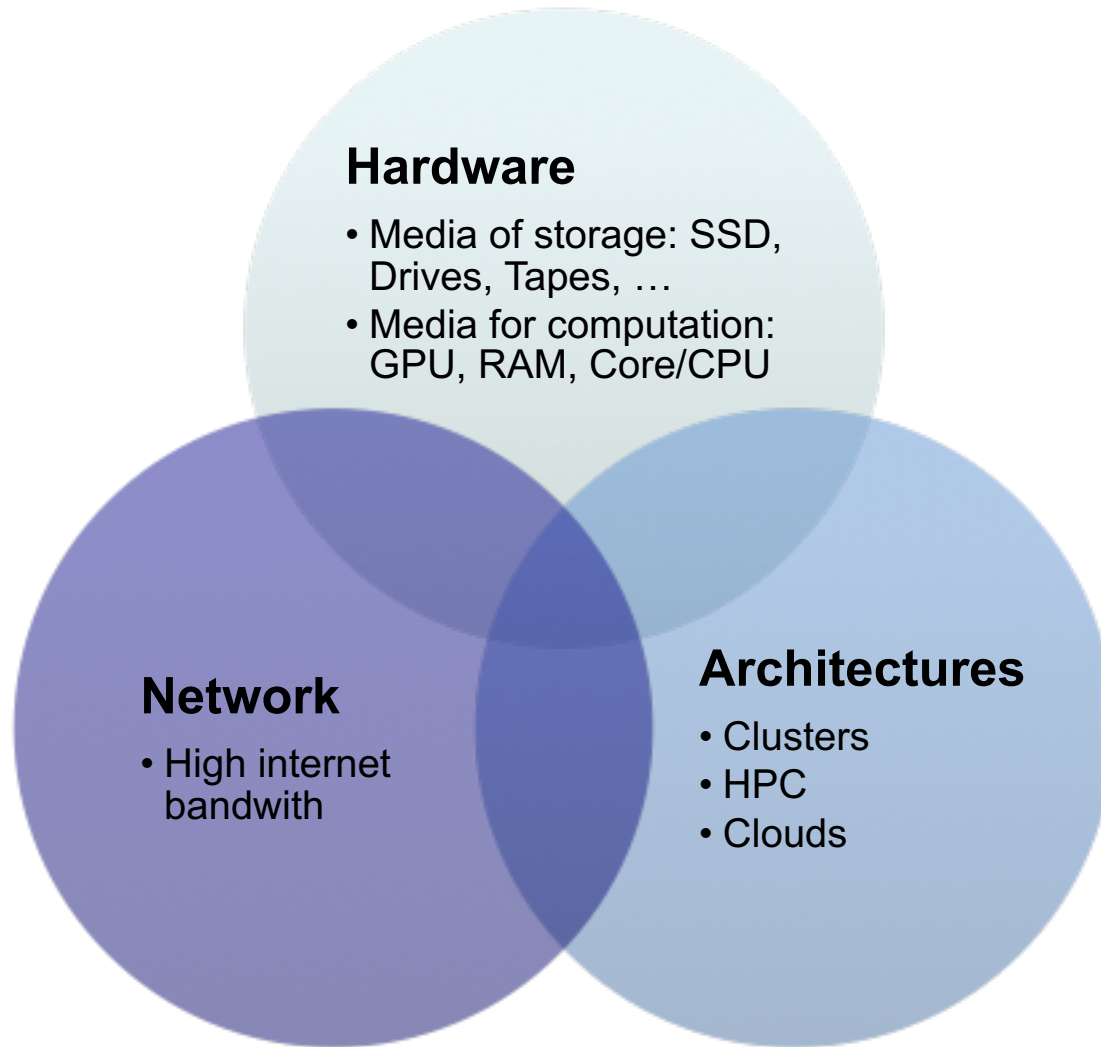
- Parallelization
- Compute time estimation

