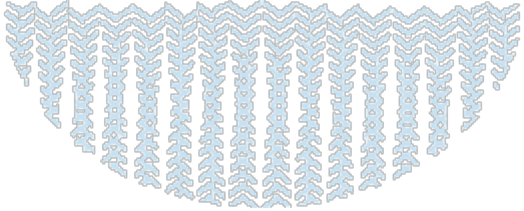
Plant Cyber-infrastructures: wheatIS & French elixir node

Hadi Quesneville

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

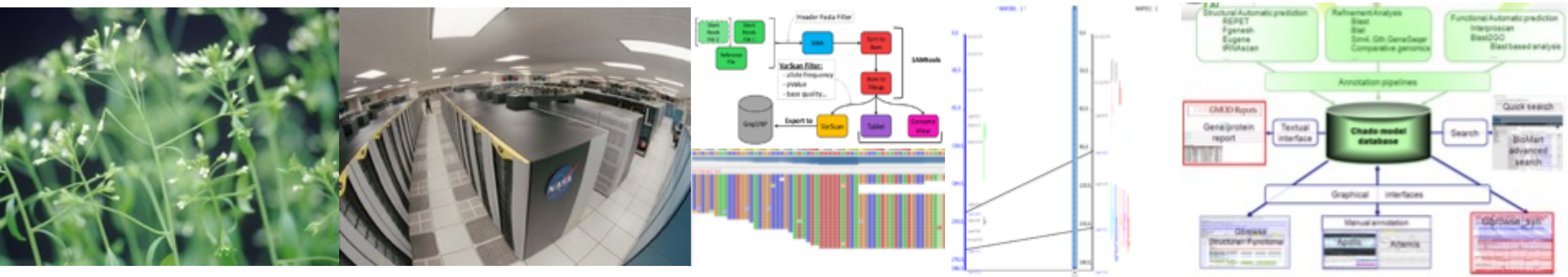
Submitted on 5 Jun 2020



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

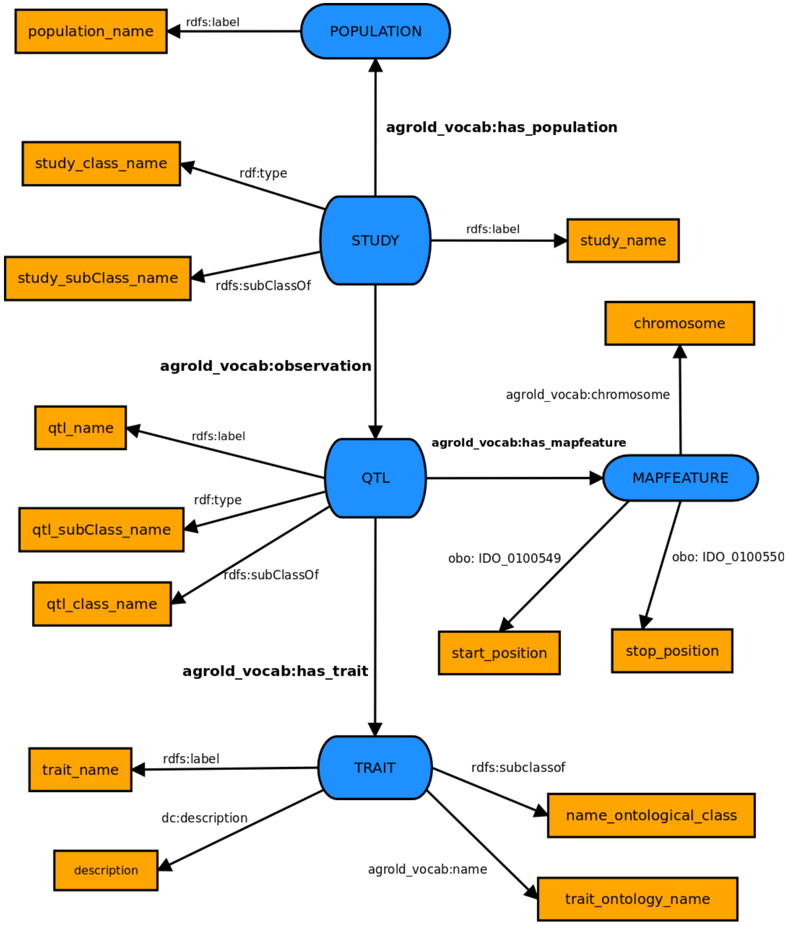
Serve a community

- 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

transPLANT → excelerate

elixir-europe.org

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)

URGI

Genetics and genomics resources for plants and crop parasites

(INRA)

South Green bioinformatics platform

Genomic resource for southern and mediterranean plants.

(CIRAD, INRA, IRD)

Contributors (data, tools, and expertise)

Lipm

Resources for plants, symbionts and pathogens

(INRA,CNRS)

ABiMS

Marine biology analysis

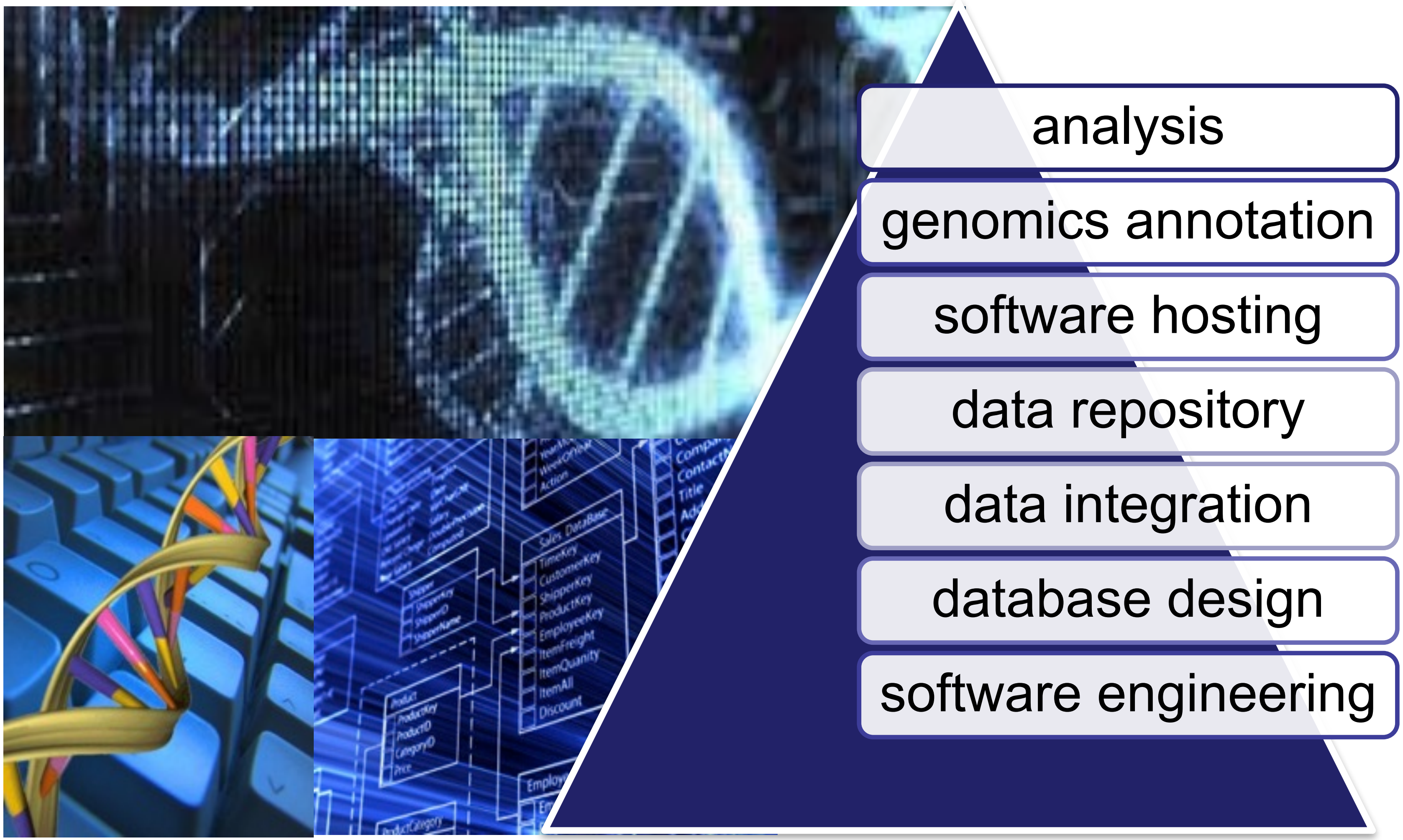
(CNRS, UPMC)

