# The ELIXIR infrastructure for plant phenotyping-genotyping data management

Anne-Françoise Adam-Blondon

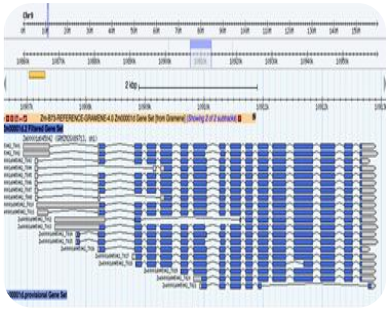
# The ELIXIR infrastructure for plant phenotyping-genotyping data management

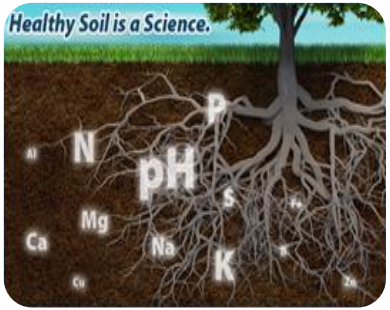
A-F Adam-Blondon, INRA, ELIXIR-FR


# Context


Genetic resources
Taxons
Genetic analyses
Phenotypes
Polymorphisms

# Predictive biology: Genomic x Environment x Phenomic



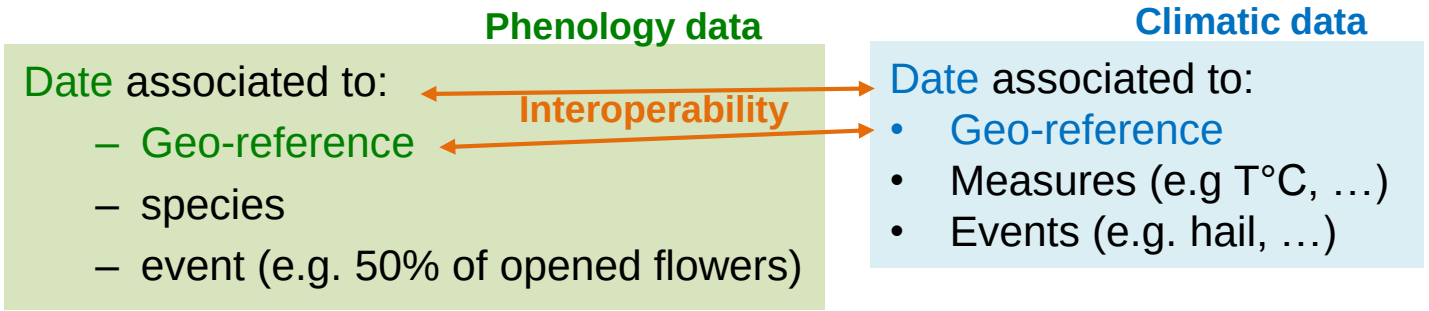



**Interoperability of heterogeneous data sets**

From C. Miguel and C. Pommier

# Reusing and integrating the data : e.g. 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 :** 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…

# Europe has been continuously reinforcing its policy for facilitating open access to data

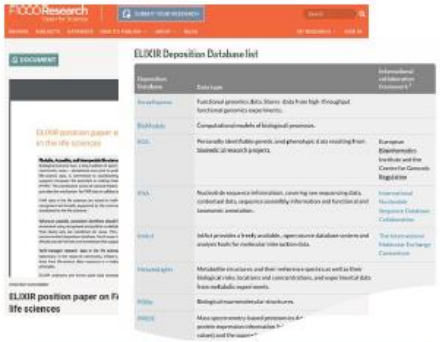


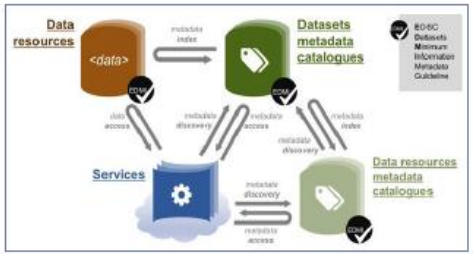
"**Facilitating access to results encourages the re-use of research outputs** and supports Open Science. This is essential for Europe's ability to enhance its economic performance and improve the capacity to compete through knowledge. […] Results of publicly-funded research can therefore be disseminated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

