Data management in large French Public-Private Projects

Anne-Françoise Adam-Blondon

Unit of Research in Genomic-Info (URGI), INRA

Large french PPP

Agence Nationale de la Recherche ANR

AKER Betterave2020

aMaizing

Breed Wheat

BFF Biomass For the Future

PeaMUST

RapsodyN

SUNRISE UNE CULTURE POUR LE FUTUR

Partners involved in data management

IRISA

URGI

IGEPP Institut de Génétique, Environnement et Protection des Plantes

Agroécologie Dijon Unité de Recherche

GDEC Génétique Diversité Ecophysiologie des Céréales

agap

LIRIS

Lipm Laboratoire Interactions Plantes Micro-organismes

LIMOS

INRA SCIENCE & IMPACT

cirad

cnrs

FLORIMOND DESPREZ

biogemma

syngenta SEEDS

Data management in PIA projects

Data management has been organized across six 2012-2020 projects

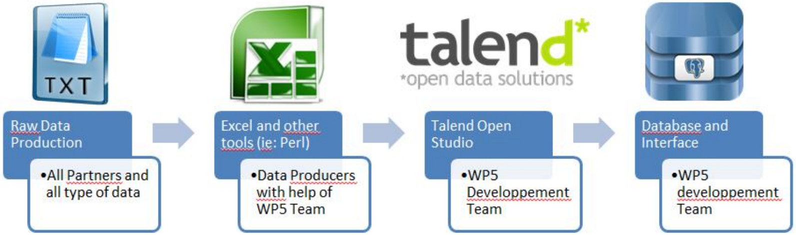
- **Tranverse actions in relation with data management : AKER (sugar Beet), AMAIZING (maize), BREEDWHEAT (wheat), PEAMUST (pea), RAPSODYN (rapeseed), BFF (energy crops) + PHENOME = EPPN-FR**
- **The principles were set up by the public and private partners together:**
  - **INRA GnpIS information system for the integration of heterogeneous public and private data**
  - **Optimisation of the developments of GnpIS between projects.**

Different possible architectures:

- GnpIS in complement of local information systems (e.g. Amaizing)
- GnpIS used for all the data (public and private (e.g. BreedWheat, Peamust)
- A private instance of GnpIS used for data management (e.g. AKER, Rapsodyn)
- Some projects use GnpIS only for one type of information (BFF)

Implementation of the DMP of the consortium agreement

- **Implementation of the data managements plans described in the consortia agreements in the tools and data management processes**
- **Public-private partnership to develop a suite of tools aiming at facilitating the insertion and integration of partner's data in GnpIS**



Raw Data Production
- All Partners and all type of data

Excel and other tools (ie: Perl)
- Data Producers with help of WP5 Team

Talend Open Studio
- WP5 Developpement Team

Database and Interface
- WP5 developpement Team

FLORIMOND DESPREZ

biogemma

URGI

Genetic ressources
Phenotypic data
Genotypes
SNP detection
Association genetics
Transcriptomics
Genetic maps
Genomic selection (to be done)

Creation of a community of practices

- **These projects were a very efficient lever to evolve the practices of the french crops communities in terms of data standardization**

  - Common standards of metadata (data about the experiment)
  - Plant material identified under the responsibility of the genebanks
  - Development of ontologies for phenotypic variables under a crop ontology format

Variable = Trait + unit + protocol

Objectives of using the same information system (or file repository) for data management

- Facilitate access to data for the partners of the project : private data

- Give access to public data to an international community of researchers and breeders : data that follows the FAIR principles (***F****indable,* ***A****ccessible,* ***I****nteroperable,* ***R****eusable)*

- To integrate the data with other data: data has to be interoperable and reusable