BIPAA

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

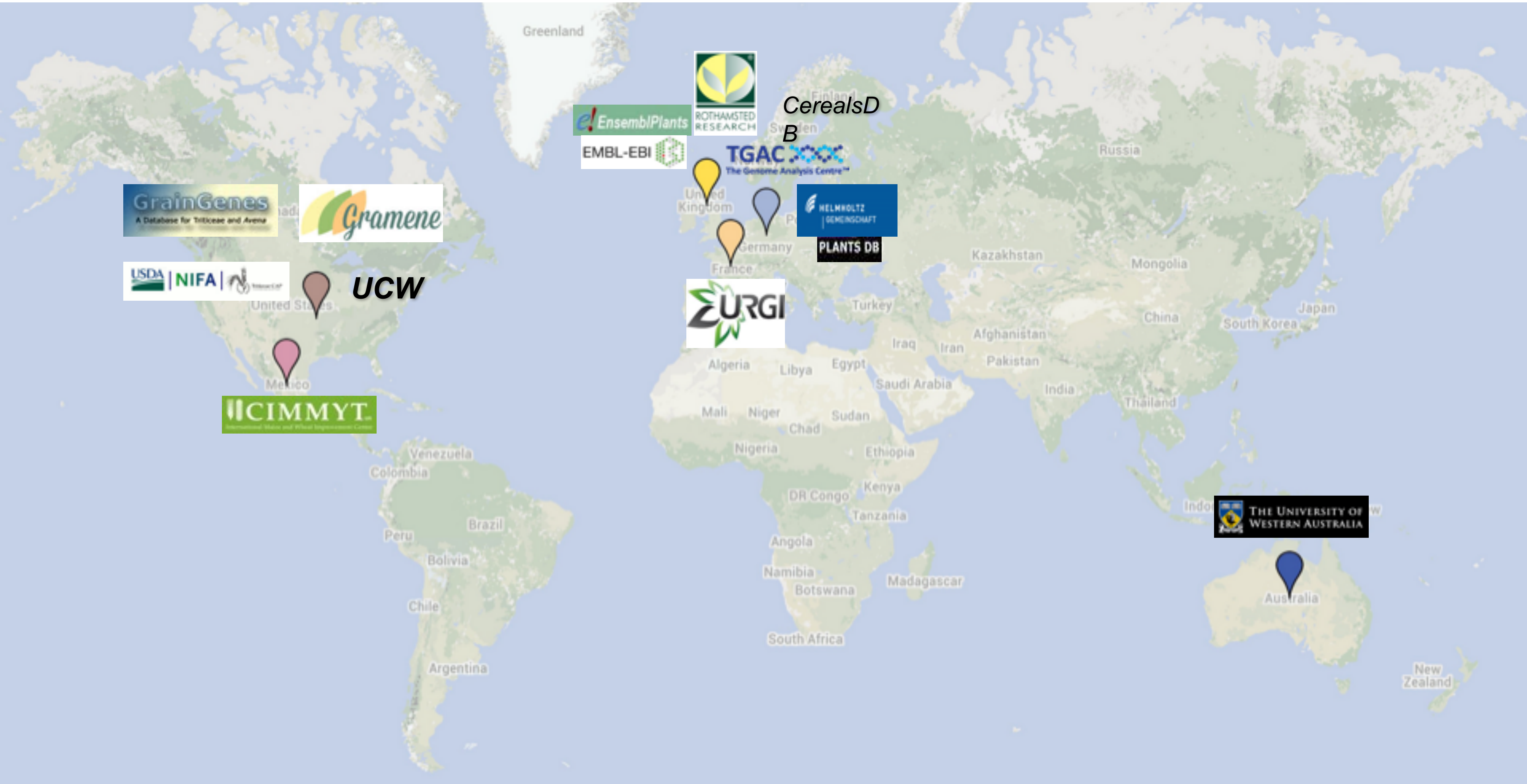
WheatIS Expert Working Group

Wheat initiative commissioned a WheatIS EWG

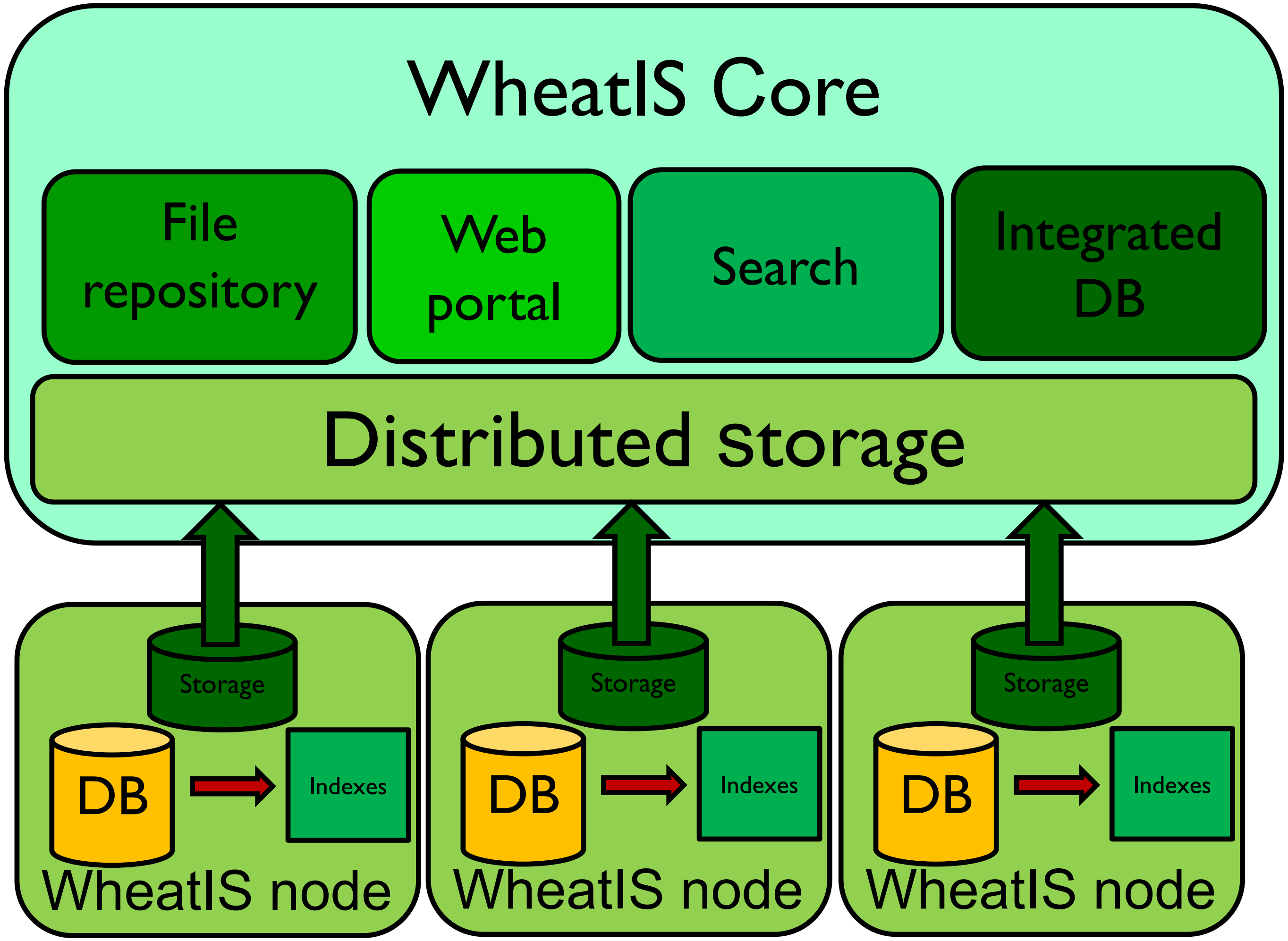
Expert Working Group

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

WheatIS nodes (#12)

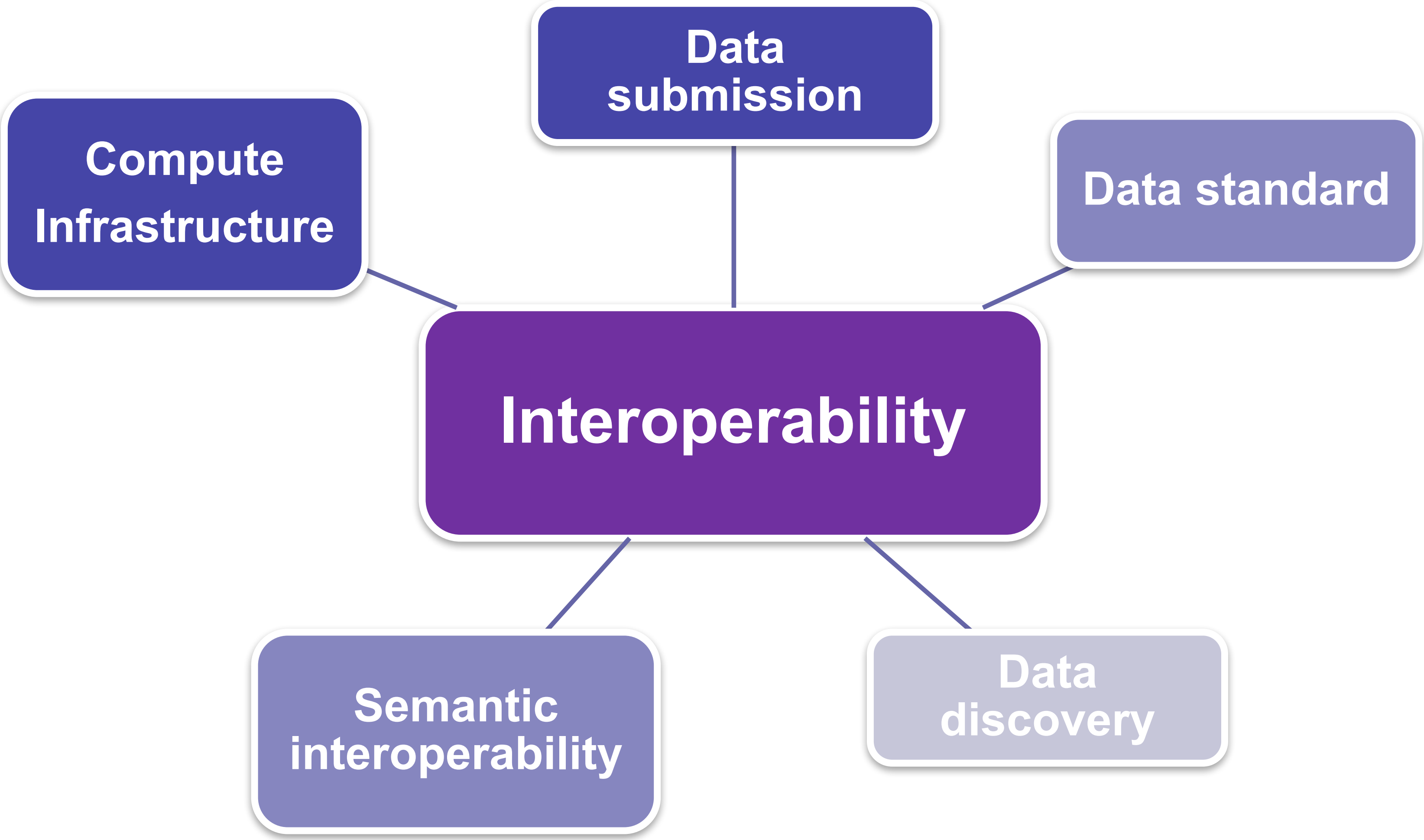


WheatIS Architecture



Hadi Quesneville

Cyber-infrastructures key issue



Interoperability

Data submission

Compute Infrastructure

Data standard

Semantic interoperability

Data discovery

Hadi Quesneville

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.

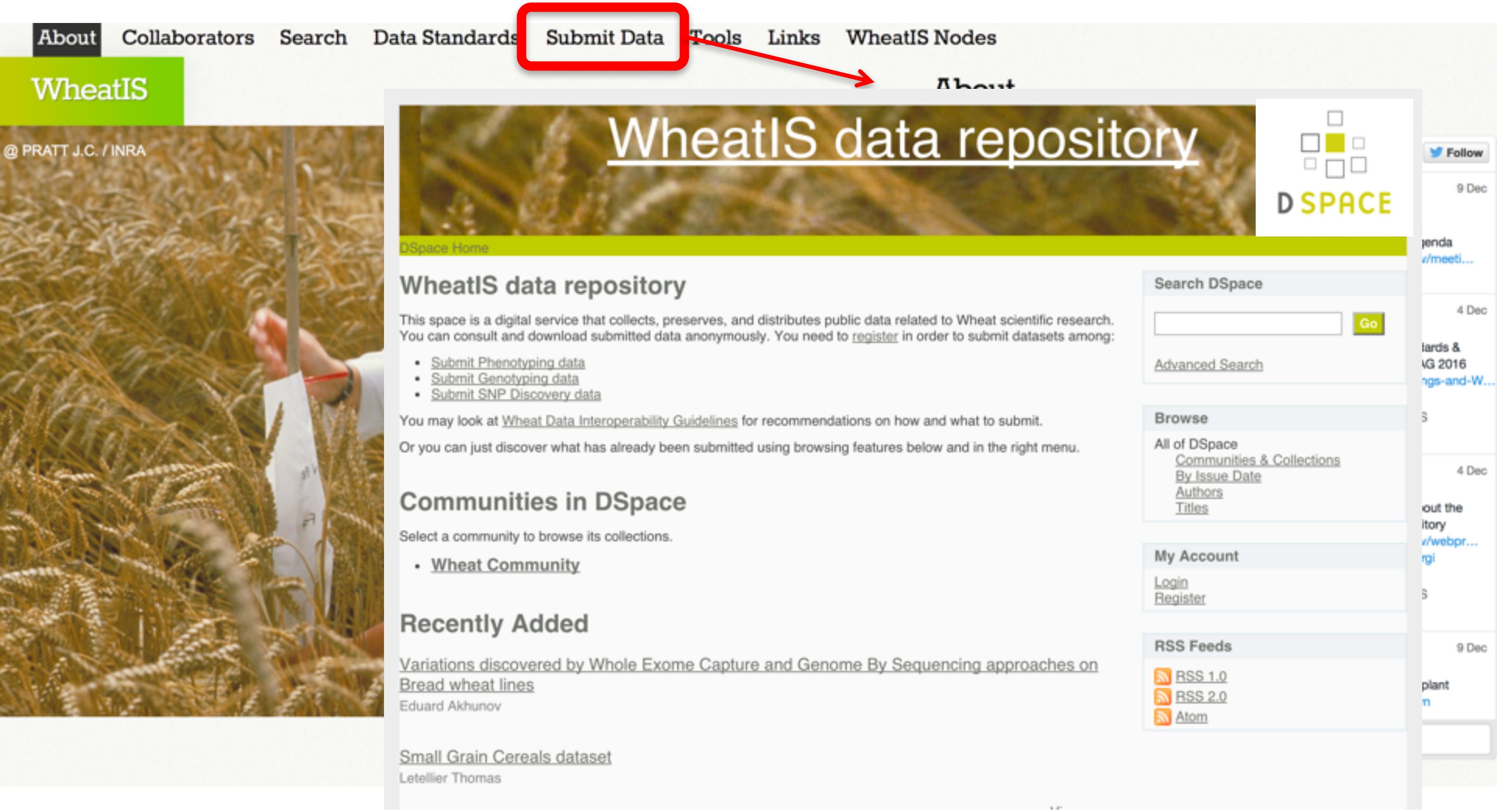
To be Re-usable:

- 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



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.