Data management plans are evaluated in EU projects proposals

# 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**

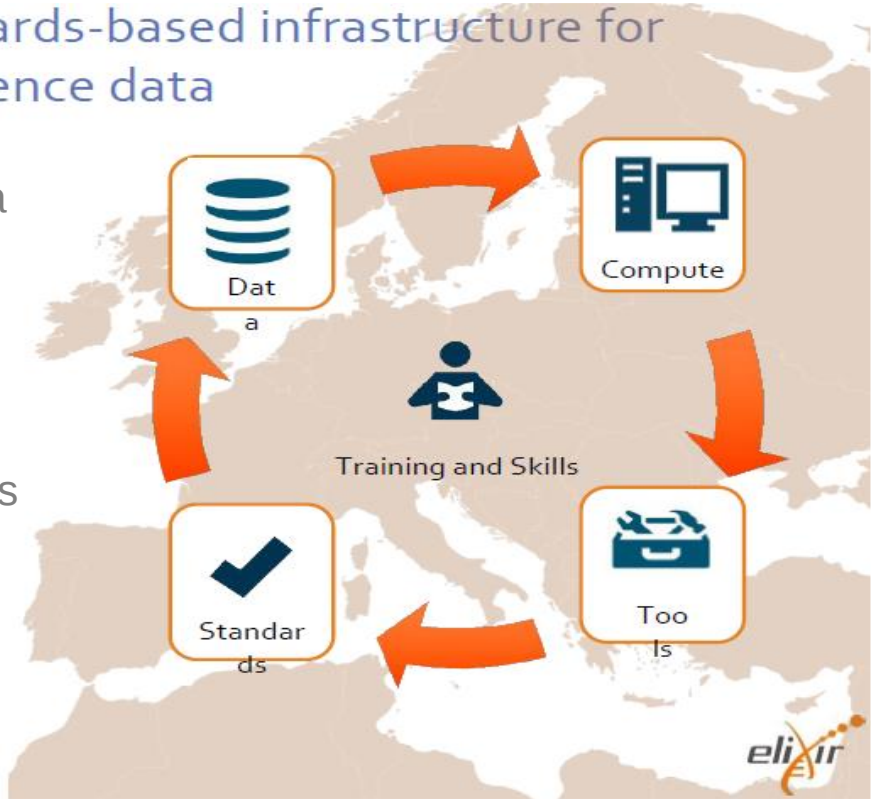
Metabolomics

Galaxy

Proteomics

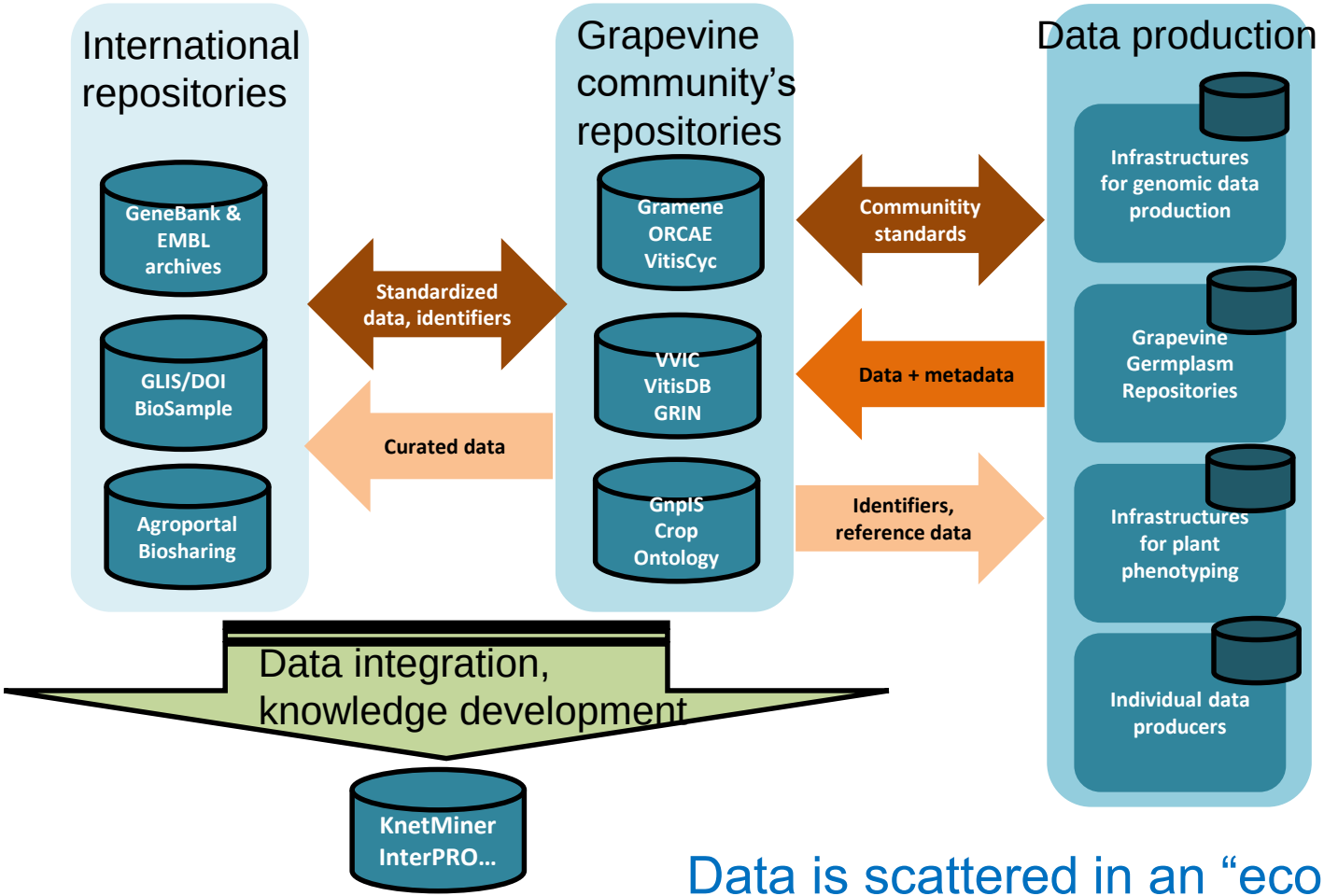
Rare diseases

...delivered in partnership with research communities




**Contacts Plants: Cyril Pommier (FR), Astrid Junker (DE),Kristina Gruden (SL)**

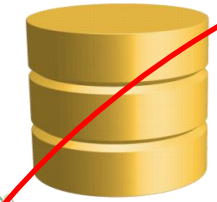
**From N. Blomberg**




Data is scattered in an "ecosystem" of information systems

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

# Insertion of data repositories in federations of information systems



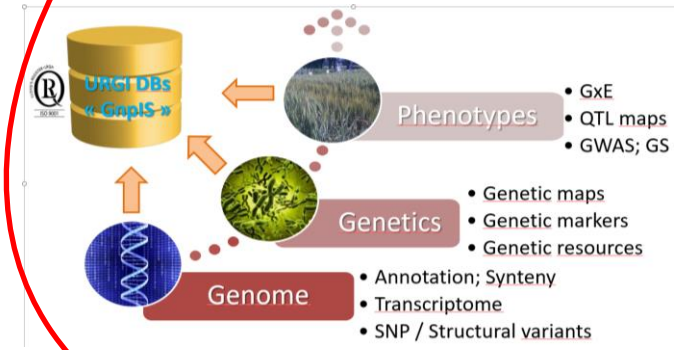
GnpIS-coreDB

2002: Single database

2008: DB Interoperability

Internal standards/identifiers are fine

2010: set of interoperable databases



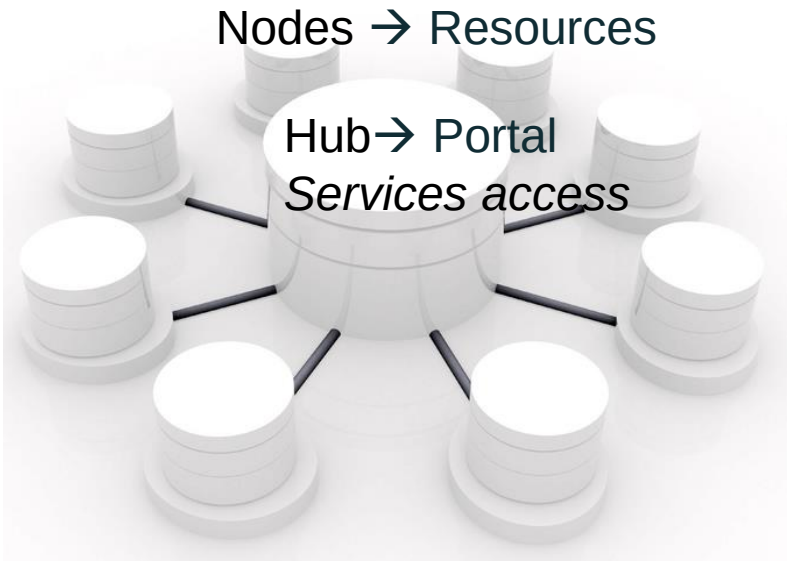

GnpIS Information System

2015: Distributed information system

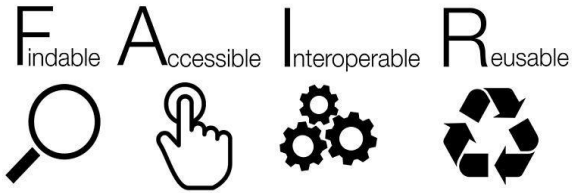
Internationally agreed standards and identifiers

# 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***

# Examples of federations


# Three “use cases” of federations

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

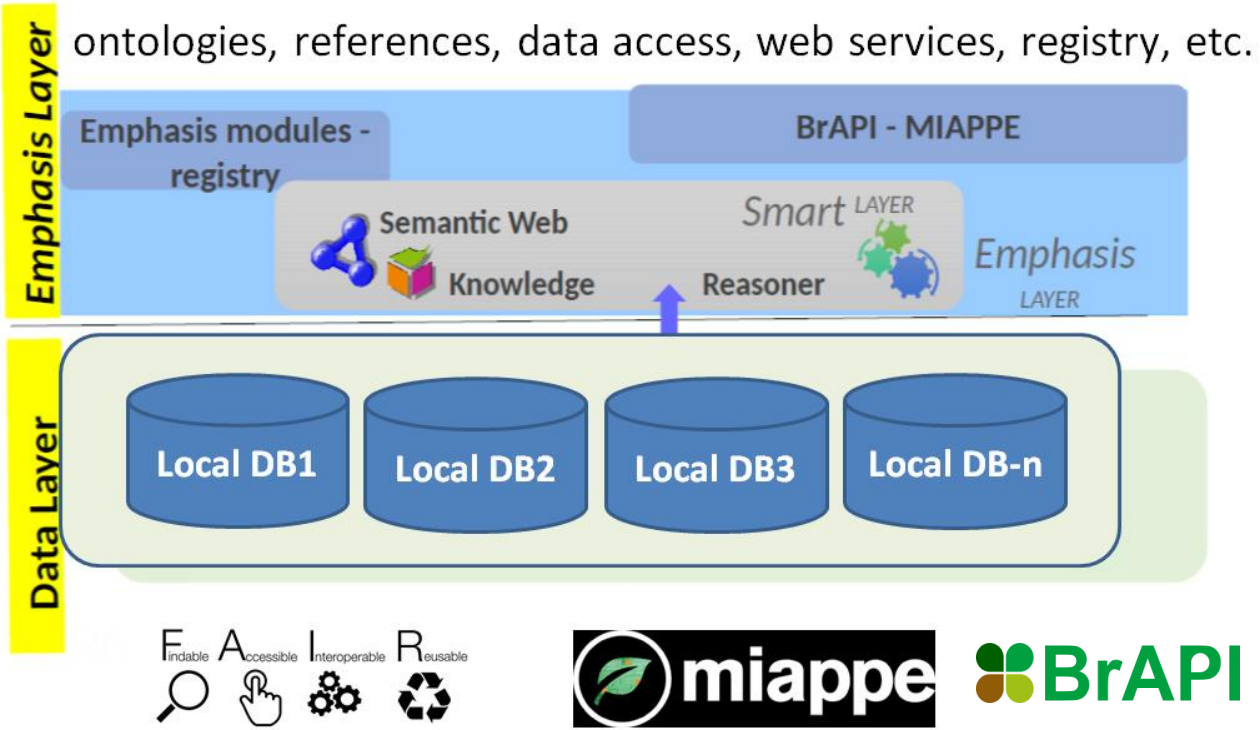

- The European Infrastructure for Multi-scale Plant Phenomics and Simulation (EMPHASIS) and its information system (https://emphasis.plant-phenotyping.eu/e-Infrastructure). In the frame of a strong collaboration with ELIXIR


- The data federation of the ELIXIR Plant Science community (https://elixir-europe.org/communities/plant-sciences )

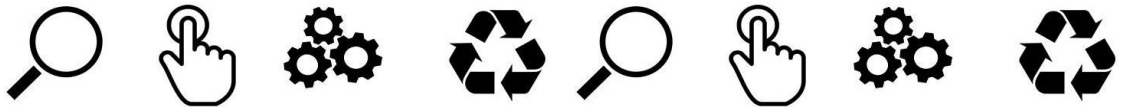
https://emphasis.plant-phenotyping.eu/e-Infrastructure

