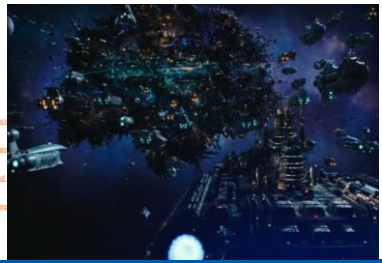
Developing infrastructures for FAIR data in plant biology

Cyril Pommier, Guillaume Cornut, Michael Alaux, Windpouire Esther Dzale Yeumo, Raphaël Flores, Thomas Letellier, Célia Michotey, Nacer Mohellibi, Sophie Durand, Erik Kimmel, et al.

Developing infrastructures for FAIR data in plant biology

A-F Adam-Blondon, URGI, INRA





Context


Genetic resources
Taxons
Genetic analyses
Phenotypes
Polymorphisms

INRA's objectives necessitate interdisciplinary researches and data re-use

Example: modelling the impact of climate change using plant phenology

Global plant phenology data portal: www.plantphenology.org
Pan European PEP725 Plant phenology database: http://www.pep725.eu/

Phenology data

Date associated to:
- Geo-reference
- species
- event (e.g. 50% of opened flowers)

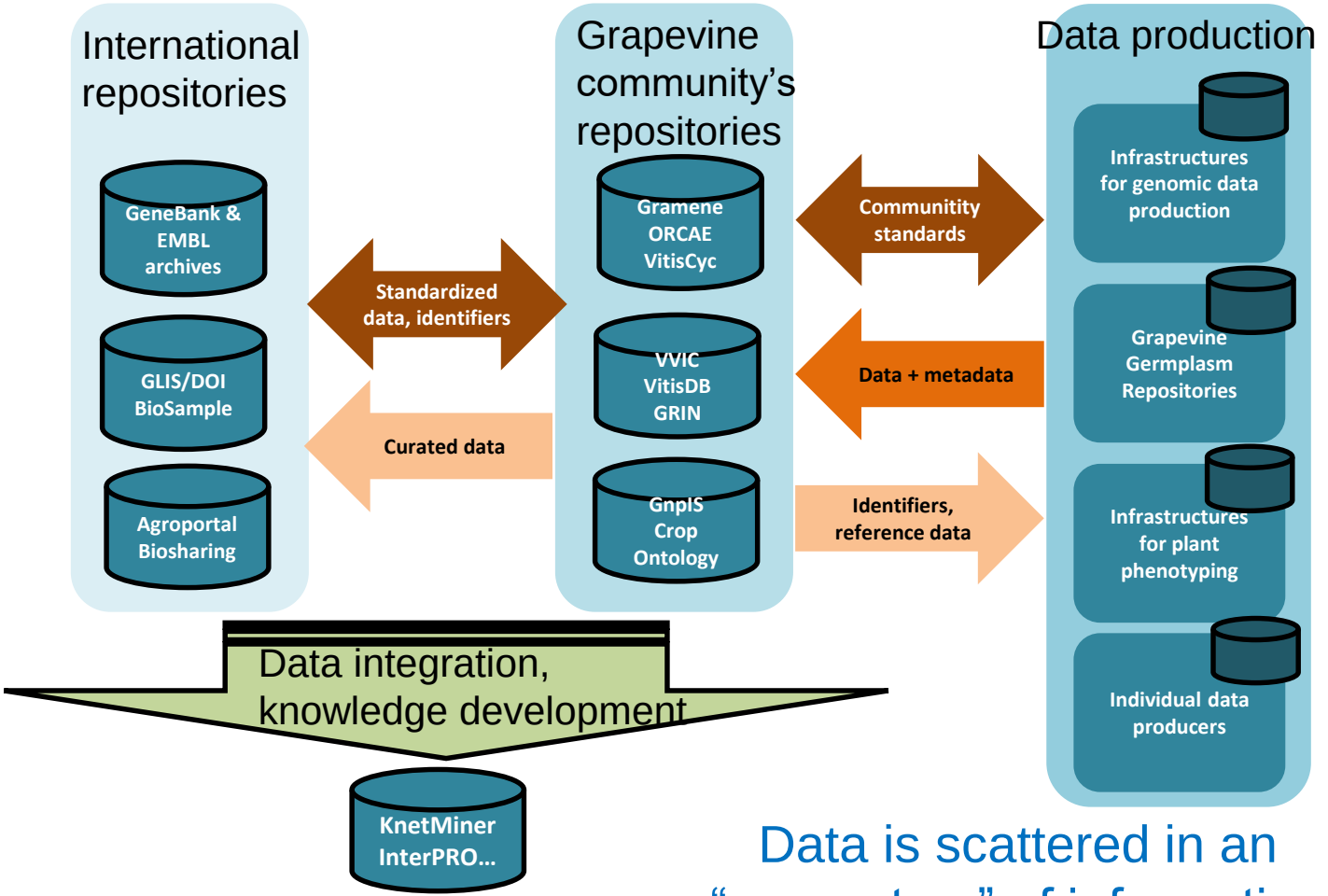
Interoperability

Climatic data

Date associated to:
- Geo-reference
- Measures (e.g T°C, …)
- Events (e.g. hail, …)

Phenology data : different sources, different accuracy in terms of identification of the species, scoring methods, record formats

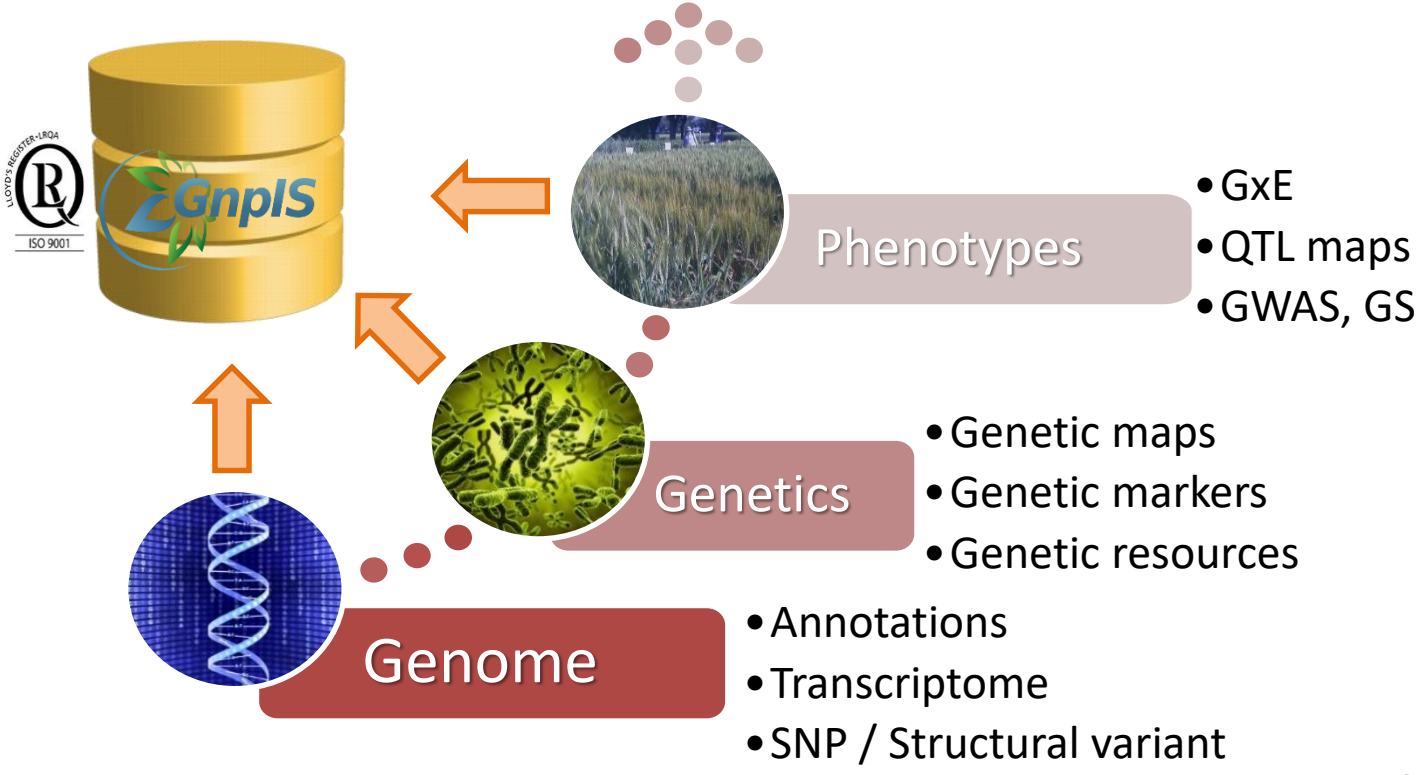
- Modelers of the impact of climate change
- Geneticists, Breeders
- Genbank managers
- Experimental station managers
- Civil (economic) society: e.g. vintage dates, cherry blooming date…