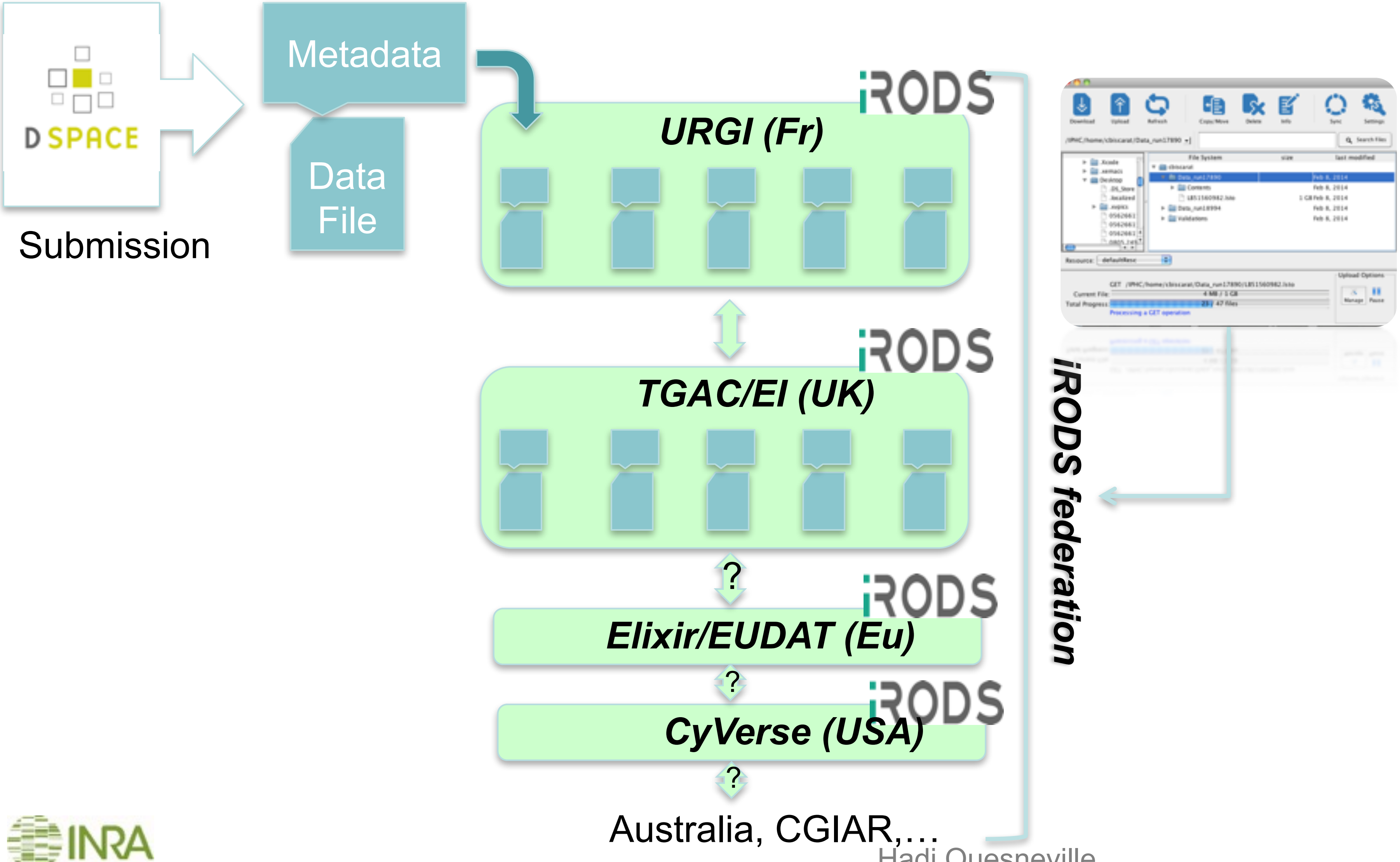
- **Wheat Community**

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

Data workflow



Submission

Metadata

Data File

URGI (Fr)

TGAC/EI (UK)

?

Elixir/EUDAT (Eu)

?

CyVerse (USA)

?

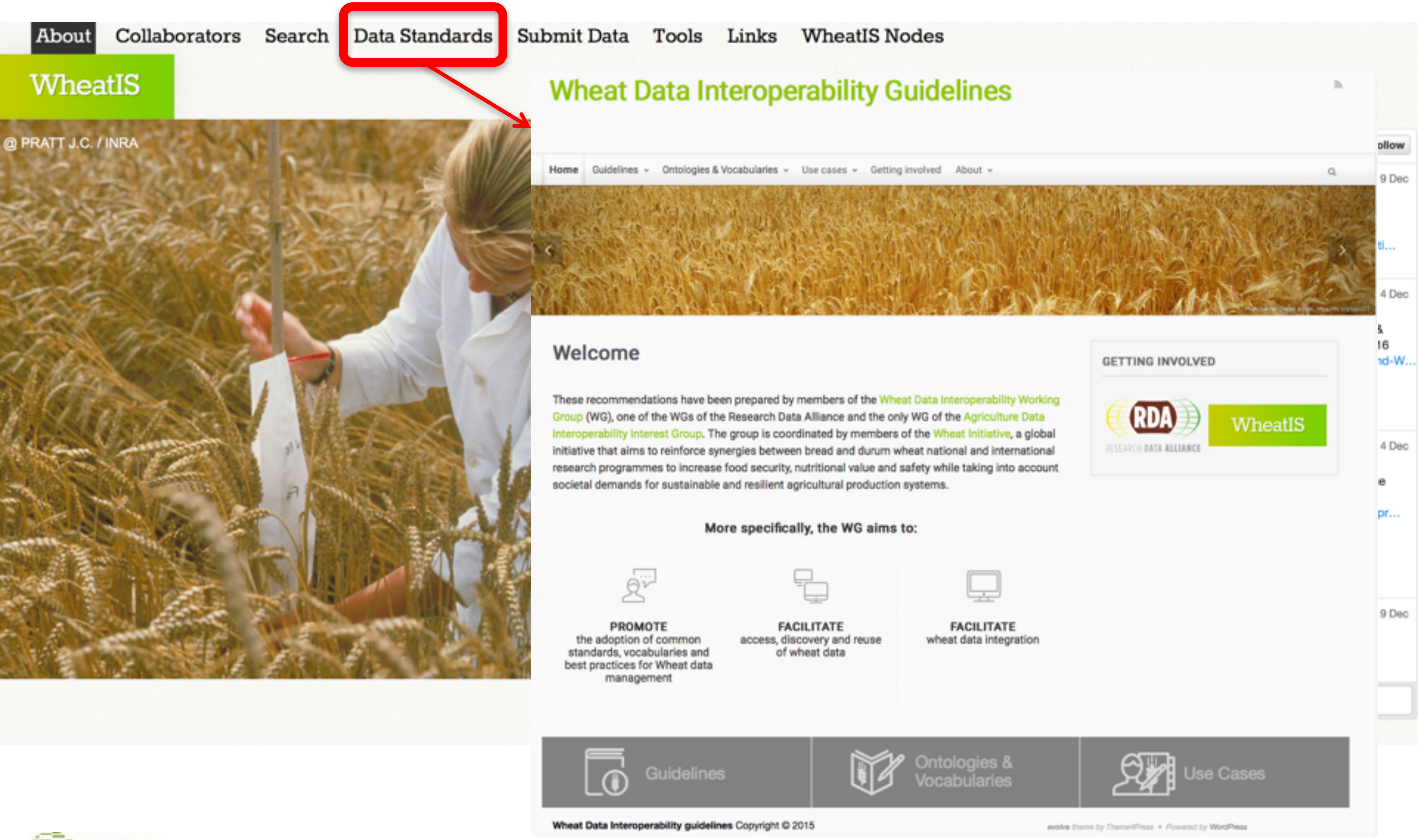
Australia, CGIAR,…

iRODS federation

DATA STANDARDS

Hadi Quesneville

wheatis.org



Wheat Data Interoperability Guidelines

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.

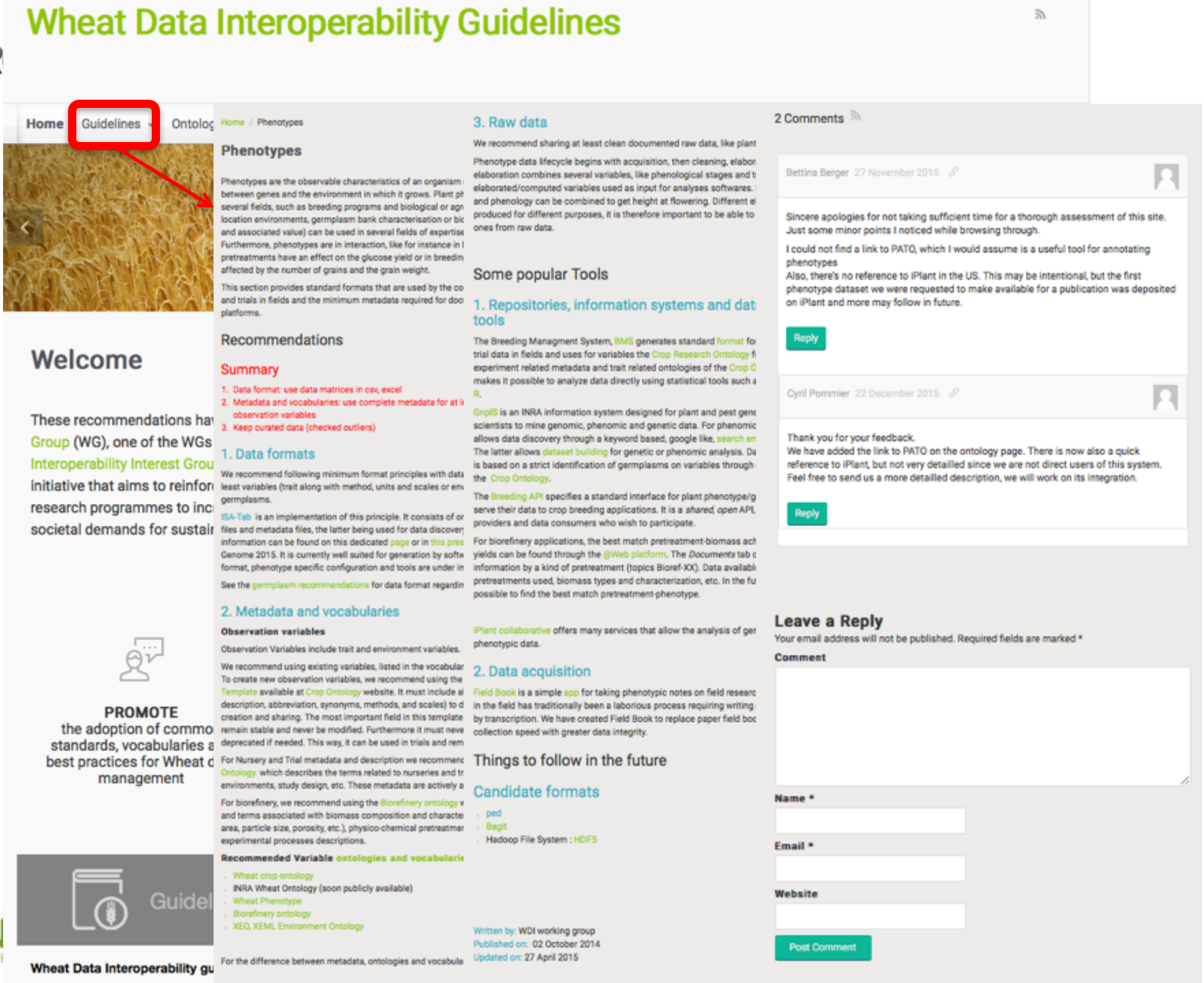
More specifically, the WG aims to:

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

Wheat Data Interoperability guidelines Copyright © 2015



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 d
management

Guidel

Wheat Data Interoperability gu

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
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 discover
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 w
and terms associated with biomass composition and characte
area, particle size, porosity, etc.), physico-chemical pretreatmen
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 el
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. Da
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 ach
yields can be found through the @Web platform. The *Documents* tab o
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.