Services

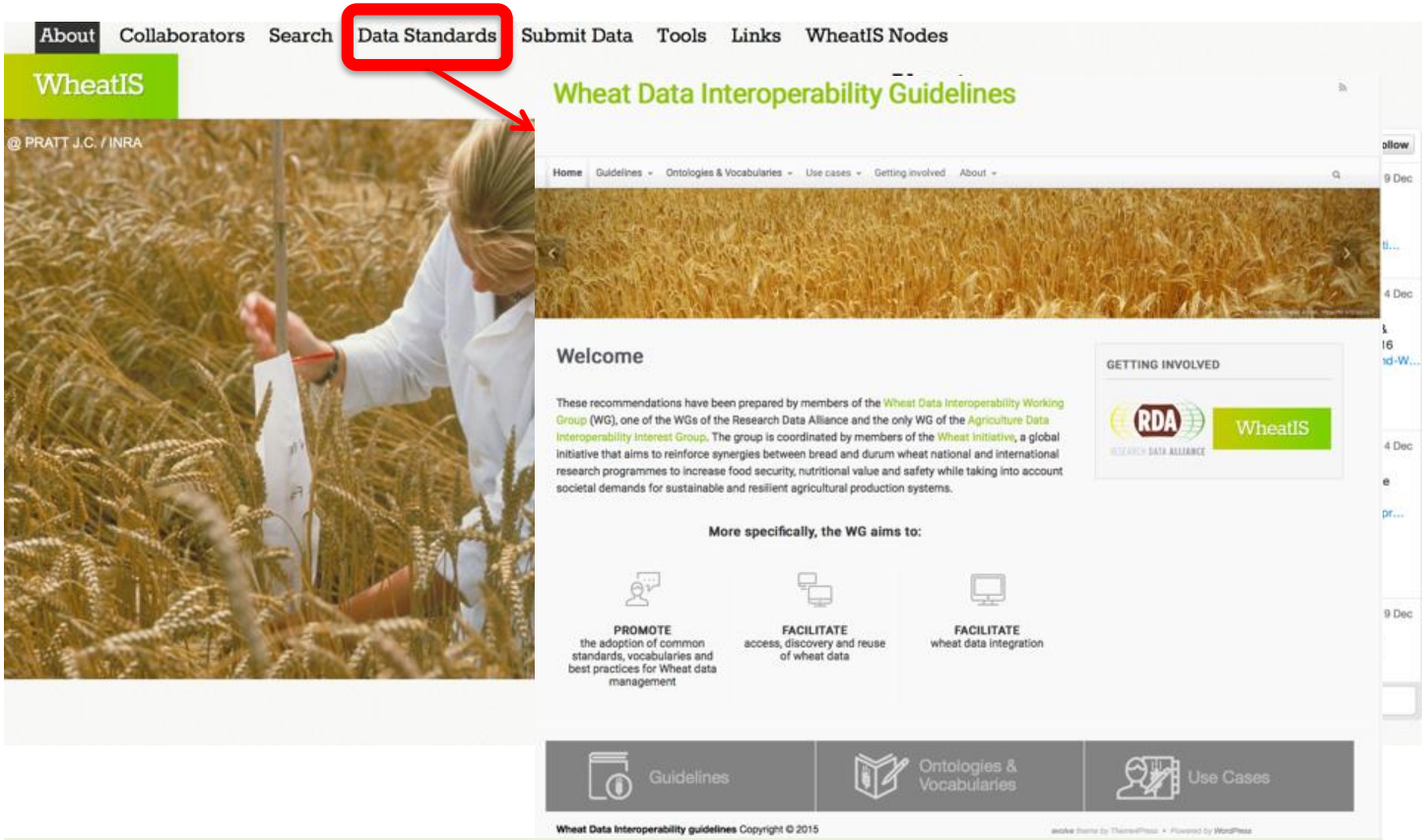
« Consumable » data



# Developing a federation of FAIR plant data repositories



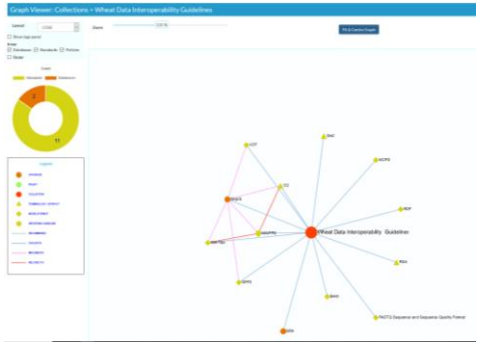
# Development of guidelines: e.g. www. wheatis.org

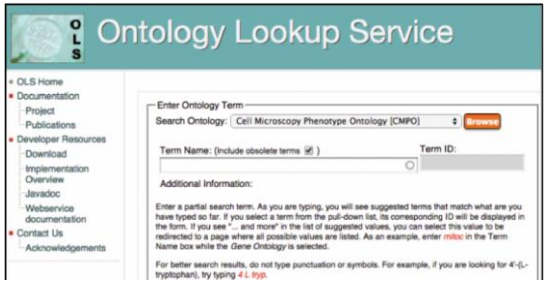



https://ist.blogs.inra.fr/wdi/

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

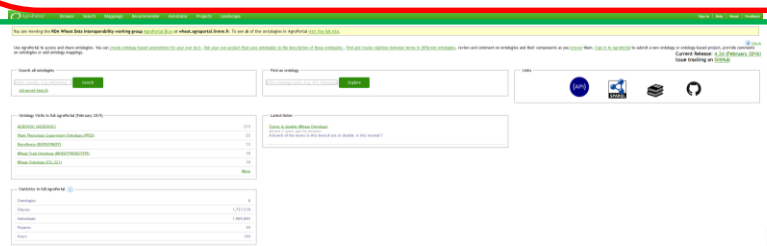
# Registries of standards and guidelines






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

International registries



http://agroportal.lirmm.fr/

Community registries

# Example of needed resources for a federation of phenotyping data

## Semantic

- Description of the data
- Controlled vocabularies: term name and definitions
- Ontologies: semantic links between terms
- ***Biologist driven***

## Structure

- Formatting and Organizing the data
- Data Models
- Standards : VCF, GFF, MIAPPE (www.miappe.org) , etc…
- ***Biologist & Computer scientist driven***

## Technical

- Data integration and sharing
- Interoperability : tools and systems
  - GA4GH
  - Breeding API www.brapi.org
- ***Computer scientist* driven**

**From C. Miguel and C. Pommier**

# 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: entity and quality), …
- Environments facilitating the development and curation of ontologies are needed

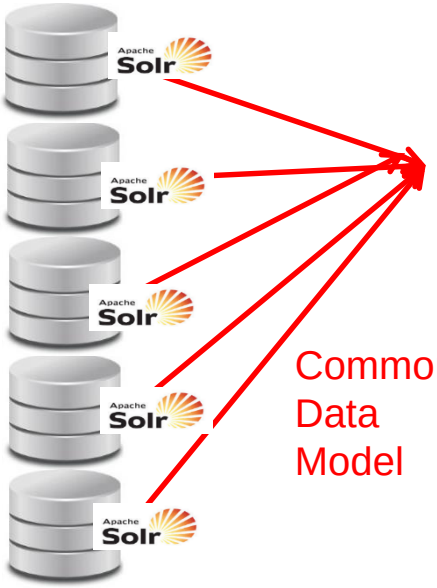
Bioversity International Planteome RDA RESEARCH DATA ALLIANCE WHEAT INITIATIVE excelerate EMPHASIS