Data is scattered in an "ecosystem" of information systems

From Adam-Blondon et al (2016) HortRes, 3:16056

Development of an INRA information system for crops, forest trees and pathogens



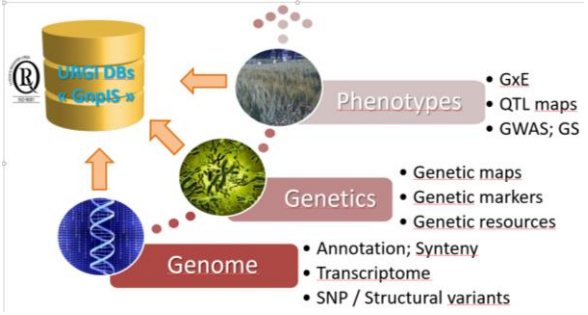
Insertion of GnpIS in federations of information systems

GnpIS-coreDB

2002: Single database

2008: DB Interoperability

2010: set of interoperable databases

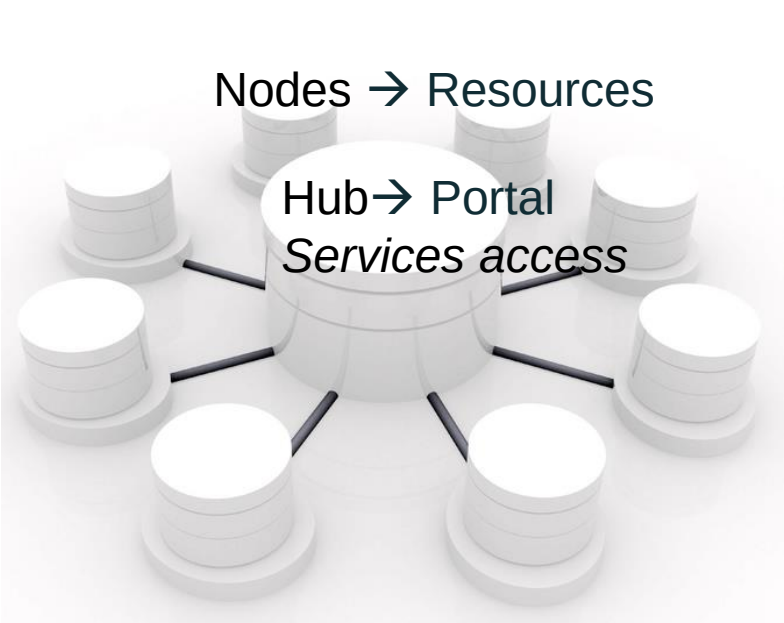


GnpIS Information System

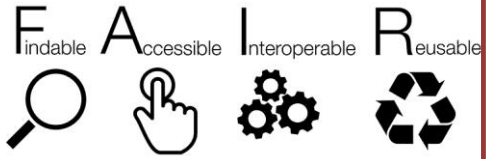
2015: Distributed information system

Featuring federations of information systems

– A network of (stable/sustainable) nodes
– A central portal offering services (e.g. search data)



Backbone of good practices enabling such infrastructures



Wilkinson et al (2016) SCIENTIFIC DATA, 3:160018, DOI: 10.1038/sdata.2016.18

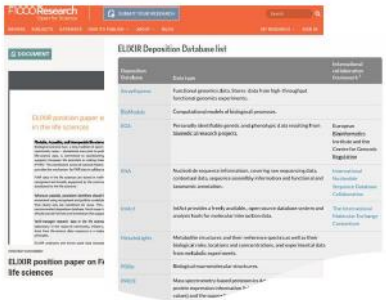
Europe has also been continuously reinforcing its policy for facilitating open access and interoperability

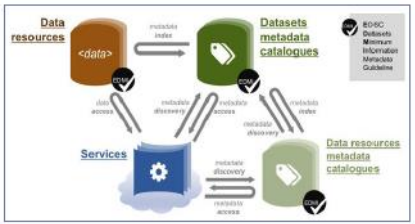
"Facilitating access to results encourages the re-use of research outputs and supports Open Science. This is essential for Europe's ability to enhance its economic performance and improve the capacity to compete through knowledge. […] Results of publicly-funded research can therefore be disseminated more broadly and faster, to the benefit of researchers, innovative industry and citizens.

Recently funded projects were asked to add a WP supporting a FAIR compliant Data Management Plan

This policy is implemented via the European Open Science Cloud (EOSC)

- The European infrastructures in Life Science should contribute to EOSC
- E.g. European Infrastructure of Bioinformatics for Life-sciences: ELIXIR (23 countries)





FAIR data management in the life sciences
10.7490/f1000research.1114985.1

Gateway for User access and mechanism for exposing life-science services (via *ELIXIR Registries*)

Compatible Cloud / Workflows / Reference Data Set Distribution Service

From N. Blomberg

ELIXIR's objectives

In 2023: Continent-scale, standards-based infrastructure for accessing and analysing life-science data