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.

Leave a Reply

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

Comment

Name *

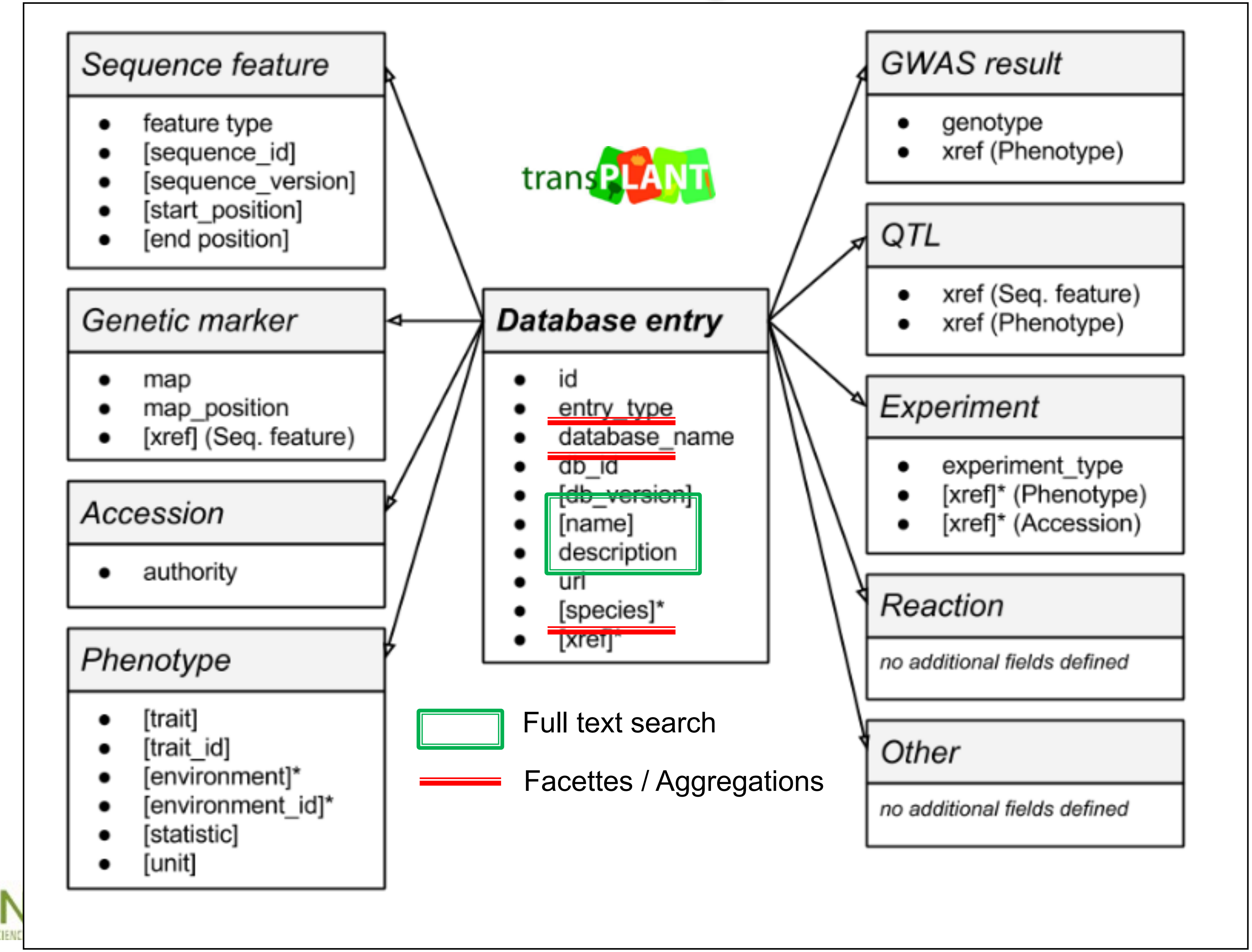
Email *

Website

DATA DISCOVERY

Hadi Quesneville

Data Discovery data model



Sequence feature

- feature type
- [sequence_id]
- [sequence_version]
- [start_position]
- [end position]

Genetic marker

- map
- map_position
- [xref] (Seq. feature)

Accession

- authority

Phenotype

- [trait]
- [trait_id]
- [environment]*
- [environment_id]*
- [statistic]
- [unit]

Database entry

- id
- entry_type
- database_name
- db_id
- [db_version]
- [name]
- description
- url
- [species]*
- [xref]*

GWAS result

- genotype
- xref (Phenotype)

QTL

- xref (Seq. feature)
- xref (Phenotype)

Experiment

- experiment_type
- [xref]* (Phenotype)
- [xref]* (Accession)

Reaction

no additional fields defined

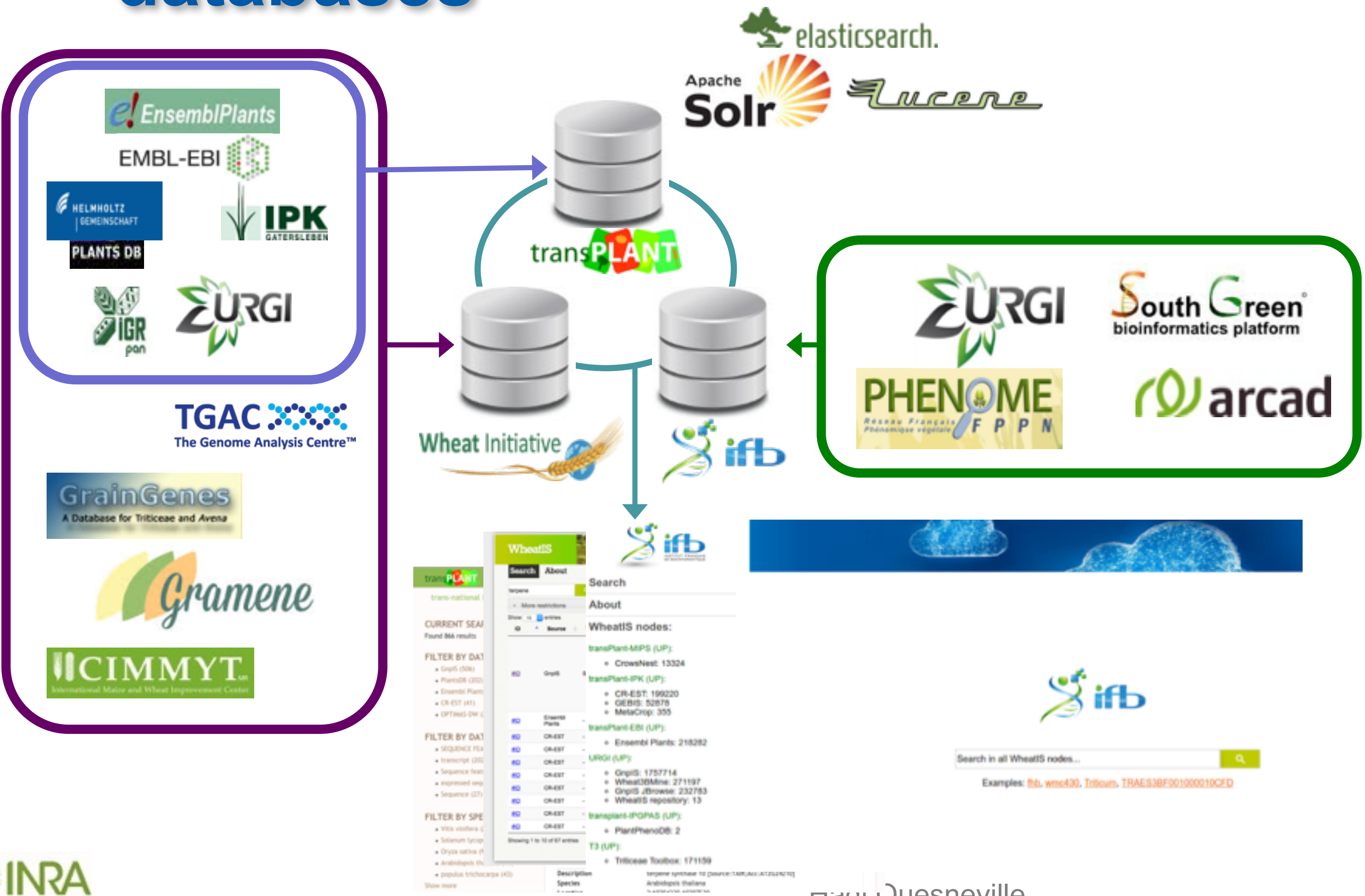
Other

no additional fields defined

🟩 Full text search

🟥 Facettes / Aggregations

Full text search of distributed databases



trans-national

CURRENT SEAR

Found 866 results

FILTER BY DAT
- GnpIS (506)
- PlantsDB (202)
- Ensembl Plants
- CR-EST (41)
- OPTIMAS-DW

FILTER BY DAT
- SEQUENCE FEA
- transcript (202
- Sequence feat
- expressed seq
- Sequence (27)

FILTER BY SPE
- Vitis vinifera
- Solanum lycop
- Oryza sativa
- Arabidopsis th
- populus trichocarpa (40)

Show more

Showing 1 to 10 of 67 entries

Description terpene synthase 10 [Source:TAIR;Acc:AT2G24210]