# 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 services across federations: e.g. search and access data

# WheatIS Generic Data Discovery Tool



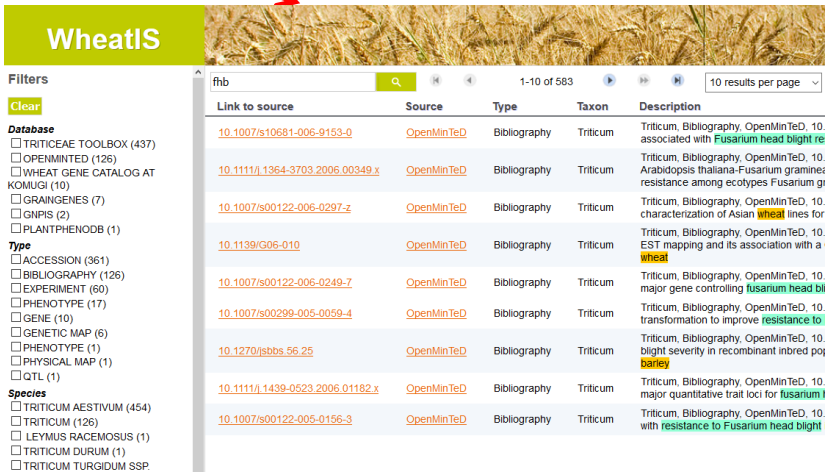

Common Data Model

User web

http://www.wheatis.org/

Google like list of results

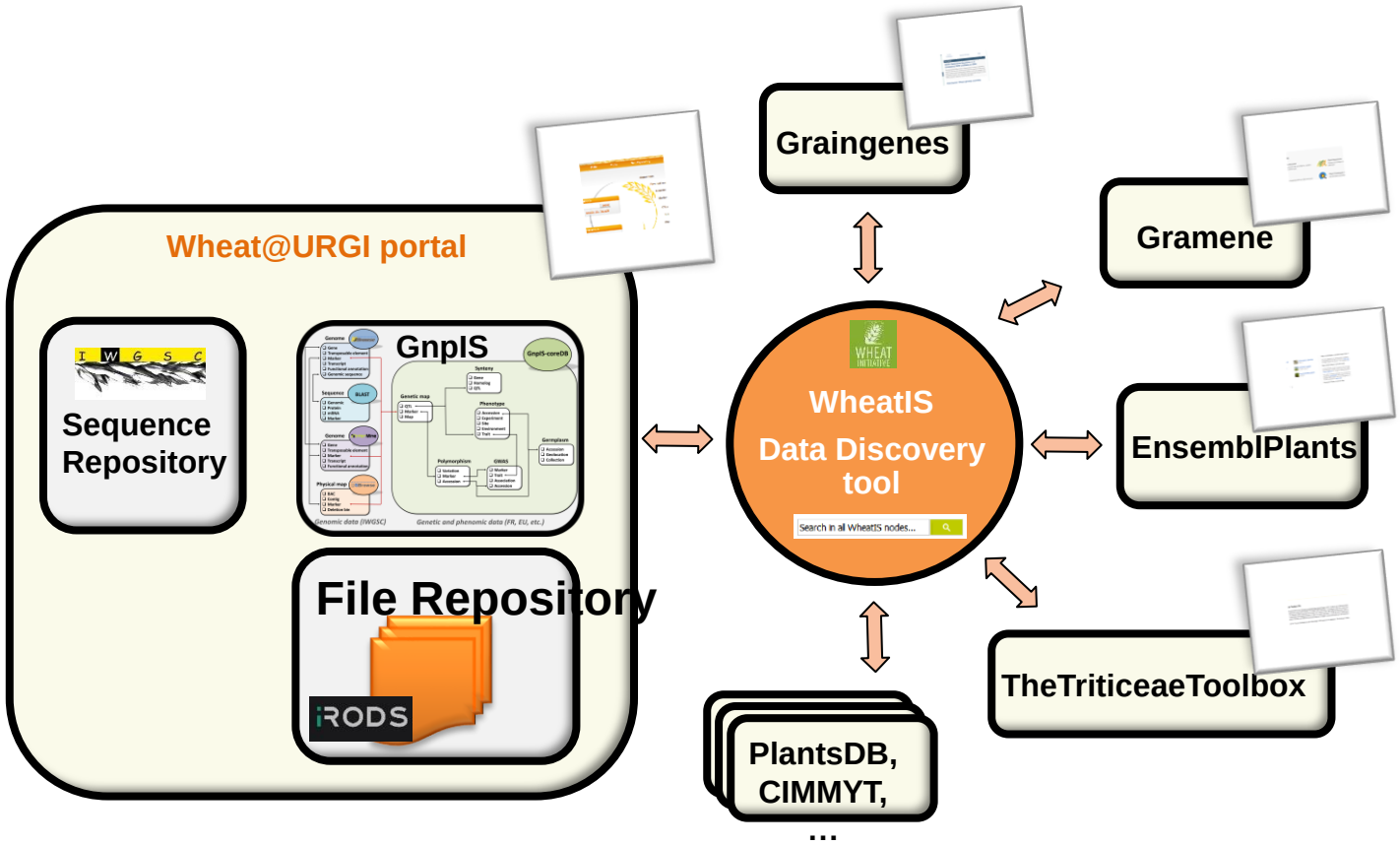
Proof of concept and success story




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

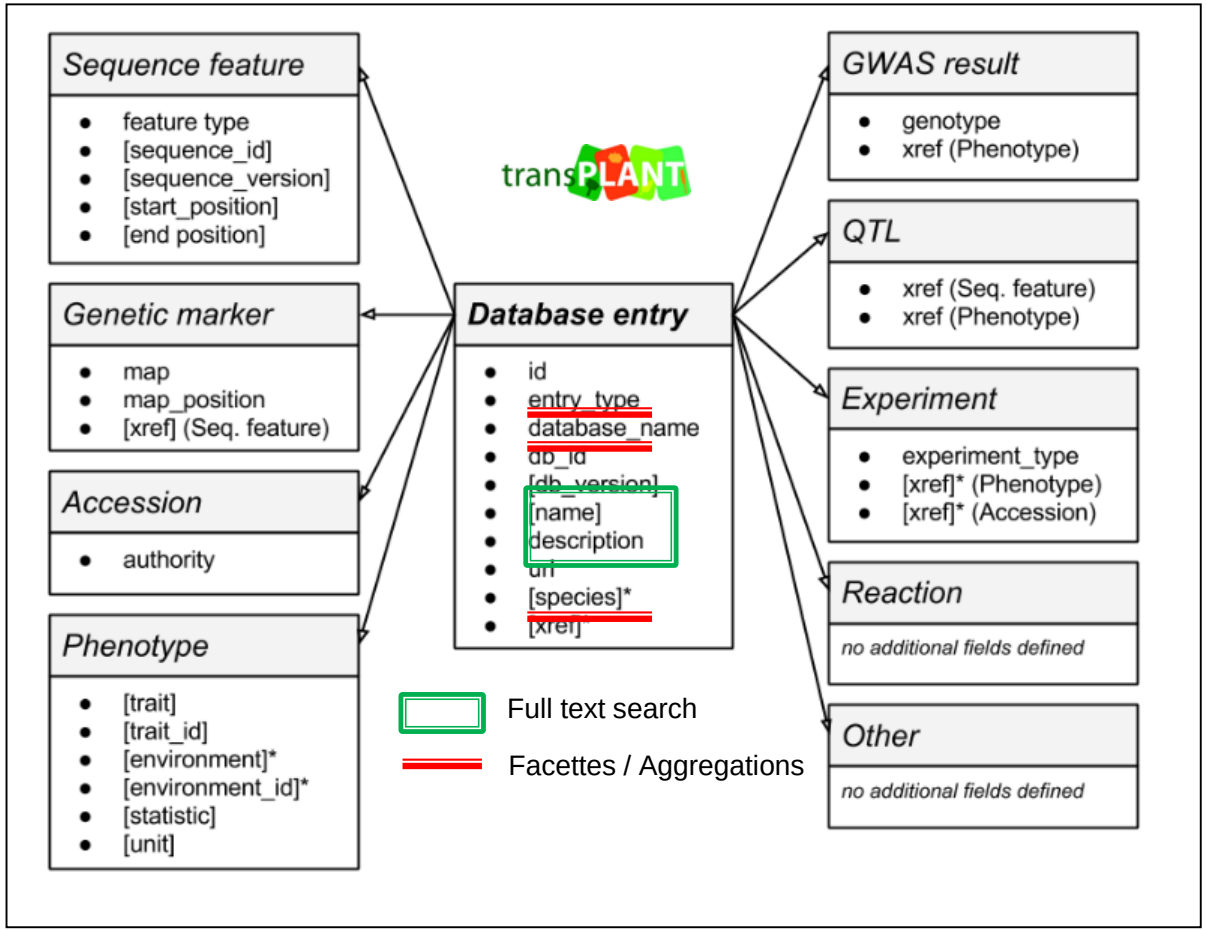
transPLANT
RDA RESEARCH DATA ALLIANCE
WHEAT INITIATIVE
ANR
IBISA INFRASTRUCTURES BIOLOGIE, SANTÉ ET AGRONOMIE