Marine metagenomics

Human data

Plants

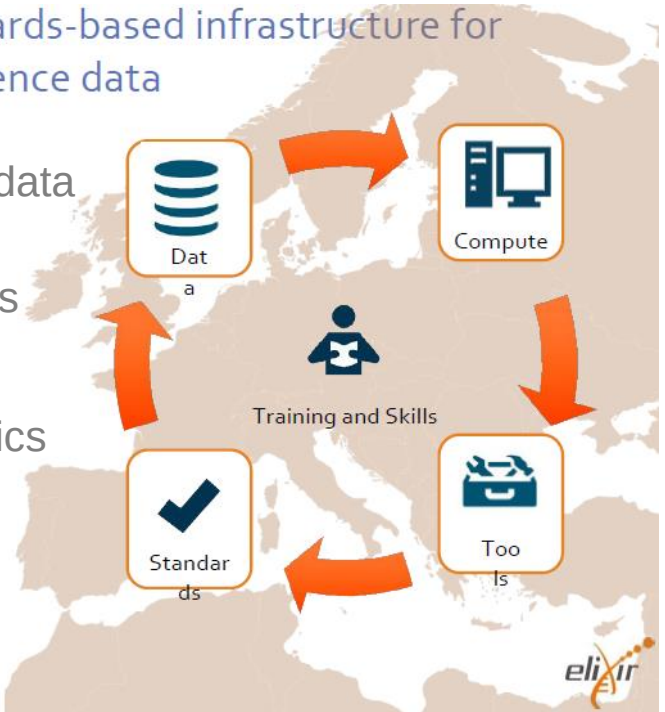
Metabolomics

Proteomics

Galaxy

Rare diseases

...delivered in partnership with research communities



Contacts Plants: Celia Miguel (IBET) & Cyril Pommier (INRA)

From N. Blomberg

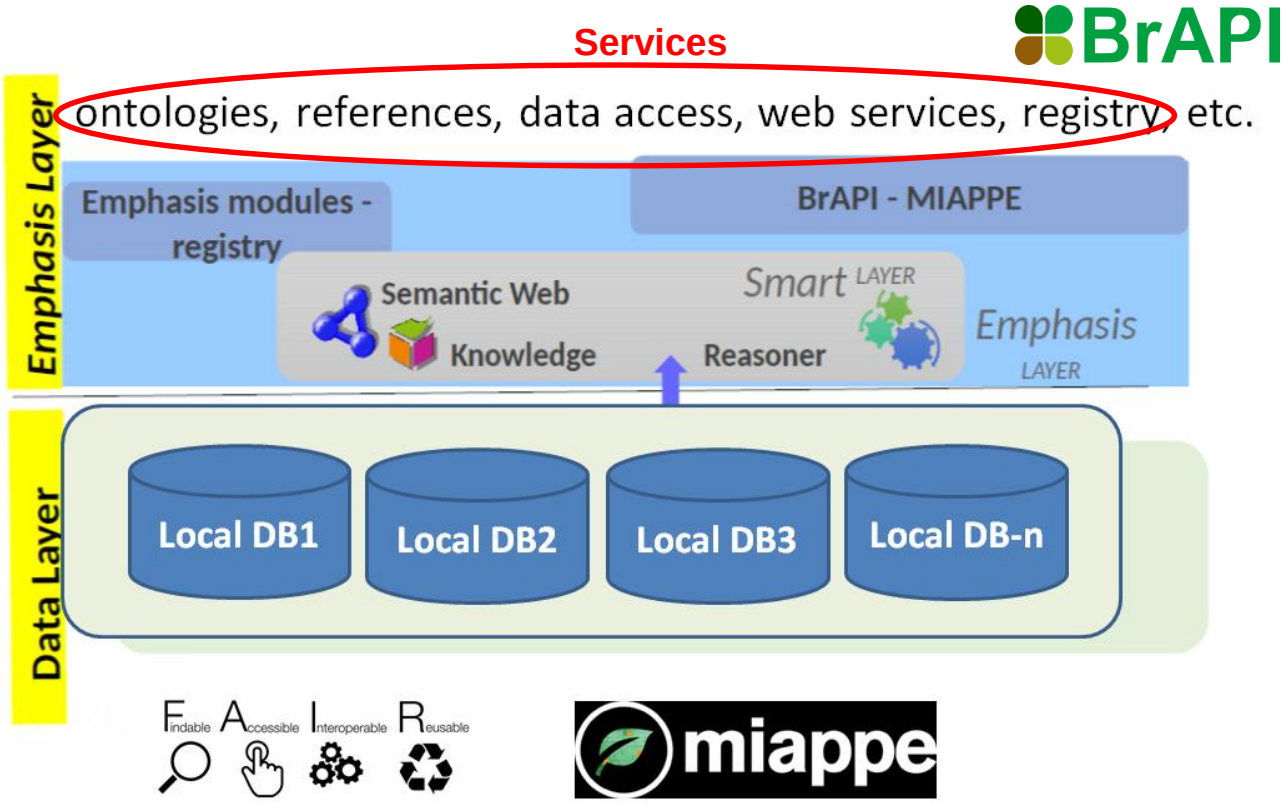
Examples of federations

Two main examples or “use cases”

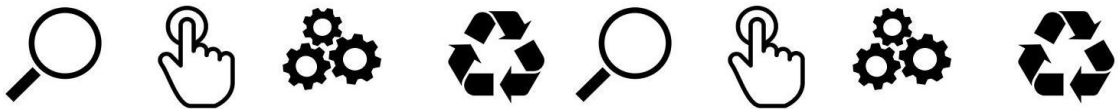
- The Wheat Initiative (G20 Initiative) and its Wheat Information System Expert Working group (www.wheatis.org). Also supported by the Research Data Alliance.

- The European Infrastructure for Multi-scale Plant Phenomics and Simulation (EMPHASIS) and its information system (https://emphasis.plant-phenotyping.eu/e-Infrastructure). In the frame of a strong collaboration with ELIXIR

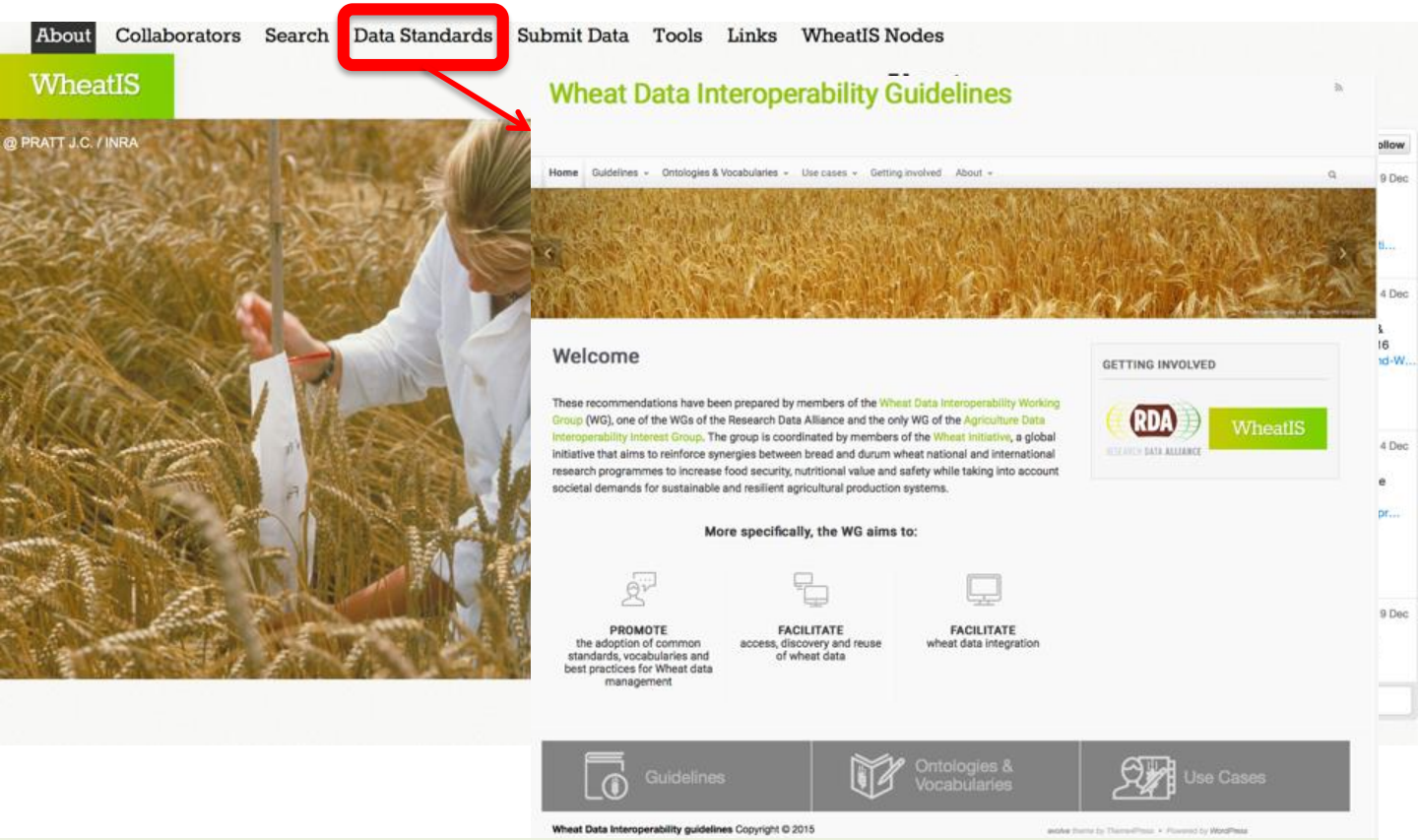
https://emphasis.plant-phenotyping.eu/e-Infrastructure