Species Arabidopsis thaliana

Search

About

WheatIS nodes:

transPlant-MIPS (UP):
- CrowsNest: 13324

transPlant-IPK (UP):
- CR-EST: 199220
- GEBIS: 52878
- MetaCrop: 355

transPlant-EBI (UP):
- Ensembl Plants: 218282

URGI (UP):
- GnpIS: 1757714
- WheatDBMine: 271197
- GnpIS JBrowse: 232783
- WheatIS repository: 13

transplant-IPGPAS (UP):
- PlantPhenoDB: 2

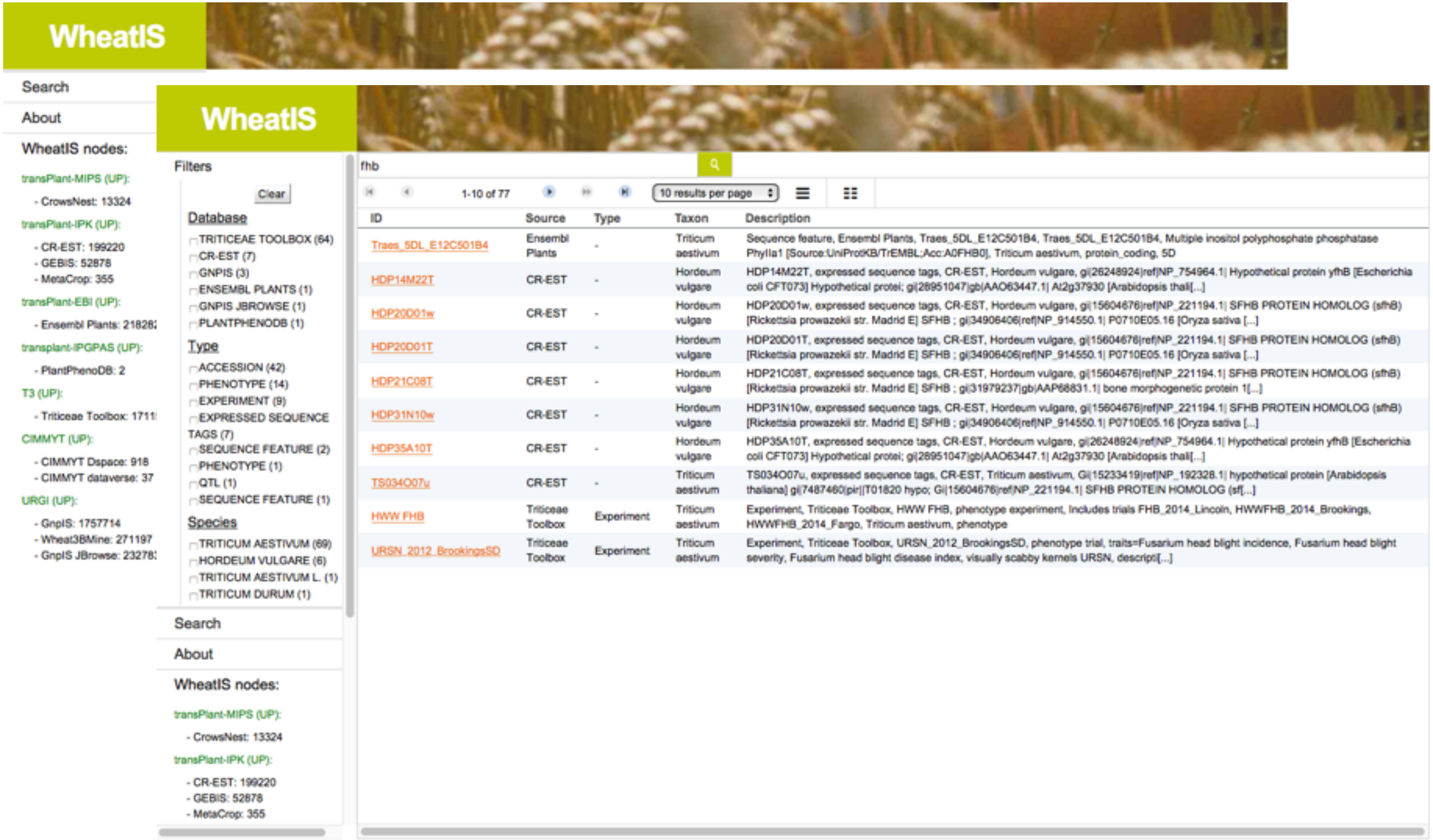
T3 (UP):
- Triticeae Toolbox: 171159

Search in all WheatIS nodes...

Examples: fhb, wmc430, Triticum, TRAES3BF001000010CFD

Hadi Quesneville

wheatis.org



WheatIS nodes:

transPlant-MIPS (UP):
- CrowsNest: 13324

transPlant-IPK (UP):
- CR-EST: 199220
- GEBIS: 52878
- 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: 1757714
- Wheat3BMine: 271197
- GnpIS JBrowse: 23278

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)

WheatIS nodes:

transPlant-MIPS (UP):
- CrowsNest: 13324

transPlant-IPK (UP):
- CR-EST: 199220
- GEBIS: 52878
- MetaCrop: 355

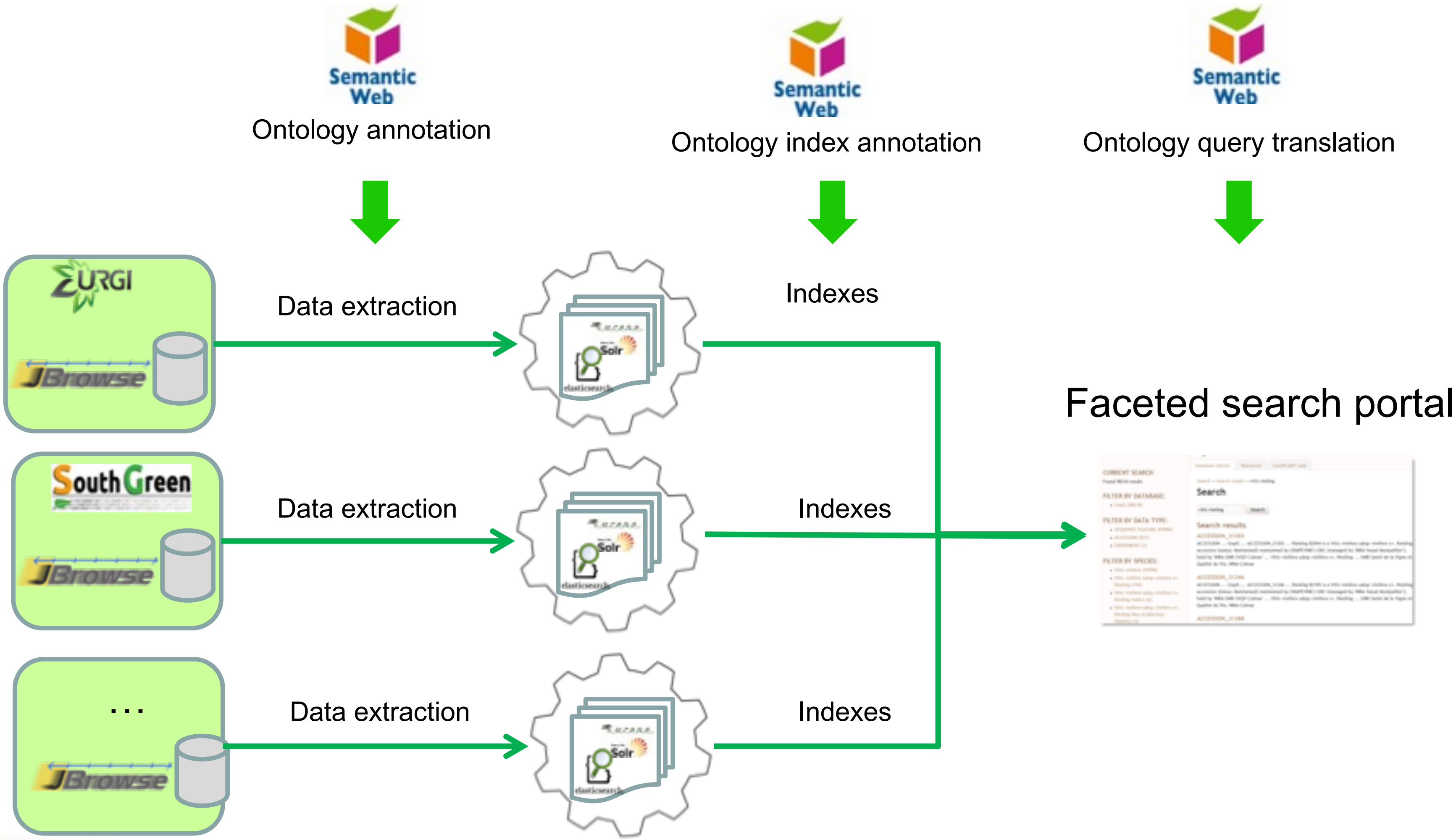
fhb

1-10 of 77

10 results per page

ID	Source	Type	Taxon	Description
Traes_5DL_E12C501B4	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
HDP14M22T	CR-EST	-	Hordeum vulgare	HDP14M22T, expressed sequence tags, CR-EST, Hordeum vulgare, gi|26248924|ref|NP_754964.1| Hypothetical protein yfhB [Escherichia coli CFT073] Hypothetical protei; gi|28951047|gb|AAO63447.1| At2g37930 [Arabidopsis thali[...]
HDP20D01w	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 [...]
HDP20D01T	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 [...]
HDP21C08T	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[...]
HDP31N10w	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 [...]
HDP35A10T	CR-EST	-	Hordeum vulgare	HDP35A10T, expressed sequence tags, CR-EST, Hordeum vulgare, gi|26248924|ref|NP_754964.1| Hypothetical protein yfhB [Escherichia coli CFT073] Hypothetical protei; gi|28951047|gb|AAO63447.1| At2g37930 [Arabidopsis thali[...]
TS034O07u	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 (sf[...]
HWW FHB	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
URSN_2012_BrookingsSD	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, descripti[...]

Next for data discovery



Ontology annotation

Ontology index annotation

Ontology query translation

Data extraction

Indexes

Data extraction

Indexes

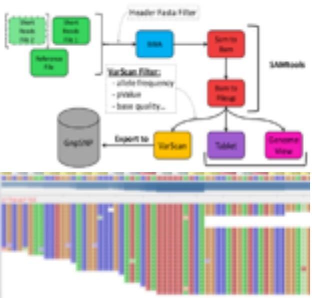
Data extraction

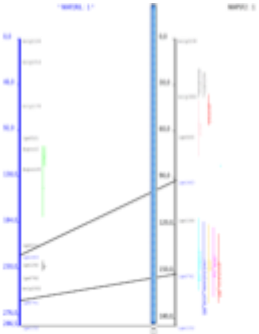
Indexes

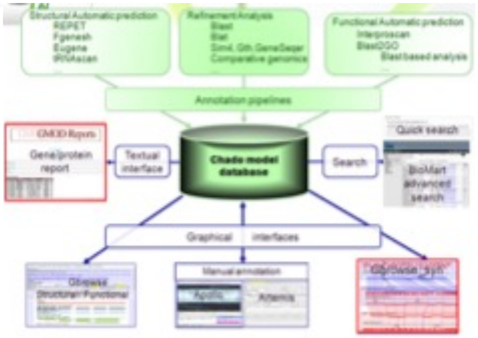
...