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




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

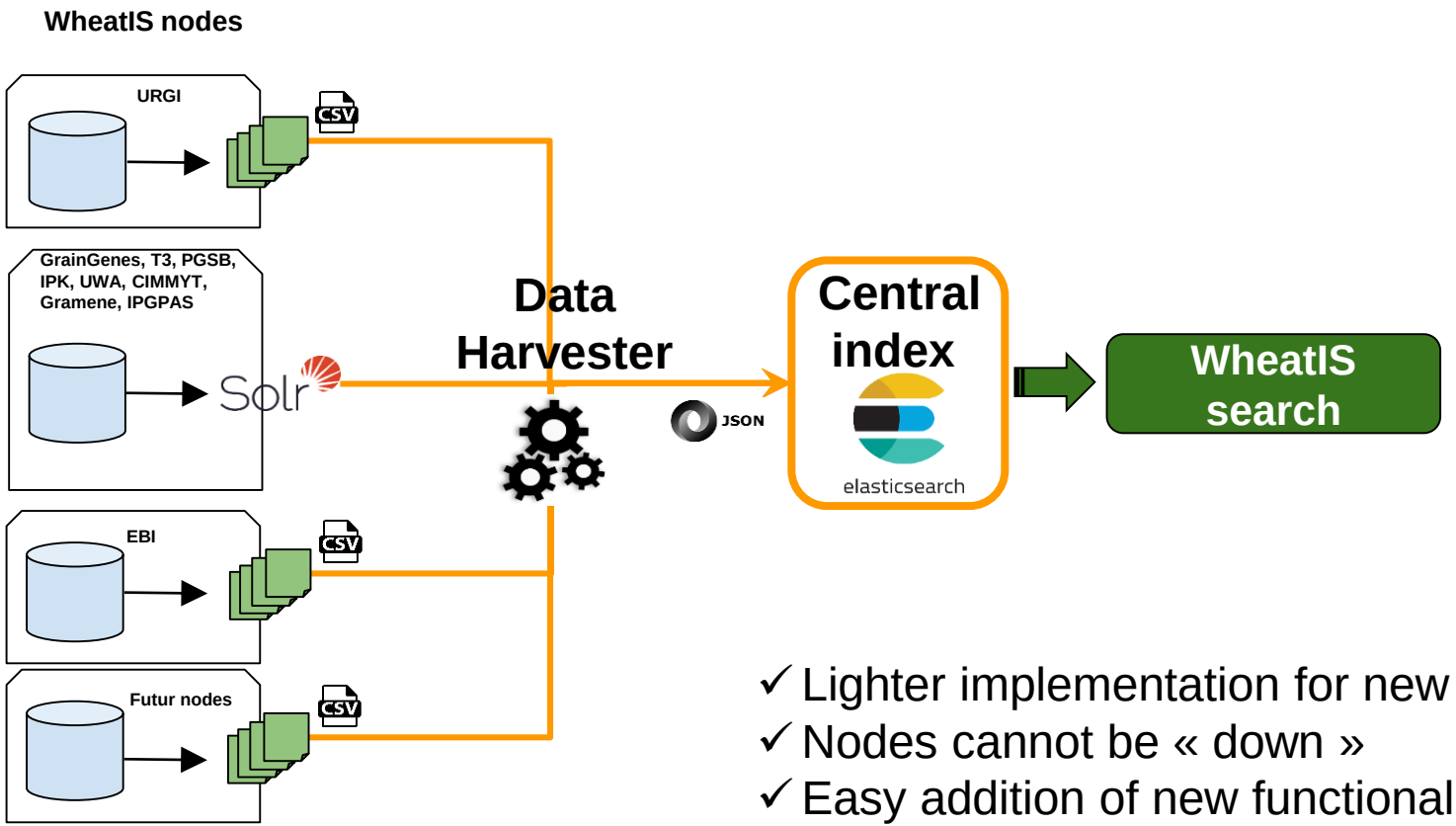
# transPLANT data model




Perspective: map the data model to Bioschema terms (schema.org) to enhance its web findability

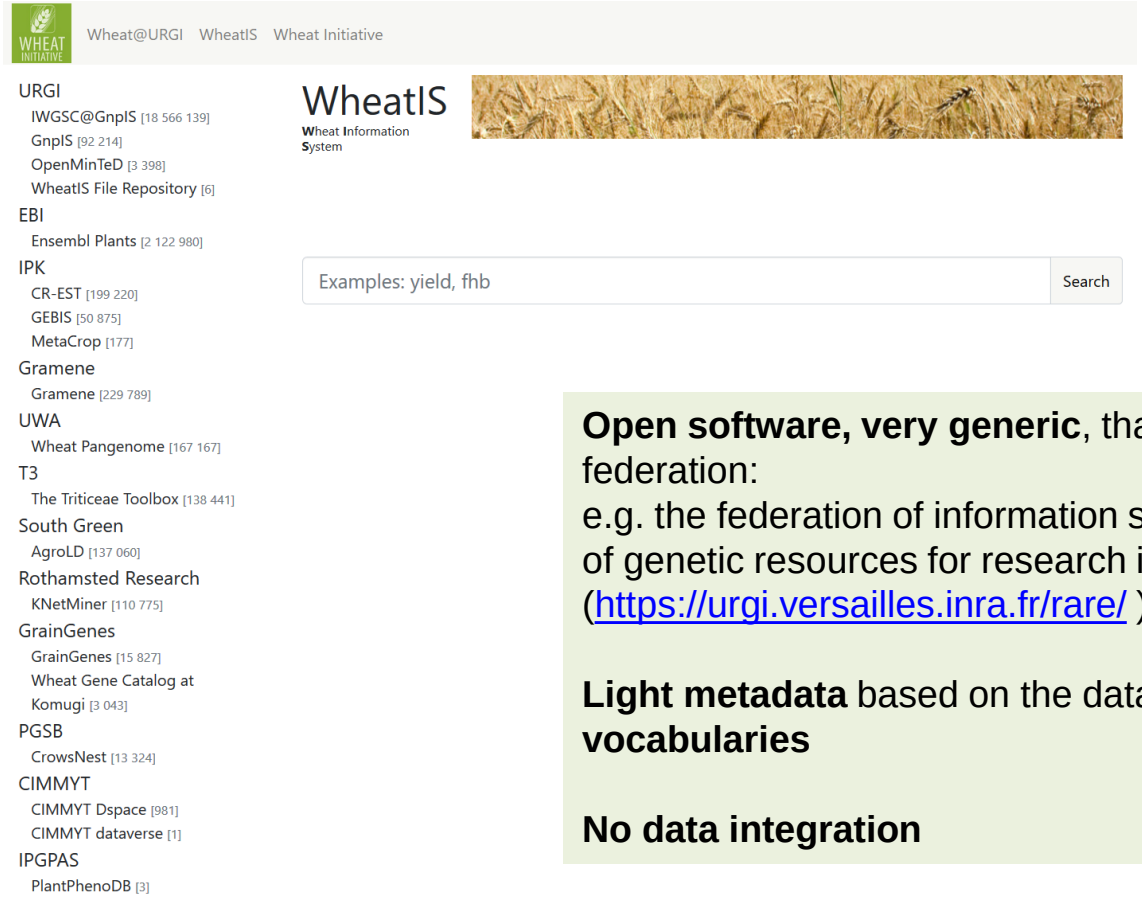
Bioschemas

# WheatIS data discovery tool: evolution




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

# WheatIS data discovery tool




**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 (https://urgi.versailles.inra.fr/rare/ )

**Light metadata** based on the data model and **no constraint on vocabularies**

**No data integration**

# Towards data integration in federations

# Breeding API initiative

http://www.brapi.org/

The Breeding API (BrAPI) Project is an **international effort** to create a RESTful specifications for a web service that enables interoperability among plant breeding databases.


**Bill & Melinda Gates Foundation**

# Breeding API initiative

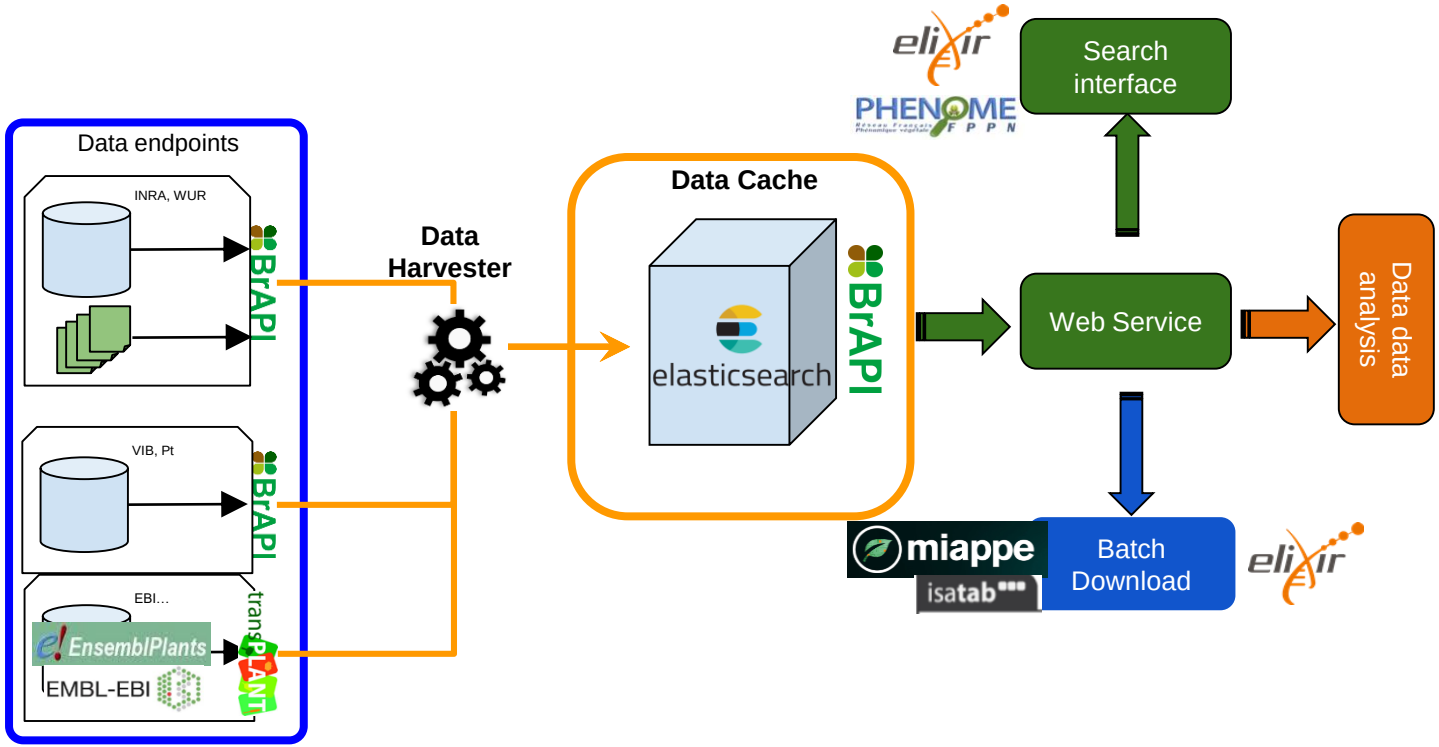
http://www.brapi.org/

- 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)