## Analysis reproducibility

- Notebooks
- Follows workflow process
- Fault tolerant → resume from previous steps

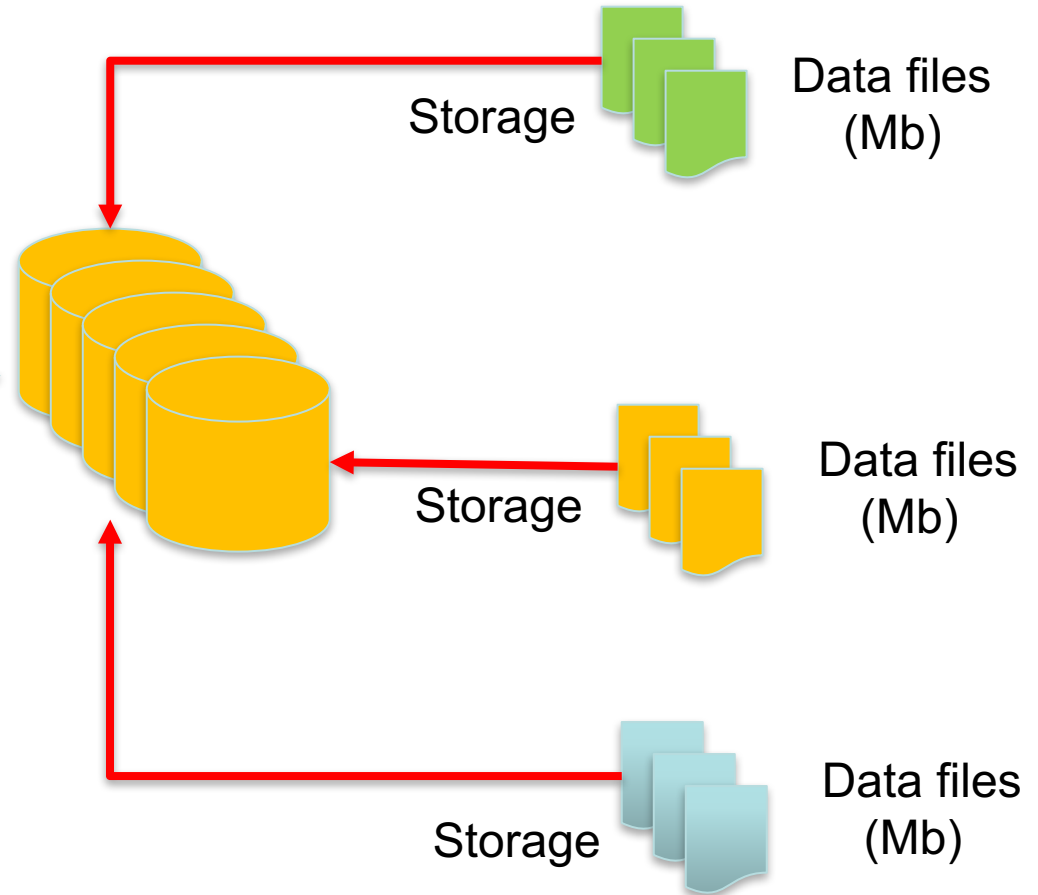


# Efficient computer resources

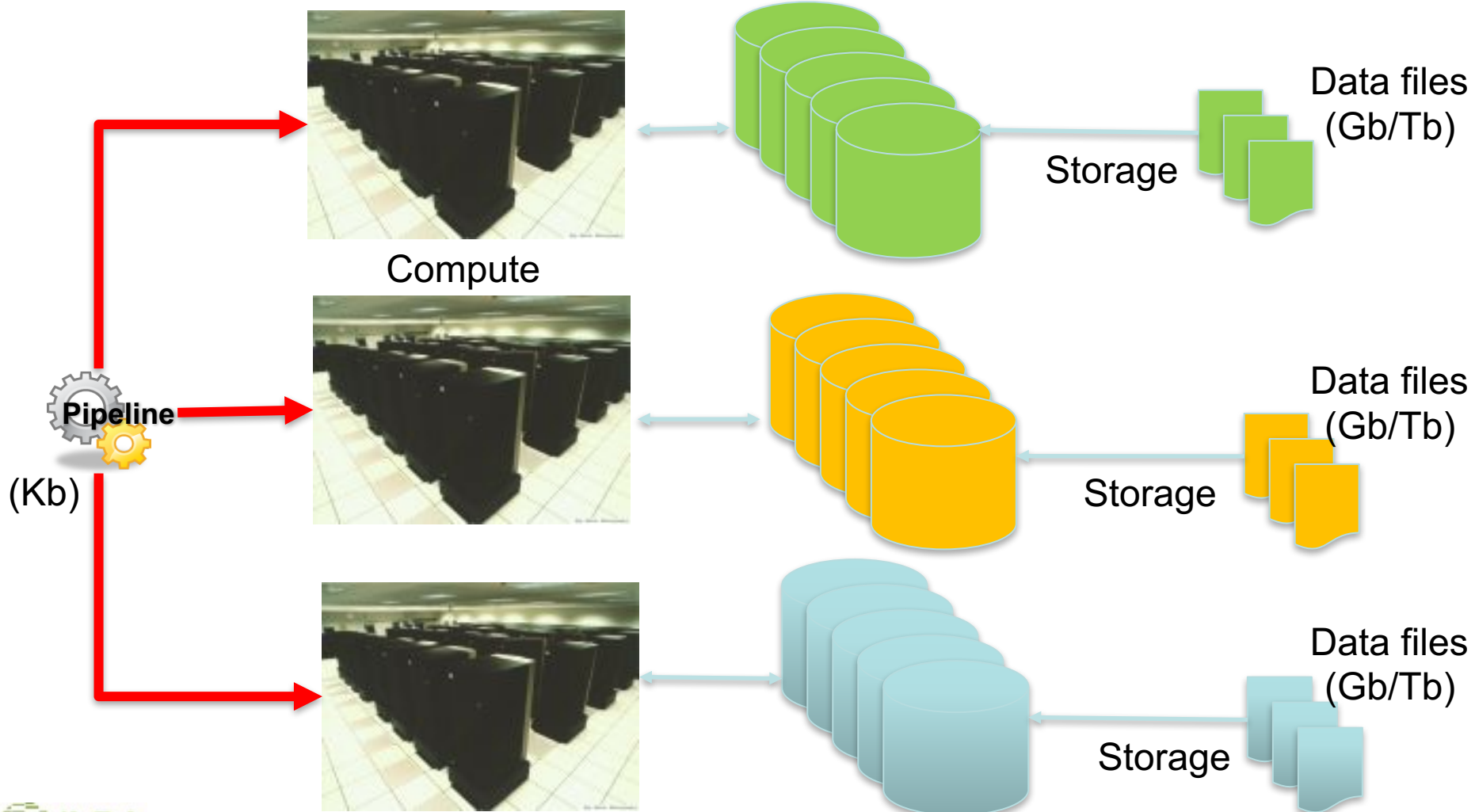


# Today's paradigm: moving data to tools

Compute