Faceted search portal

Semantic Interoperability

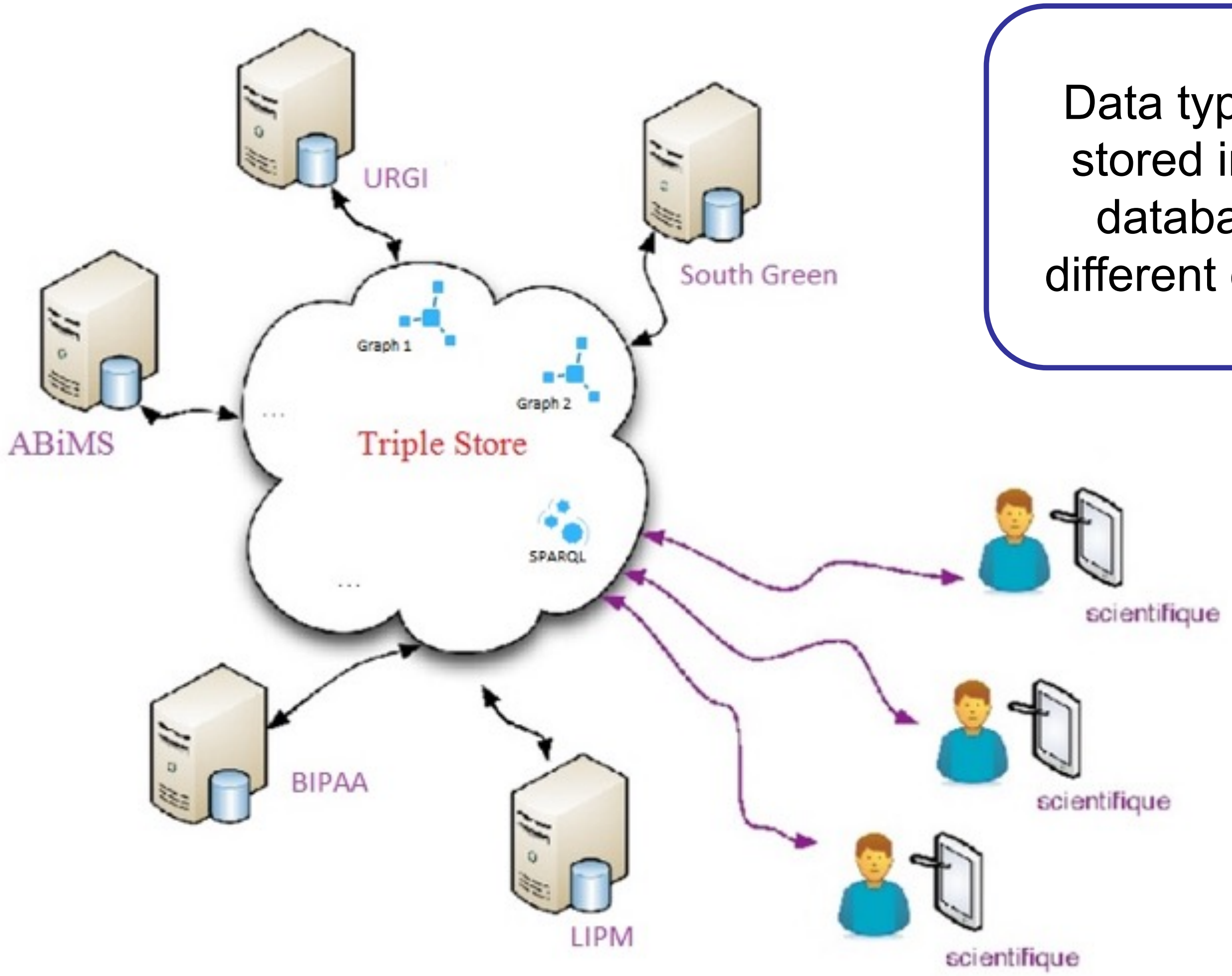






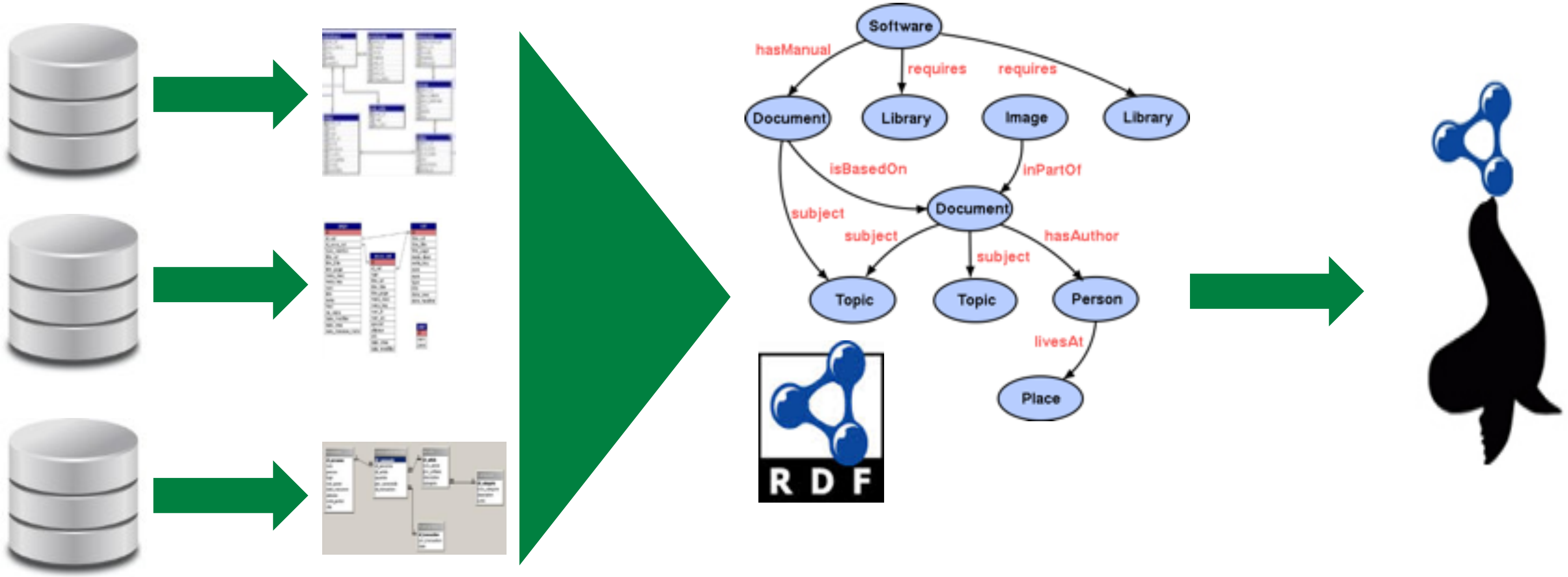
Hadi Quesneville

Why a semantic interoperability



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

Develop a web semantic interoperability



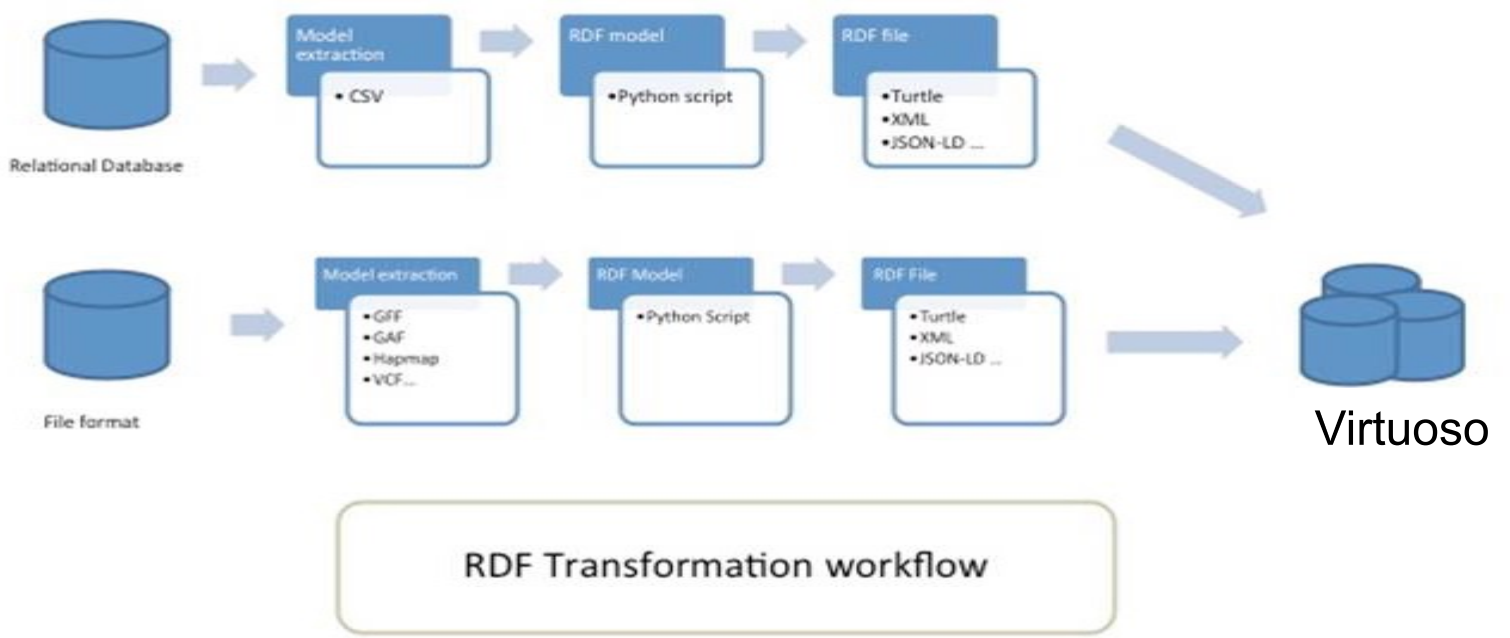
Ontology based annotation of database schemas (GO, PO, TO, CO, …)

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

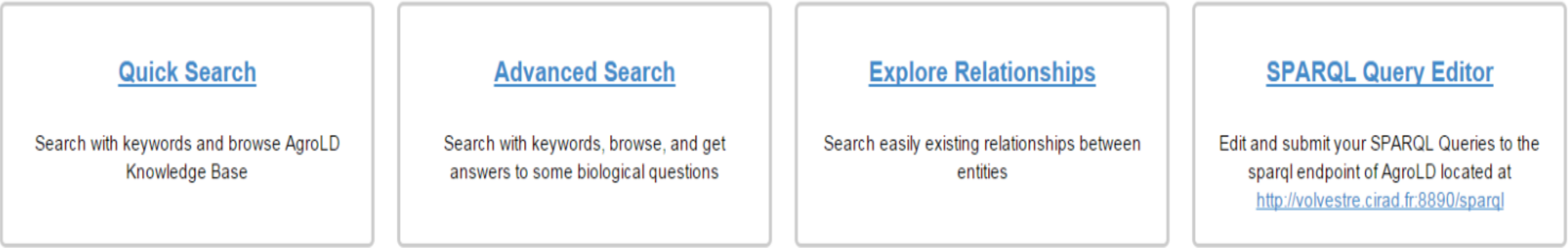


- **Some workflows are generic enough to be reused (i.e. GFF3 or HapMap)**
- **Codes available on GitHub, https://github.com/SouthGreenPlatform/AgroLD.**

AgroLD portal

(www.agrold.org)