Selby *et al. Bioinformatics* (2019), doi.org/10.1093/bioinformatics/btz190

# Enabling improvements of data services by an international community

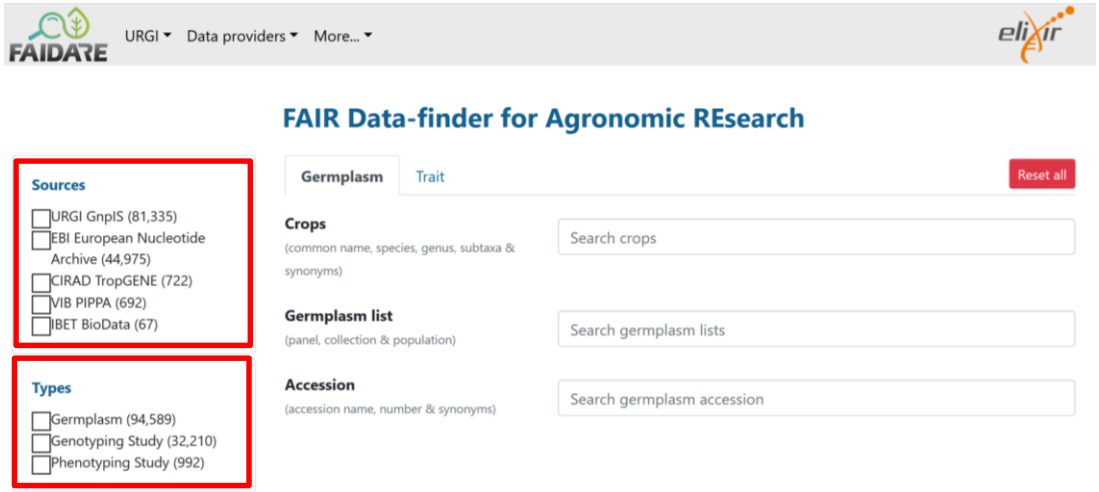


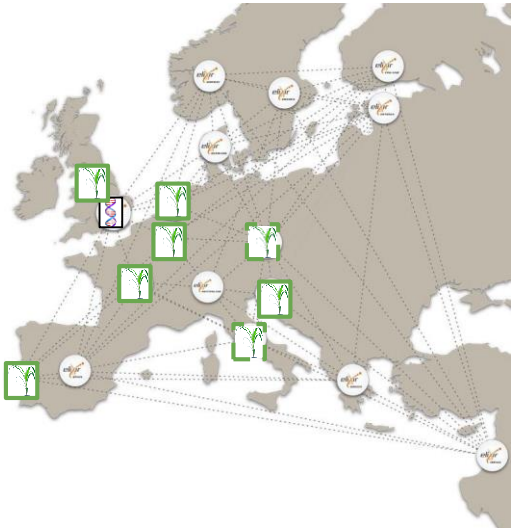

Anyone can develop a service that will « consume » the data of the federation exposed with the BrAPI standard: « machines understand your data »

excelerate EMPHASIS BrAPI

# ELIXIR Plant Data Search Service: FAIDARE

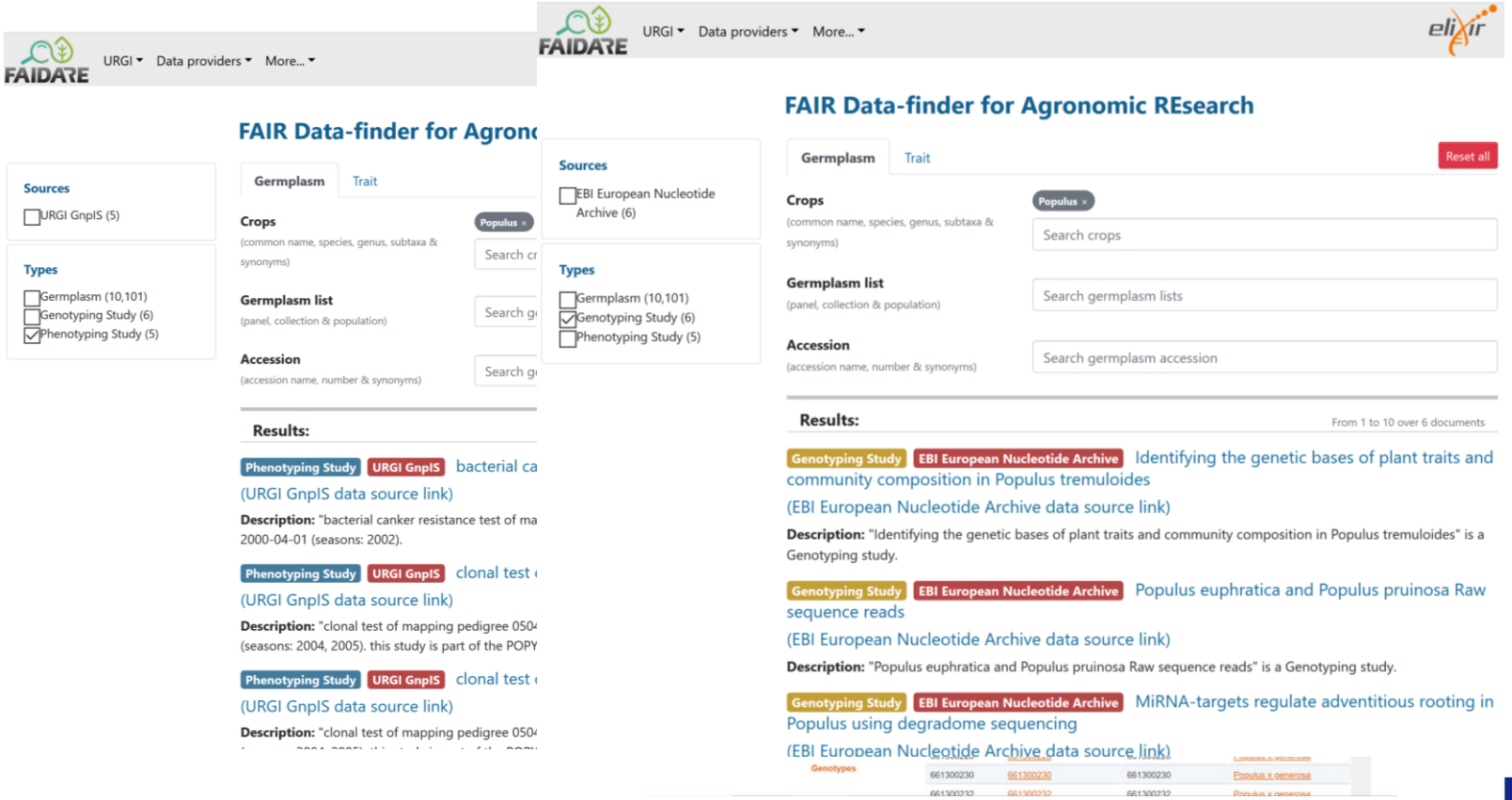
- Open source software
- Links
  - Genomic Centralised repository
  - Phenotype Distributed repositories



excelerate

# FAIDARE : Web Interface



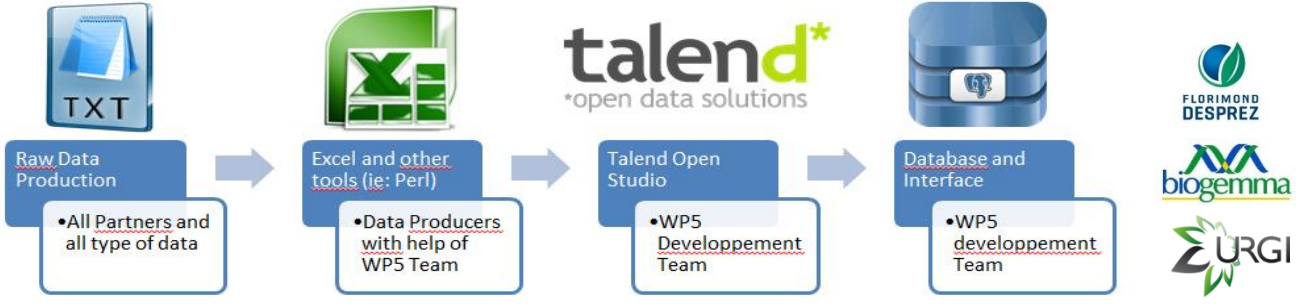
# How do we (easily) get the data FAIR into the data federation?