Developing a federation of FAIR plant phenotyping data repositories



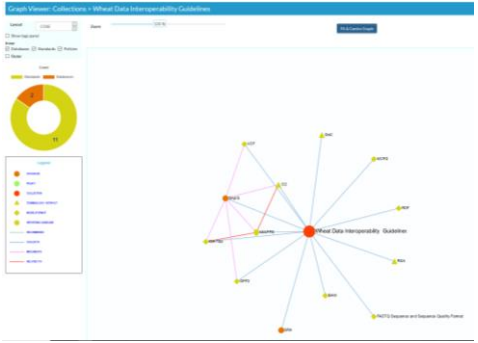
Development of guidelines: e.g. www. wheatis.org



https://ist.blogs.inra.fr/wdi/

Dzale-Yeumo et al (2017) F1000Research, 6 : 1843

Registries of standards and guidelines



elixir UNITED KINGDOM

EMBL Australia Bioinformatics Resource

RDA RESEARCH DATA ALLIANCE

FORCE11

F1000 FACULTY OF 1000

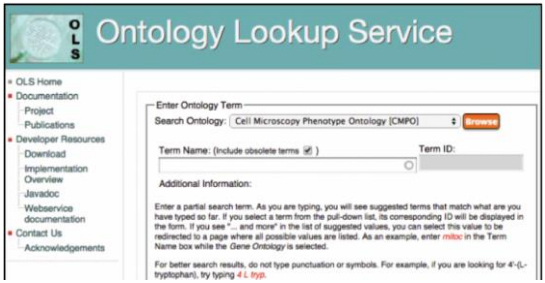
DRYAD

FAIRsharing.org standards, databases, policies

PLOS

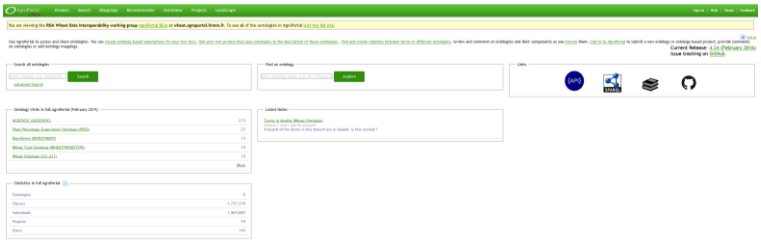
nature publishing group

EMBOpress



EMBL-EBI

https://www.ebi.ac.uk/ols/ontologies



http://agroportal.lirmm.fr/

Supported by ANR

With the collaboration of

NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY · SIFR project · LIRMM · CNRS · Bioversity International · INRA SCIENCE & IMPACT · IRD · cirad

Development of a metadata standard for phenotyping experiments

- MIAPPE: Minimum Information About Phenotyping Experiment
- www.miappe.org
- Steering committee Emphasis, Elixir CGIARs

Last release MIAPPE v1.1 (Jan. 2019). Major improvements:

- Extension to accommodate woody plants as an additional use-case.
- Specification of a data model for easier implementation in various formats and automatic validation.
- Improved compatibility with ISA-Tools and Breeding API (BrAPI).
- Provision of clear definitions and examples for all fields.

Adoption of the Crop Ontology format for the description of the phenotyping variables (www.cropontology.org)

Variable = trait + method + scale

Examples

- Woody Plant Ontology
- Wheat INRA Phenotyping Ontology v1.3 beeing merged with the Wheat Ontology developed by the CYMMIT
- Grape Ontology (v2)

- Also used to describe environment variables
- Strong curation efforts still needed: documentation of the methods, standardize the vocabulary (traits), …
- Environments facilitating the development and curation of ontologies are needed

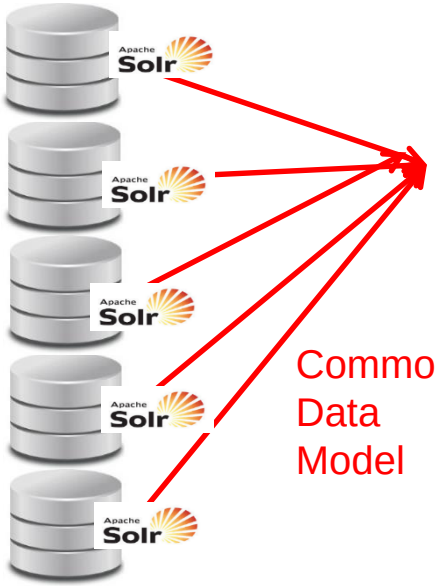
Bioversity International Planteome RDA RESEARCH DATA ALLIANCE WHEAT INITIATIVE excelerate EMPHASIS

Registries of identifiers for key objects

- DOI for plant accessions (following the FAO recommandations)
- Biosample : identifiers for samples derived from accessions
- Crop Ontology : identifiers for phenotyping variables
- DOI associated to phenotyping trial sets (and data papers)