- Different user profiles
- Provides different entry points



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

SouthGreen bioinformatics platform

INRA SCIENCE & IMPACT

Hadi Quesneville

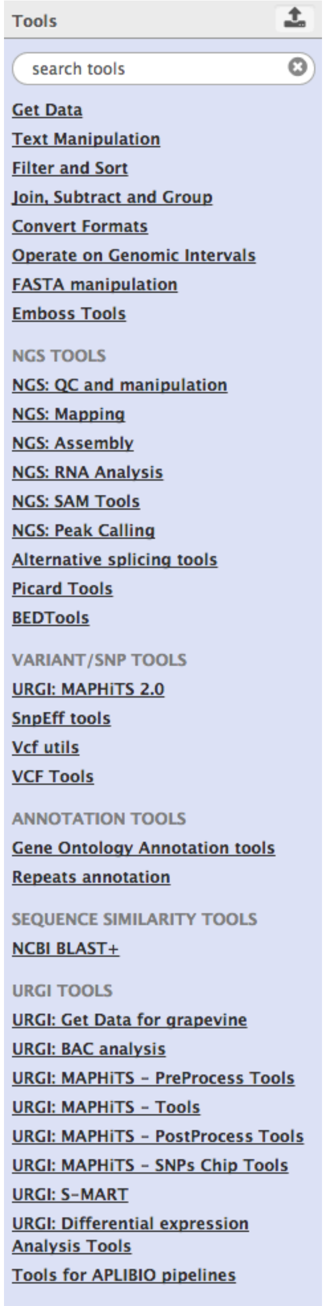
Phenotypic data integration

- **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

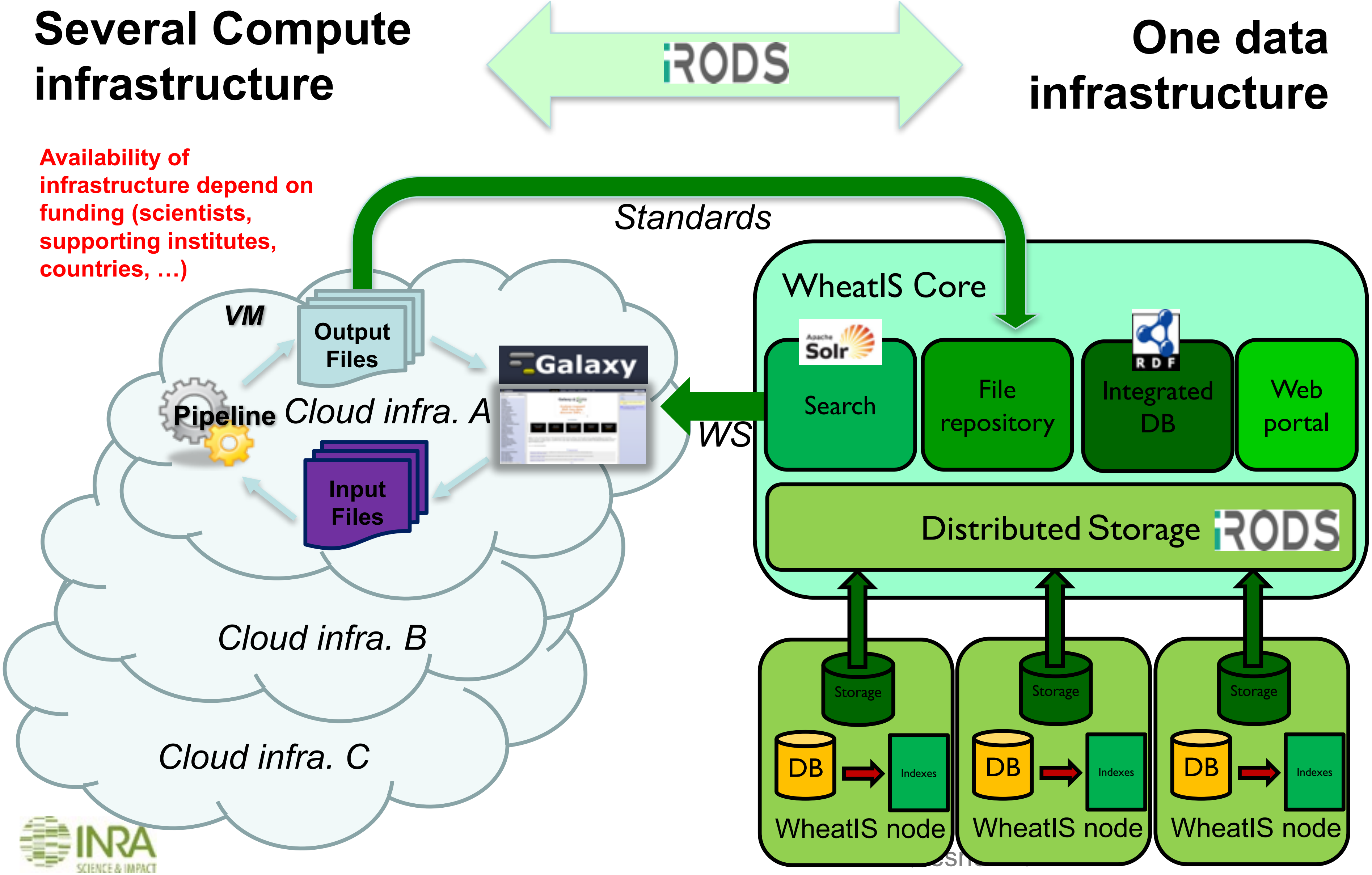
- Differential expression
- ChIP-seq, BS-seq

Genetic diversity analysis

- Population genetic statistics
- genome wide association mapping

Hadi Quesneville

Compute infrastructures



Several Compute infrastructure

iRODS

One data infrastructure

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

Standards

VM

Output Files

Galaxy

Pipeline

Cloud infra. A

Input Files

WS

Cloud infra. B

Cloud infra. C

WheatIS Core

Search

File repository

Integrated DB

Web portal

Distributed Storage iRODS

Storage

DB

Indexes

WheatIS node

Storage

DB

Indexes

WheatIS node

Storage

DB

Indexes

WheatIS node

A wheat use case

NGS data → 1Gb of genomic sequence → 50 Tb

16 Gb wheat genome → would need 800 Tb

Future standard analysis → 18 varieties (Plant J 2017 Feb 23.) → 14400 Tb = 14 Pb ?

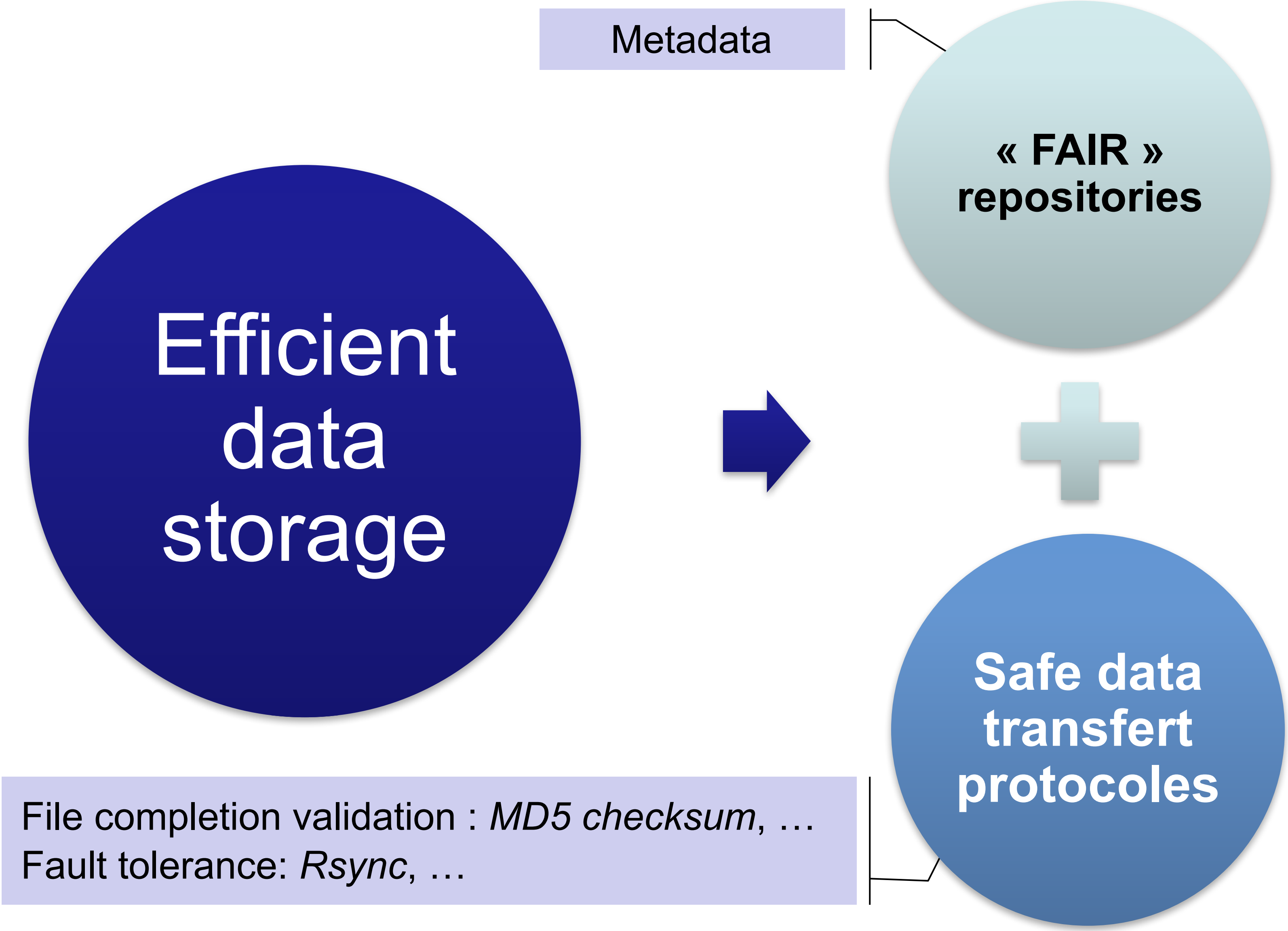
Hadi Quesneville

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

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

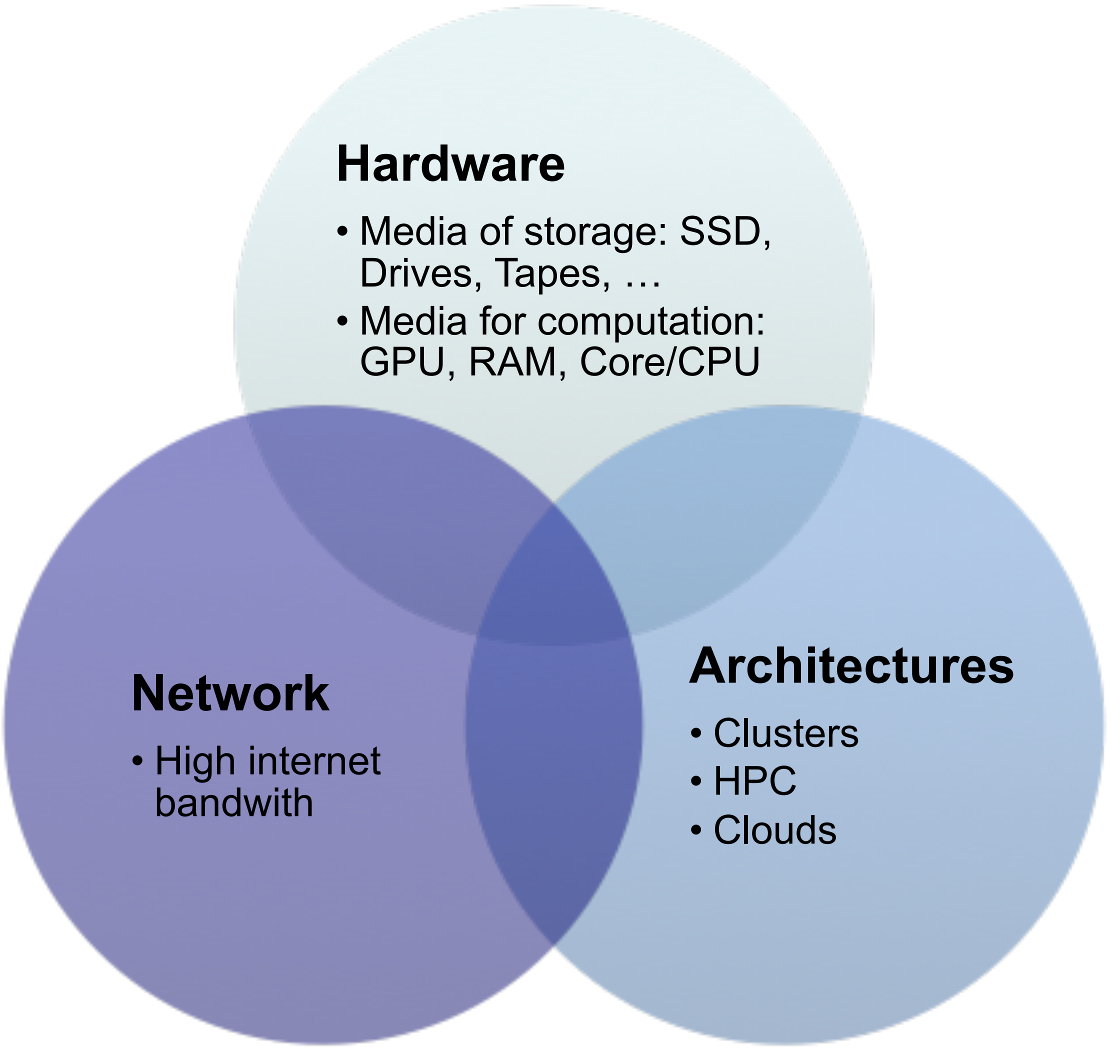
Data dependancies

- Parallelization
- Compute time estimation

Analysis reproducibility

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

Efficient computer resources



Hardware

- Media of storage: SSD, Drives, Tapes, …
- Media for computation: GPU, RAM, Core/CPU