# Example of five large crop centered French national projects on sugar beet, maize, wheat, pea and rapeseed (2012-2020)

AKER, AMAIZING, BREEDWHEAT, PEAMUST, RAPSODYN Transverse actions in relation with data management

- GnpIS for the integration of heterogeneous public and private data according to the data management plans described in the consortium agreements
- Optimisation of the developments of GnpIS between projects
- Public private partnership to develop a suite of tools aiming at facilitating the insertion and integration of partner's data in GnpIS




Nb: these projects started before we knew about the FAIR principles and at the very start of the Wheat Initiative

# Publishing FAIR data in GnpIS

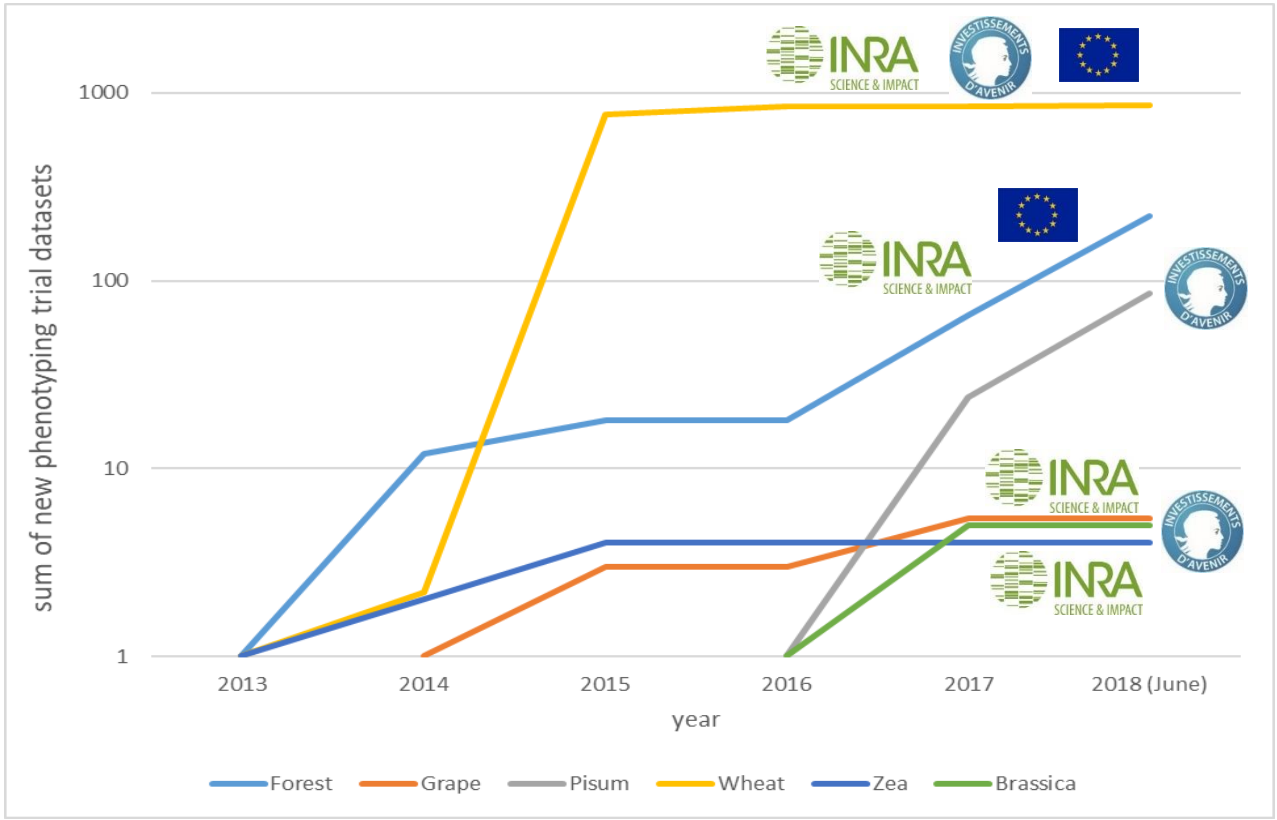
**Enforcement of good practices for data traceability in the consortia**

- Plant material identified under the responsibility of the genebanks
- Development of crop ontologies for phenotypic variables (e.g. https://www.cropontology.org/ontology/CO_333/Beet%20Ontology )

| | Genetic Resources | | Phenotyping | | Genotyping | | Association | |
|---|---|---|---|---|---|---|---|---|
| | Public | Private | Public | Private | Public | Private | Public | Private |
| **Brassica** | 981 | **18** | 5 | | | | | |
| **Beta** | 10783 | | 5 | | | | | |
| **Zea** | 1861 | **869** | 20 | 3 | 1 | **4** | 3 | |
| **Pisum** | 1015 | **872** | 0 | **86** | | **3** | | |
| **Wheat** | 10448 | **2811** | 821 | **57** | 1 | **22** | 43 | **1970** |

| | Sequence | | Genetic Map | | Genome browser | |
|---|---|---|---|---|---|---|
| | Public | Private | Public | Private | Public | Private |
| **Brassica** | 4 | **2** | 2 | | 1 | |
| **Beta** | | | | | | |
| **Zea** | 8 | **8** | 18 | **21** | 4 | **3** |
| **Pisum** | 5 | **2** | | **12** | 1 | |
| **Wheat** | 7 | **27** | 29 | **1** | 6 | |

# Publishing FAIR data in GnpIS




**PHENOTYPING DATASETS**

Active data stewardship helps a lot

# Publishing FAIR data in GnpIS

**Enforcement of good practices for data traceability in the consortia**

- Association of DOI to datasets using data.inra before integration in GnpIS (e.g. DROPS/AMAIZING dataset)


Improve/have a control on the FAIRness of the data sets that are deposited

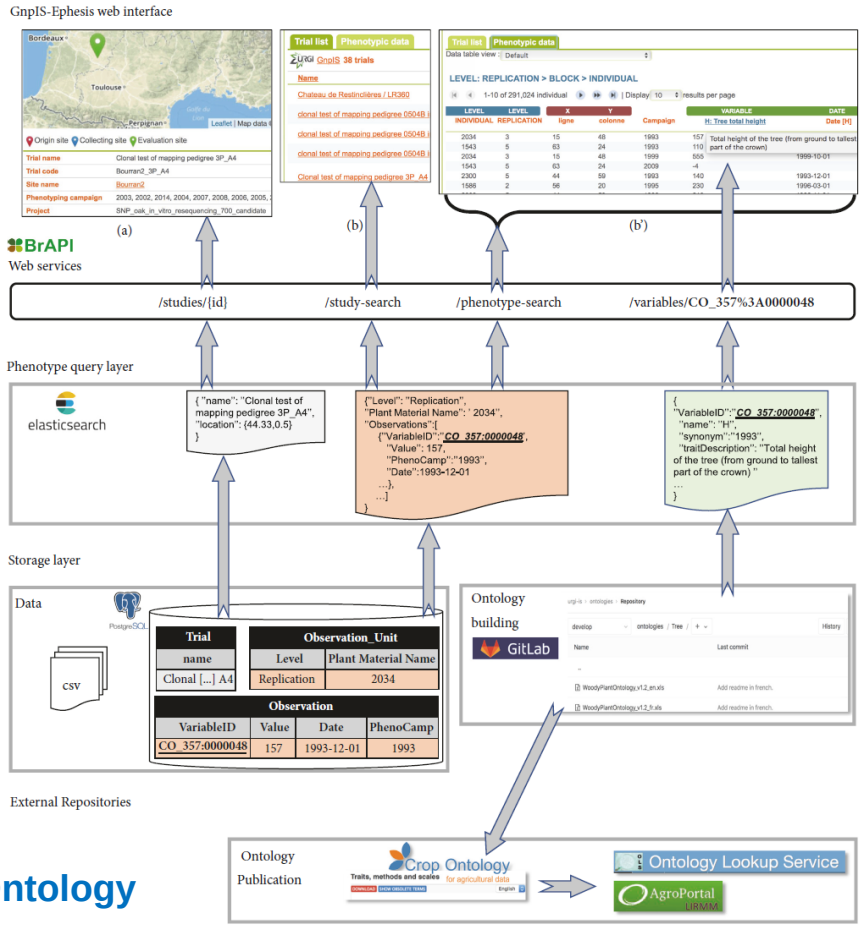
# Enabling GnpIS insertion in data federation

**Web Interface: get MIAPPE compliant ISATab data**

**BrAPI: MIAPPE compliant programatic access**

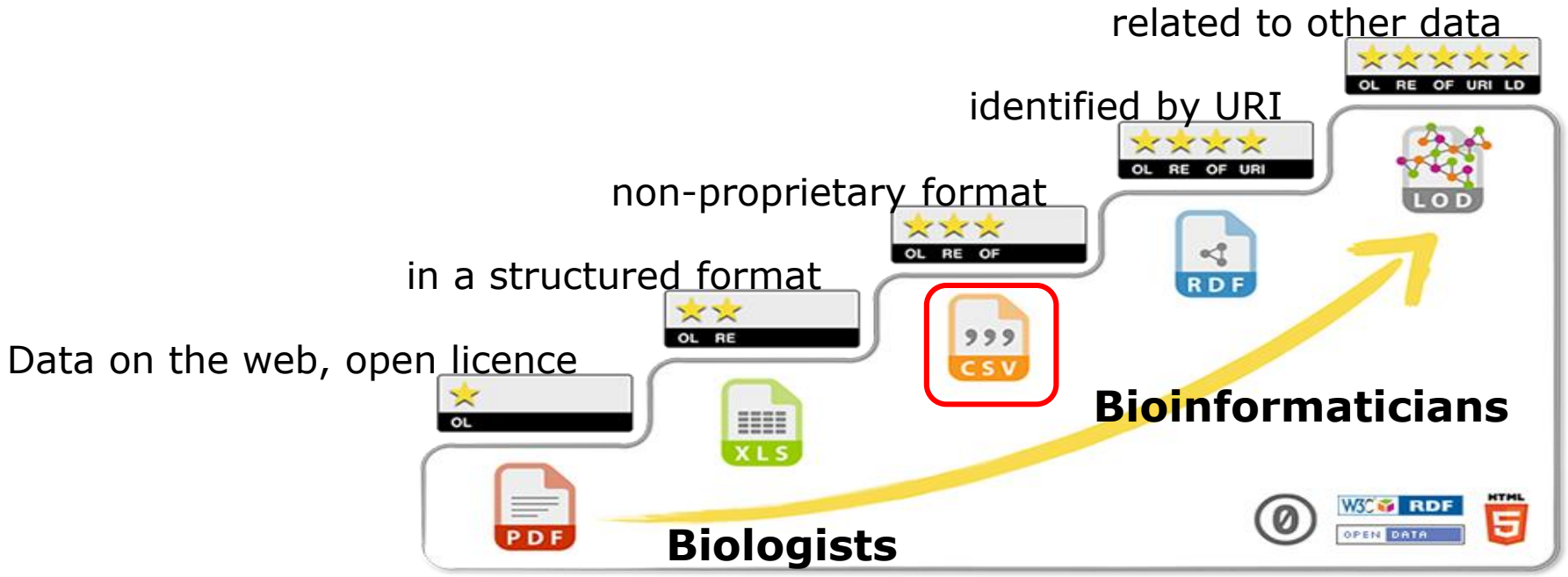
**Data submission to GnpIS compliant with MIAPPE requirements**