# Paradigm shift: moving tools to data



# Moving workflows

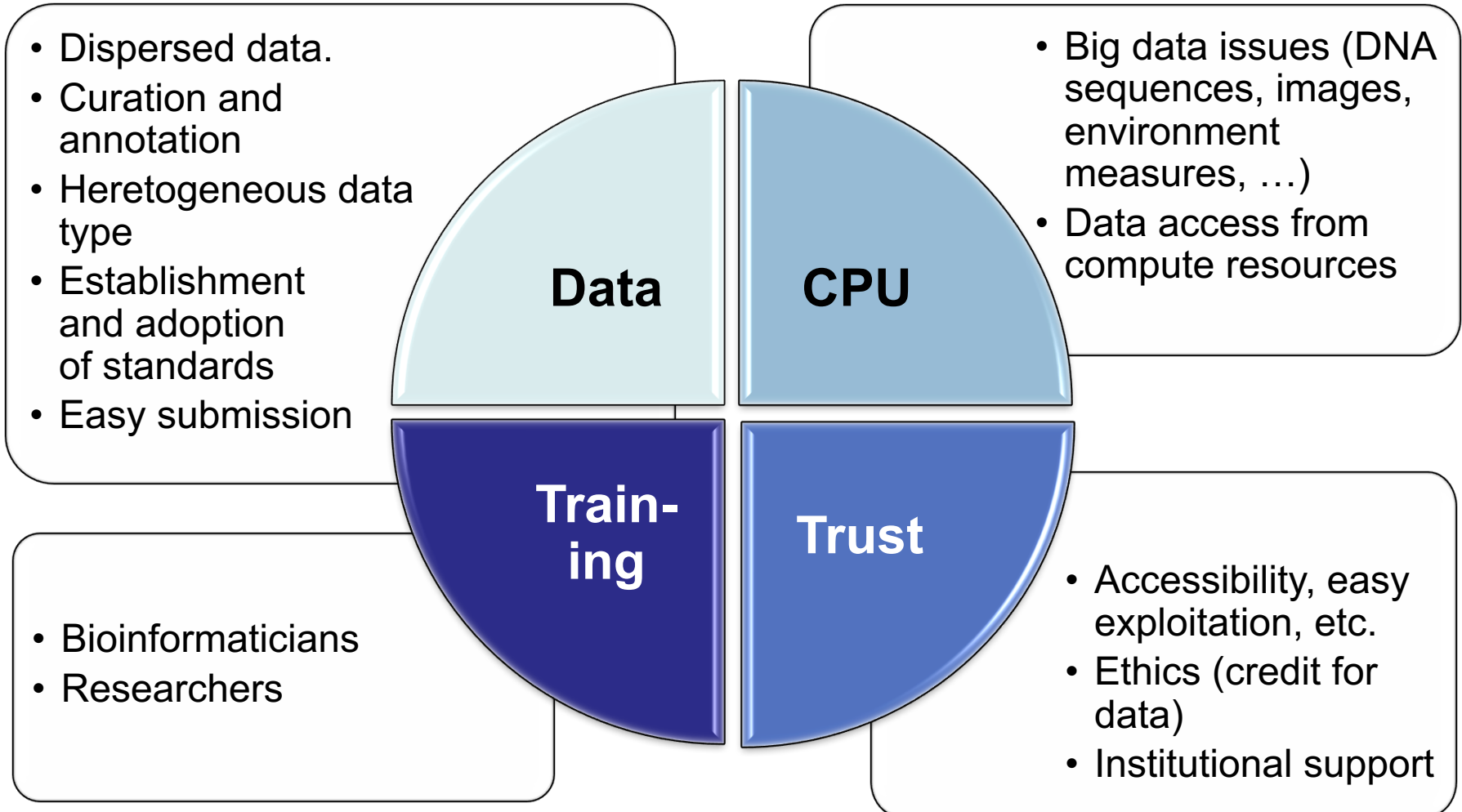
- **Moving complex workflows is challenging**
  - ◆ Hardware dependancies (Cluster)
  - ◆ OS dependancies (Unix/Windows/..., Scheduler)
  - ◆ Software dependancies
- **Technical solutions exist**
  - ◆ Virtualisation: moving an image as a file
  - ◆ Installation package: rpm, conda, ansible, puppet,...

**→ Cloud infrastructure**



# Conclusions:

## Cyber-infrastructure challenges





# 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)  
Jérôme Gouzy (LIPM)



Hadi Quesneville