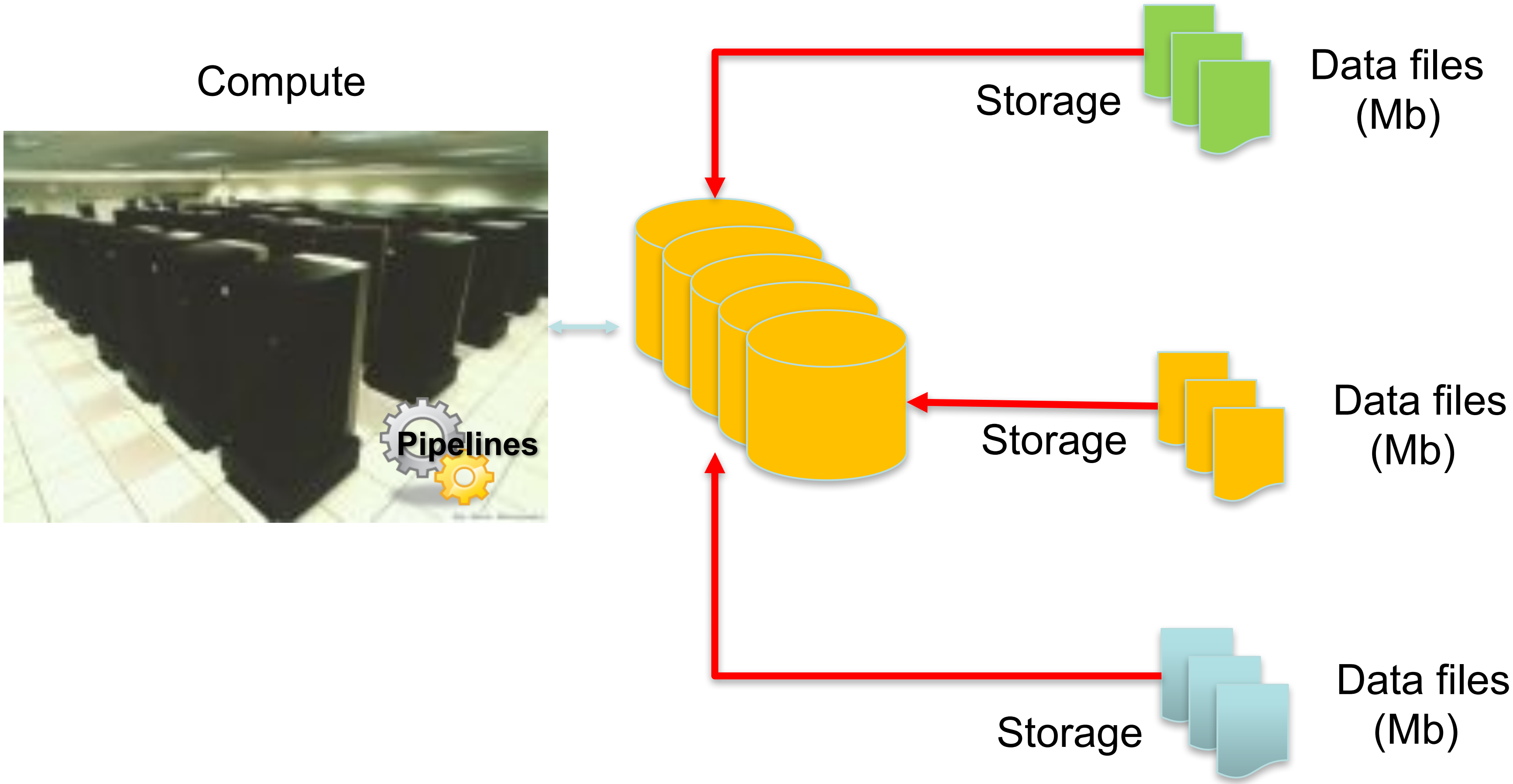
Network

- High internet bandwith

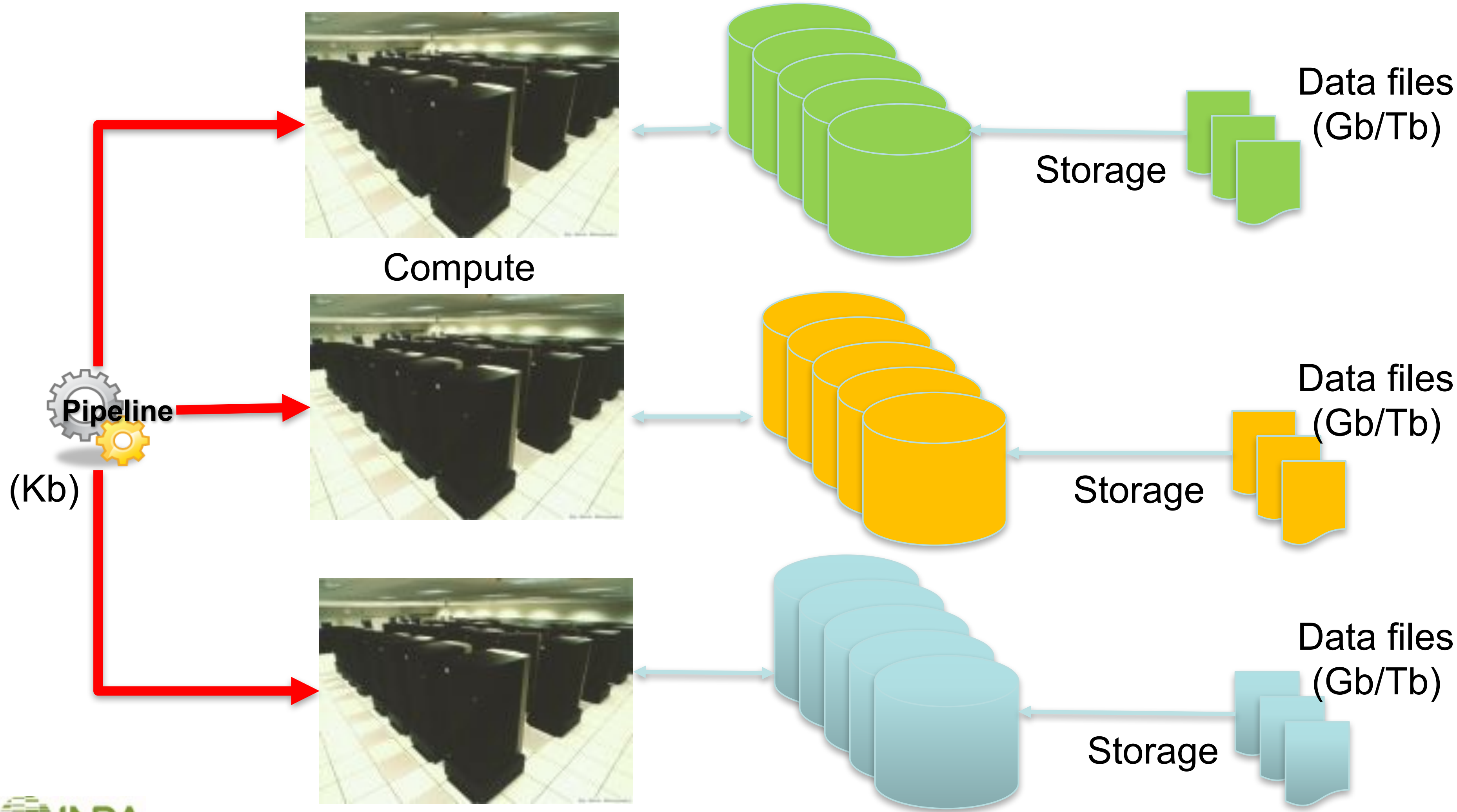
Architectures

- Clusters
- HPC
- Clouds

Today's paradigm: moving data to tools



Paradigm shift: moving tools to data



Pipeline

(Kb)

Compute

Storage

Data files
(Gb/Tb)

Storage

Data files
(Gb/Tb)

Storage

Data files
(Gb/Tb)

Hadi Quesneville

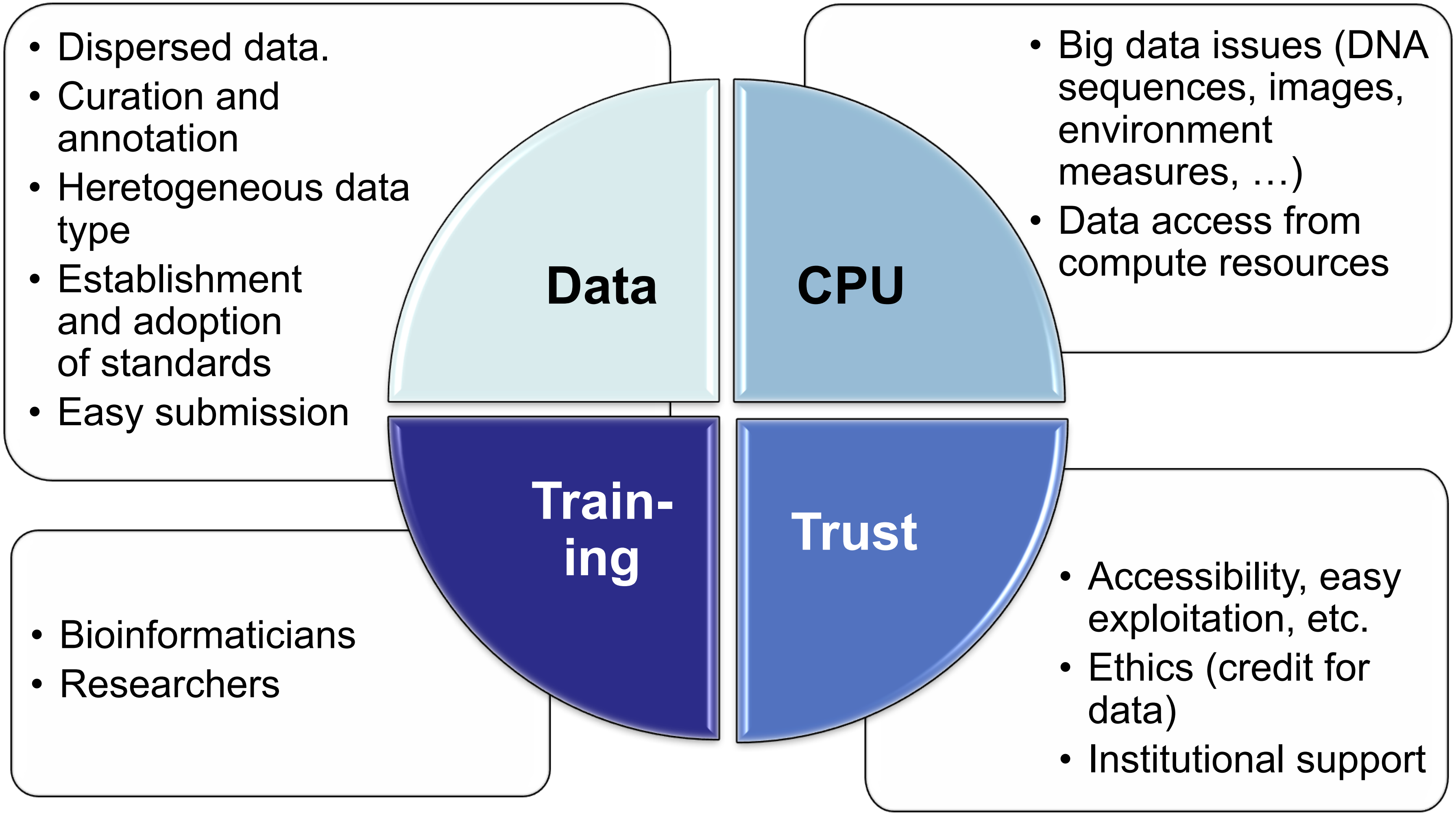
Moving workflows

- **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



Data

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

CPU

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

Train-ing

- Bioinformaticians
- Researchers

Trust

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

Acknowledments

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)