**Collaboration with the Crop Ontology**



# Conclusions


# 5 * Open Data

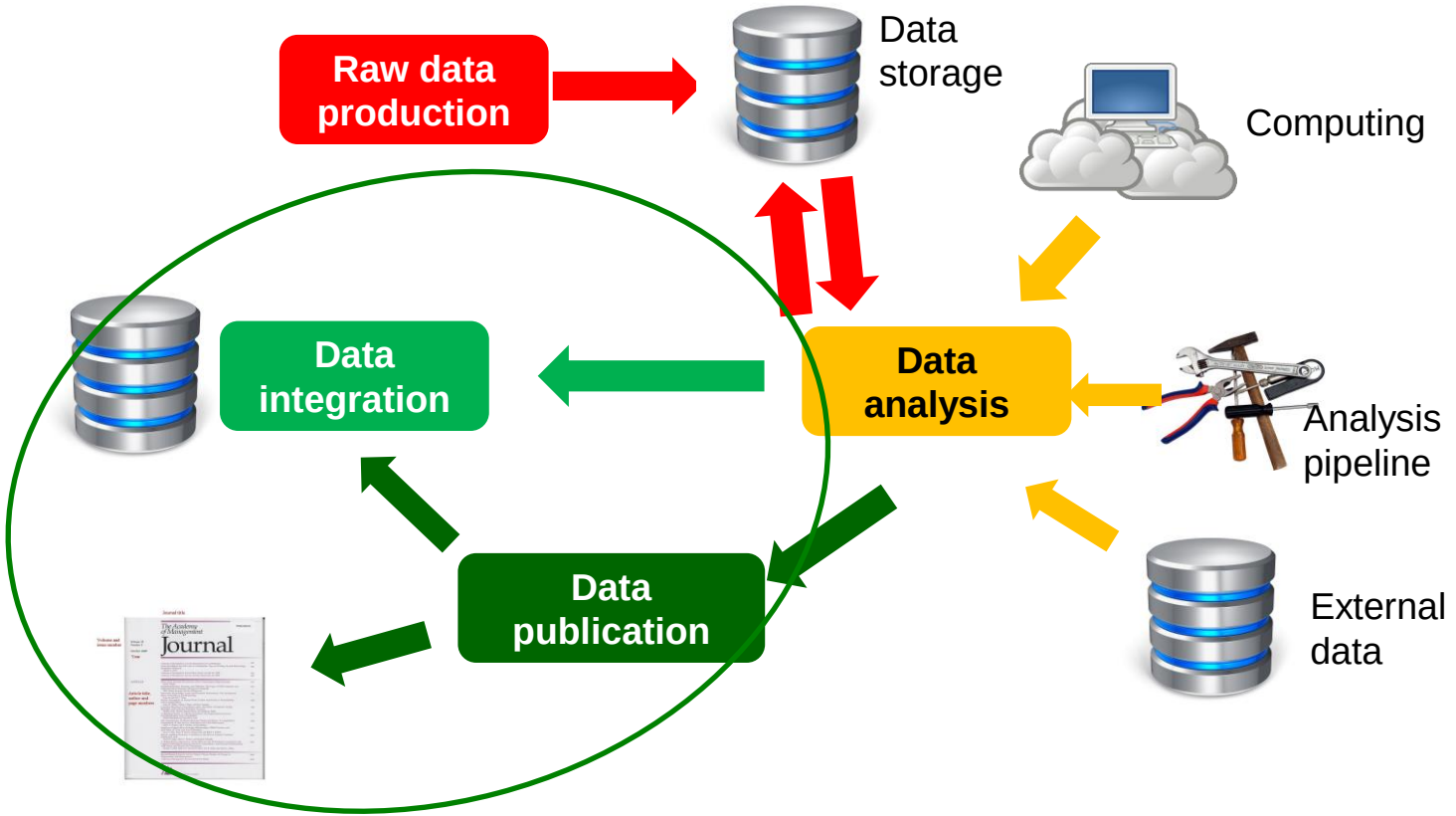



Courtesy of Daniel Jacob

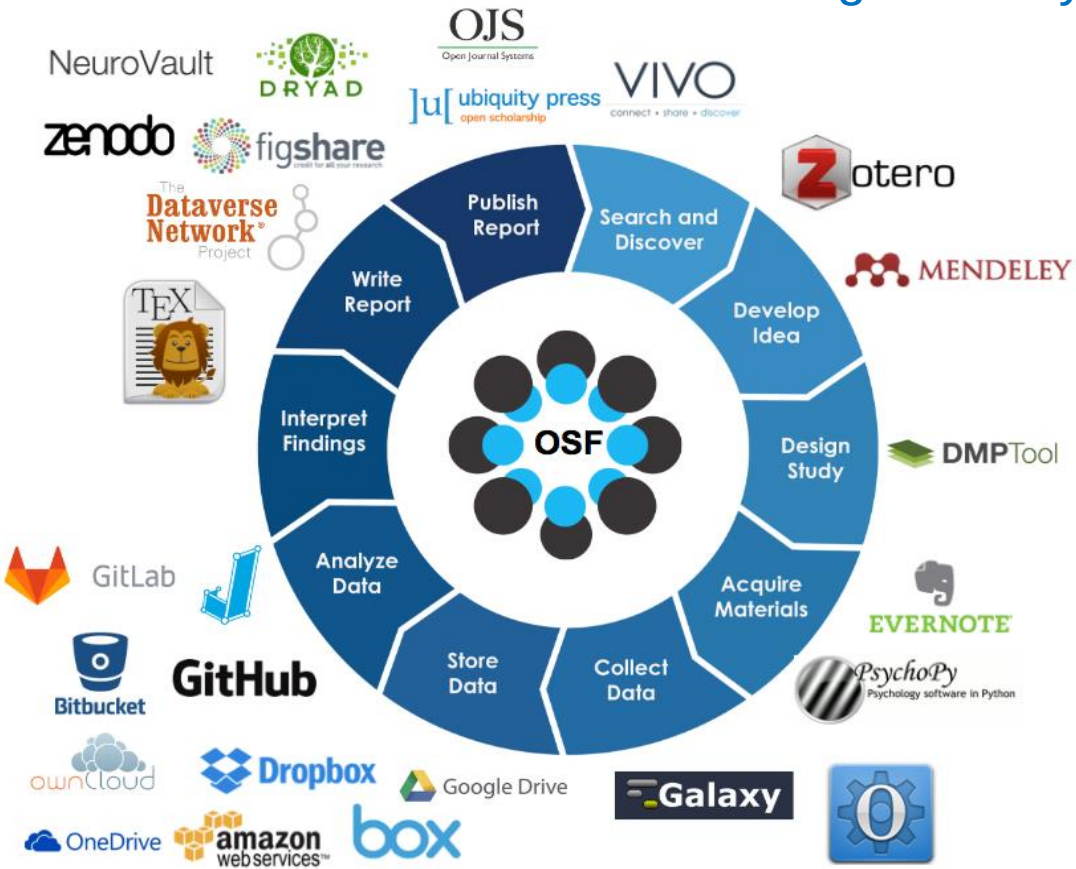
Progressing towards FAIR and Open Data requires a **multidisciplinary cooperation** :

- Biologists
- Bioinformaticians
- specialists of ontologies/semantics

# Data life cycle in (plant) biology research activities



# Challenge: provide operational environments facilitating data management all along its life cycle




**From Rowan University, NJ, USA**

elixir

**Priority 2020-2024 :** ELIXIR-Converge H2020 project.

Use case Plant coordinated by ELIXIR-PT

# ELIXIR Plant community

Thank you!

www.elixir-europe.org

@ELIXIREurope

/company/elixir-europe

# Acknowledgements

## URGI team

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
Open Mining Infrastructure for Text & Data
WHEAT INITIATIVE
International Wheat Genome Sequencing Consortium
BrAPI
Bioversity International
Planteome

## Other national and international projects

AKER Betterave2020
amaizing
Breed wheat
PeaMUST
RapsodyN
GENTREE
Whealbi
GenOak

## Financial supports

INRA SCIENCE & IMPACT
ANR
Investissements d'avenir
RDA RESEARCH DATA ALLIANCE