Developing search and data access across federations

Generic Data Discovery Tool

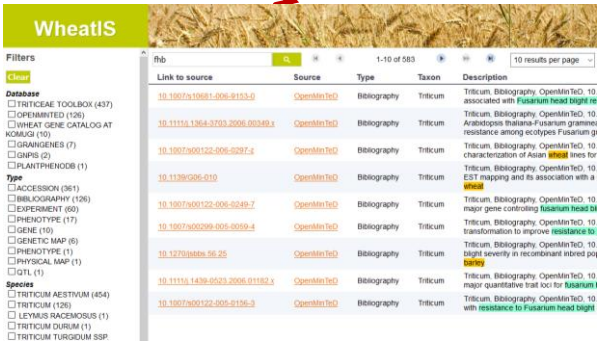


Common Data Model

User web

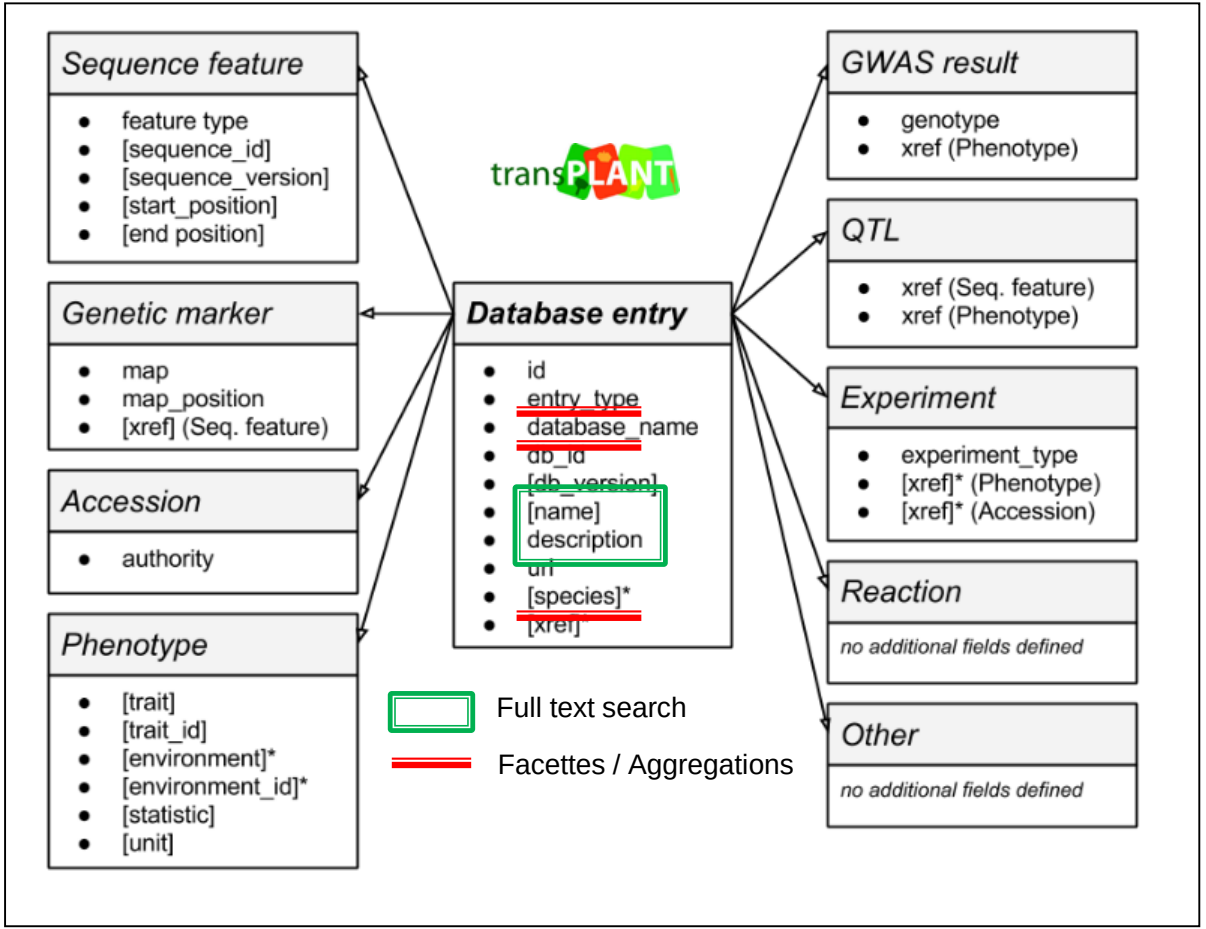
http://www.wheatis.org/

Google like list of results

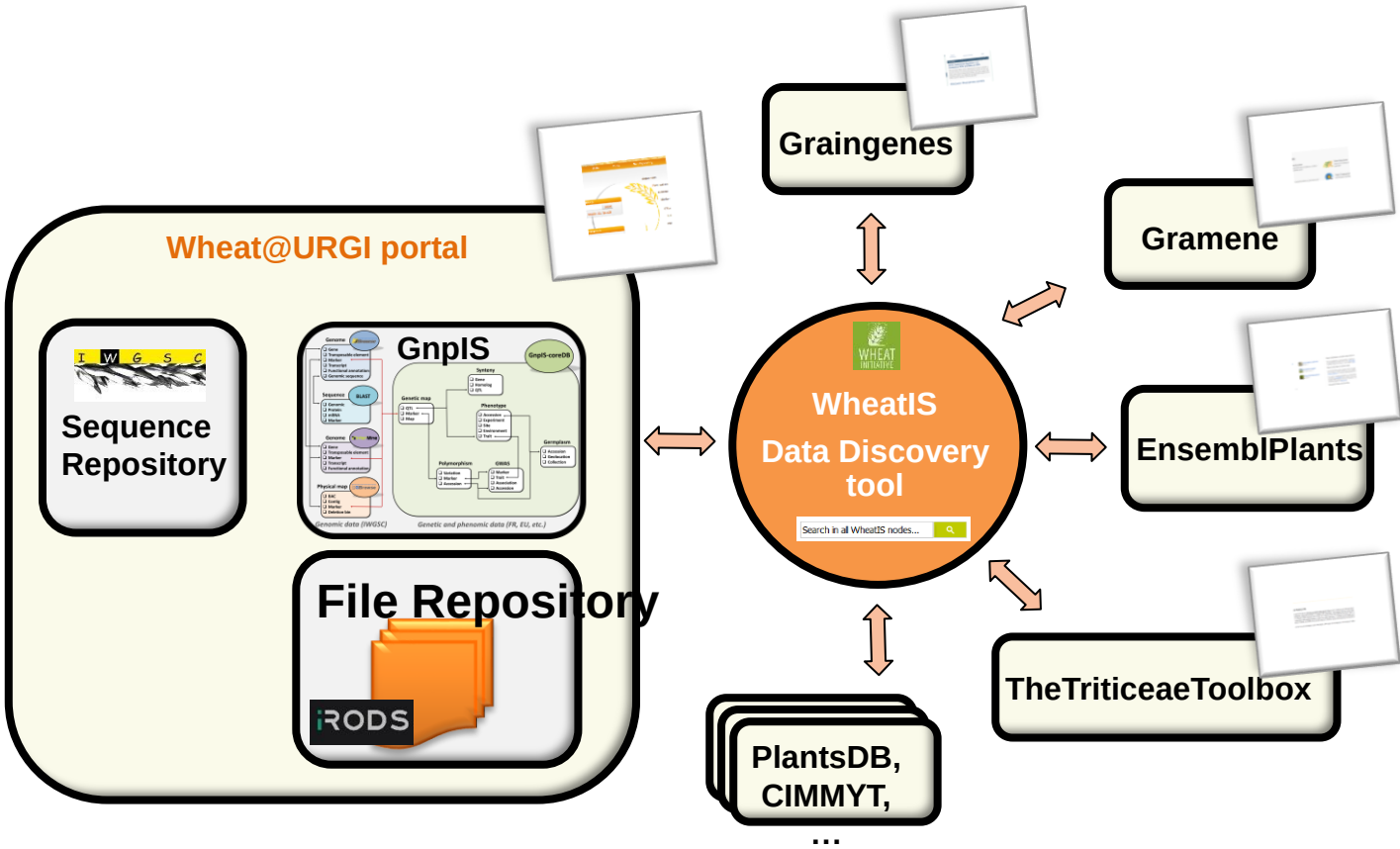


Spannagl et al. 2016, https://doi.org/10.3835/plantgenome2015.06.0038

transPLANT data model

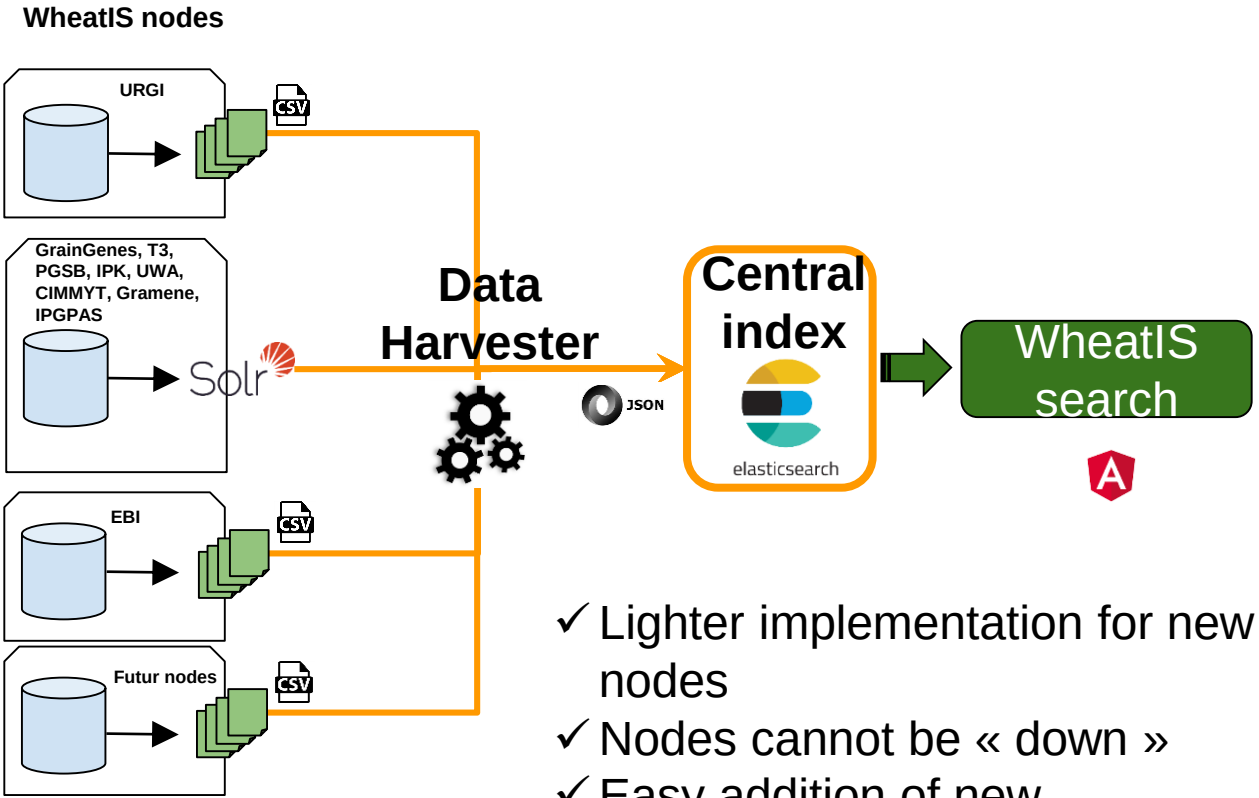


WheatIS data discovery: https://urgi.versailles.inra.fr/wheatis/



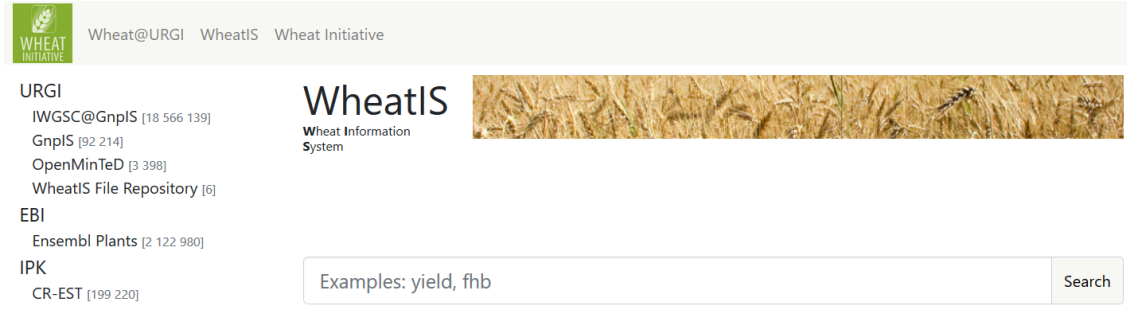
Alaux et al. Genome Biology 2018, https://doi.org/10.1186/s13059-018-1491-4

WheatIS data discovery tool: evolution



- ✓ Lighter implementation for new nodes
- ✓ Nodes cannot be « down »
- ✓ Easy addition of new functionalities

WheatIS data discovery tool: evolution



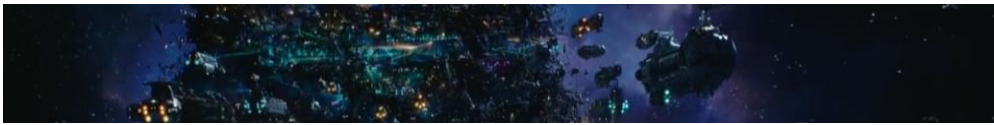
!beta! https://urgi.versailles.inra.fr/data-discovery-staging/

Based on user's remarks:

- New nodes
- New data
- New filters/data types

Open software, very generic, that can (and is) adapted to any type of federation:
e.g. the federation of information systems for of the french infrastructure of genetic resources for research in agriculture, AgroBRC-RARe **(!beta!** https://urgi.versailles.inra.fr/rare-beta/)

Deeper interoperability in federations



Breeding API initiative

http://www.brapi.org/

The Breeding API (BrAPI) Project is an international effort to create a RESTful specification that enables interoperability among plant breeding databases.

Bill & Melinda Gates Foundation

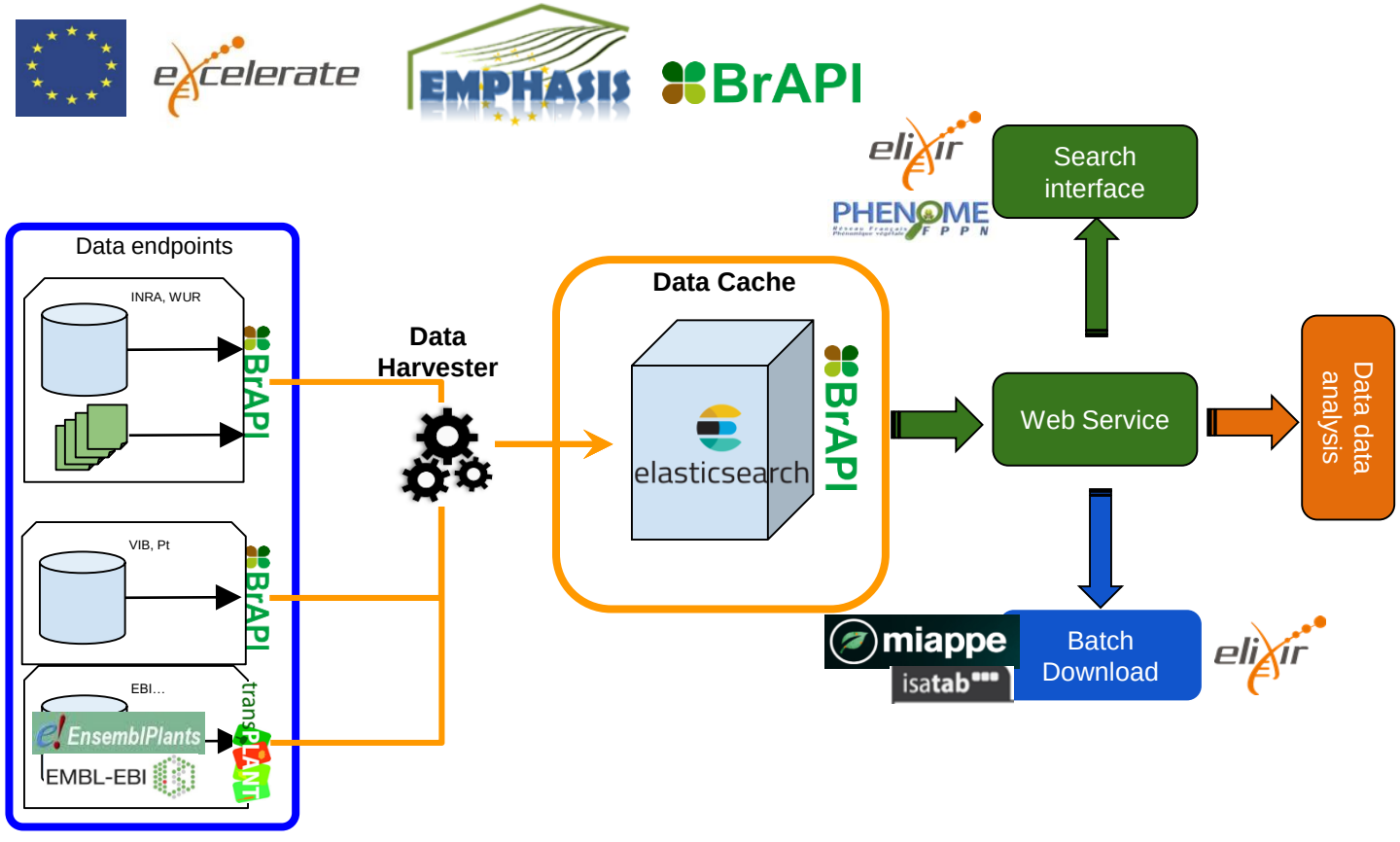
Breeding API initiative

http://www.brapi.org/

- Development of a **standard API :**
  - Calls for plant material aligned with MCPD
  - Calls for phenotyping experiments aligned with MIAPPE and of a supporting data model

- Next steps: develop the same type work on the calls for genotyping data (coordination with GA4GH)

Enabling improvements of data services



Conclusions



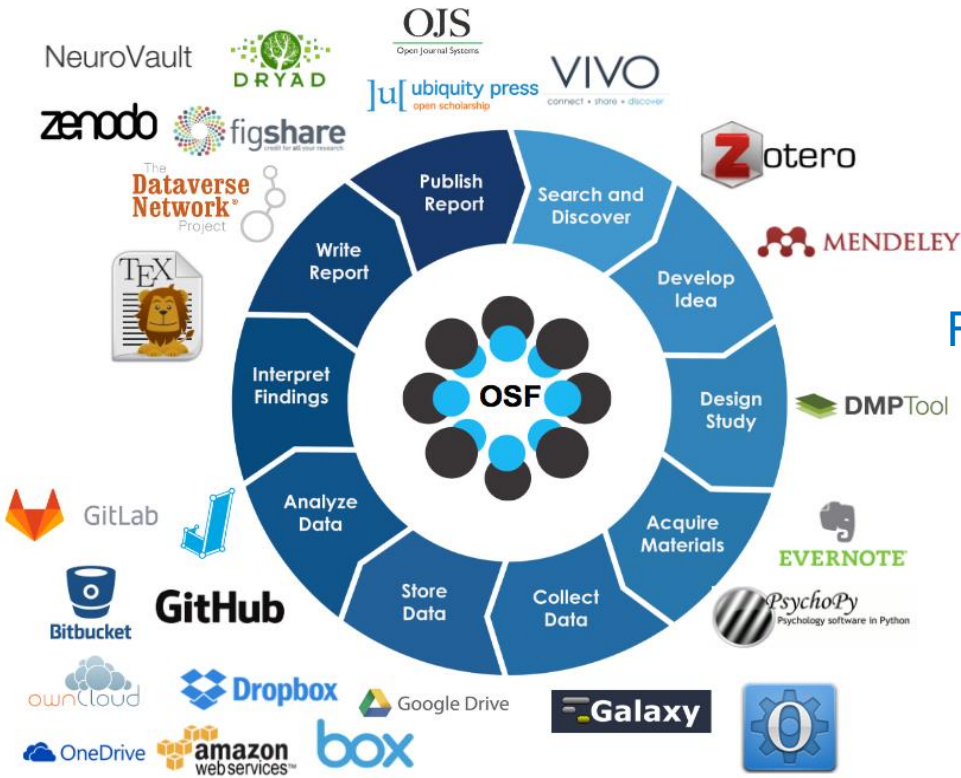
Conclusions and perspectives

- Generic search tool used for the WheatIS: light tool that allows community building -> Monitoring of the interest of end users to be done in more detail

- BrAPI : same good properties + the possibility to diversify services

- One big challenge still ahead: get the data and get it FAIR during all its life cycle.

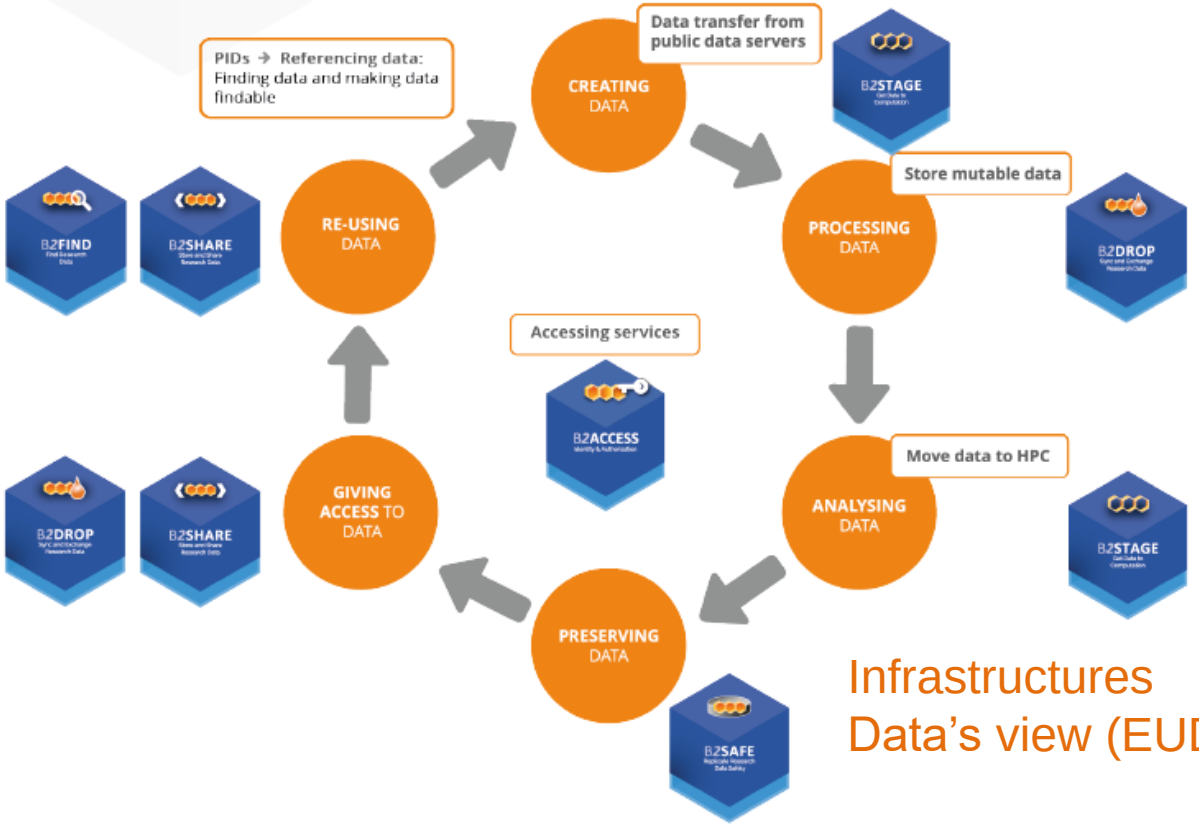
There is still a challenge in providing operational environments facilitating data management all along its life cycle



Researcher's view

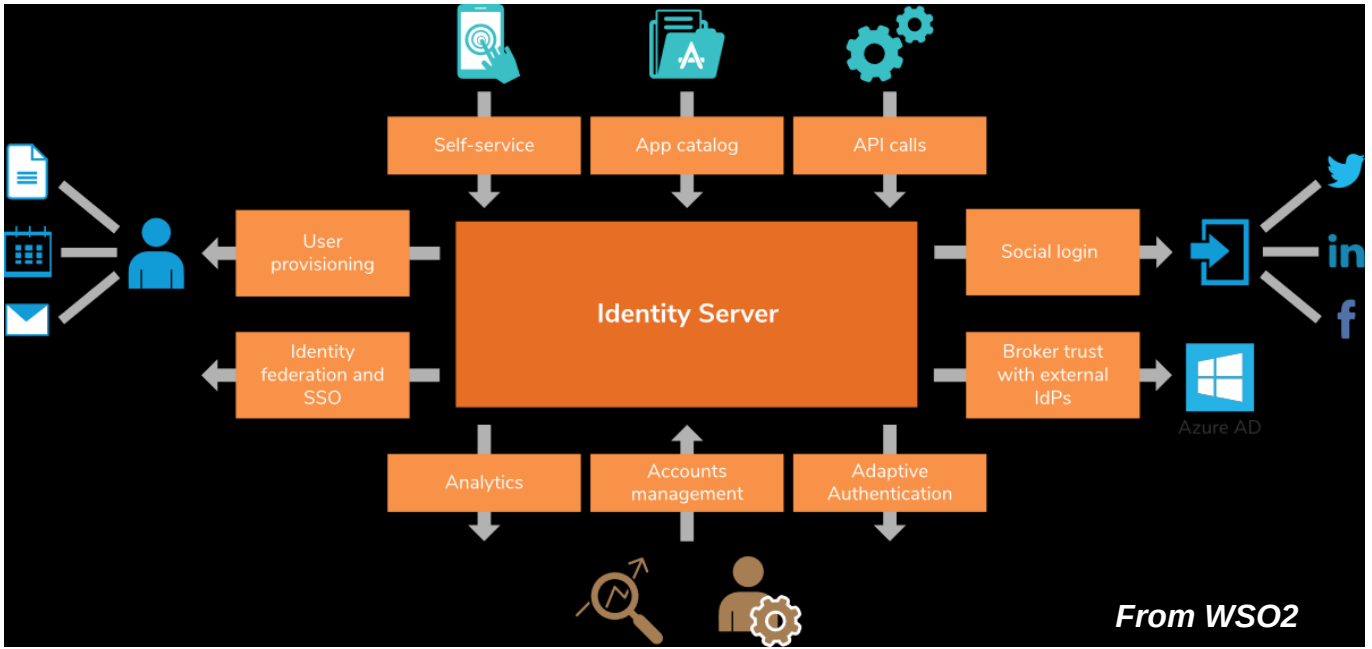
From Rowan University, NJ, USA

There is still a challenge in providing operational environments facilitating data management all along its life cycle



Infrastructures
Data's view (EUDAT)

There is still a challenge in providing operational environments facilitating data management all along its life cycle



Infrastructures
Identity and access view

Acknowledgements

H. Quesneville
C. Pommier
M. Alaux
M. Buy
D. Charruaud
G. Cornut
J. Destin
S. Diagne
S. Durand
B. El-Houdaigui
R. Flores
C. Guerche
E. Kimmel
T. Letelllier
C. Michotey
N. Mohellibi

National and International infrastructures/initiatives

PHENOME FPPN

transPLANT

elixir

EMPHASIS

openMINTED

WHEAT INITIATIVE

International Wheat Genome Sequencing Consortium

BrAPI

Bioversity International

Planteome

National and international Wheat projects

Breed wheat

Whealbi

Financial supports

INRA SCIENCE & IMPACT

ANR

Investissements d'avenir

RDA RESEARCH DATA ALLIANCE

Thank you!

WheatIS search usage metrics

2016

Summary			
Reported period	Year 2016		
First visit	01 Jul 2016 - 03:57		
Last visit	31 Dec 2016 - 00:42		
	Unique visitors	Number of visits	Pages
Viewed traffic *	**<= 460** Exact value not available in 'Year' view	**689** (1.49 visits/visitor)	**3,609** (5.23 Pages/Visit)
Not viewed traffic *			**110**

* Not viewed traffic includes traffic generated by robots, worms, or replies with special HTTP status codes.

2017

Summary			
Reported period	Year 2017		
First visit	01 Jan 2017 - 15:36		
Last visit	31 Dec 2017 - 08:09		
	Unique visitors	Number of visits	Pages
Viewed traffic *	**<= 1,556** Exact value not available in 'Year' view	**2,236** (1.43 visits/visitor)	**17,568** (7.85 Pages/Visit)
Not viewed traffic *			**996**

* Not viewed traffic includes traffic generated by robots, worms, or replies with special HTTP status codes.

2018

Summary			
Reported period	Year 2018		
First visit	07 Jan 2018 - 07:20		
Last visit	27 Nov 2018 - 07:33		
	Unique visitors	Number of visits	Pages
Viewed traffic *	**<= 1,931** Exact value not available in 'Year' view	**3,307** (1.71 visits/visitor)	**27,371** (8.27 Pages/Visit)
Not viewed traffic *			**604**

* Not viewed traffic includes traffic generated by robots, worms, or replies with special HTTP status codes.