Data Findable, Accessible, Interoperable and Reusable

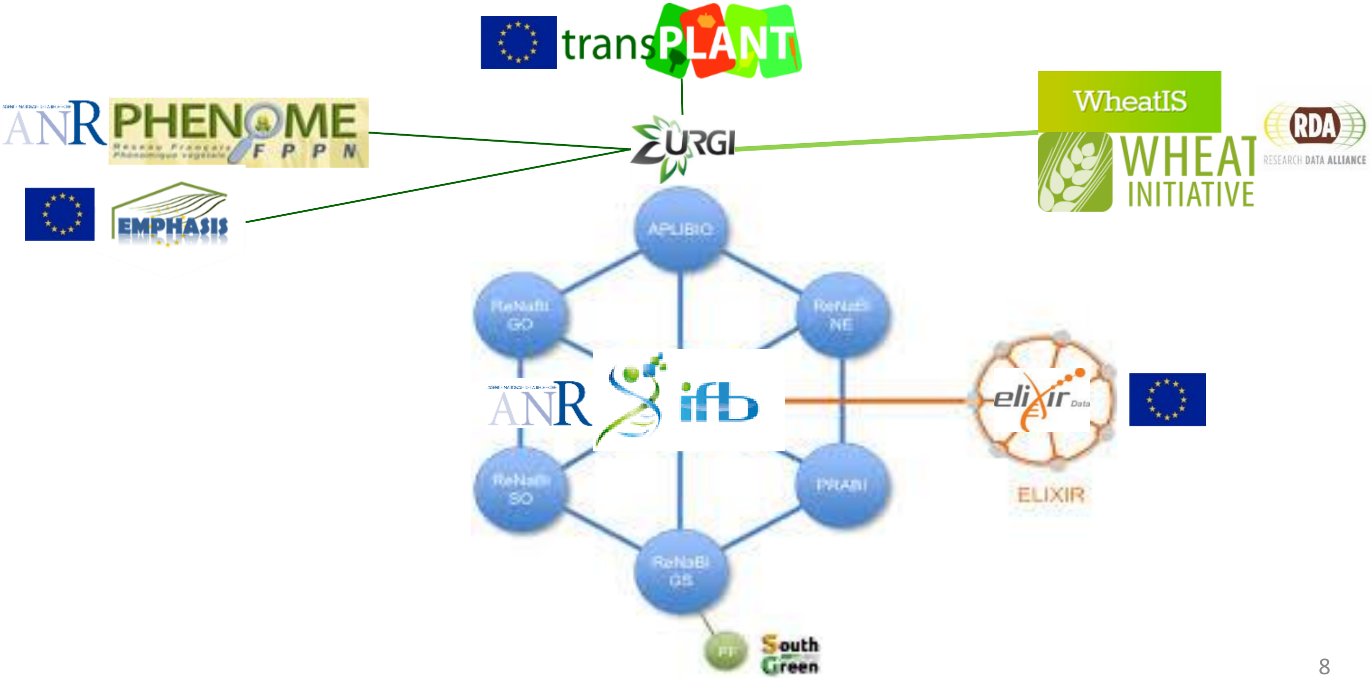
- Use of international (meta)data standards as much as possible (Reusability)

- Use of a consistent identification system for key objects for interoperability: plant material, genes, markers, phenotypes, …(Interoperability)

- Develop search web tools based on generic data models, especially when dispersed in different information systems (Findability and Accessibility)

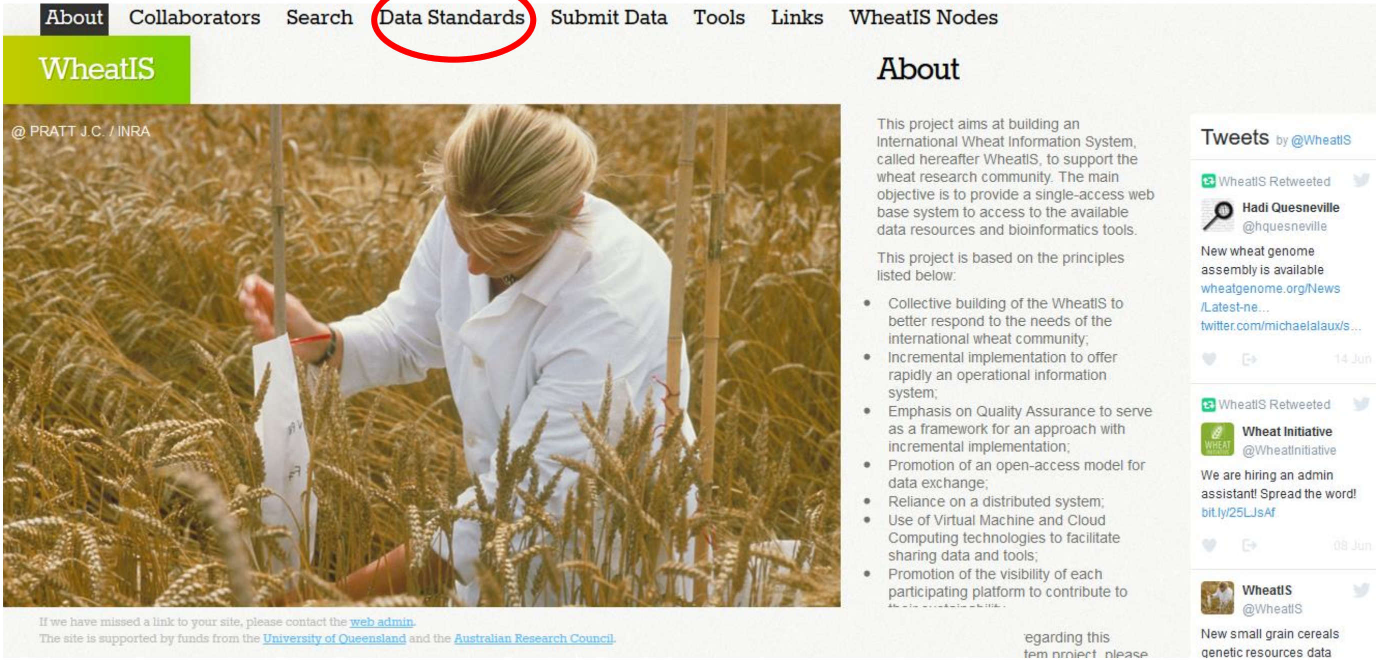
- Keep the link between the data sets and their authorship (ex. using DOI) (Findability + authorship)

Alignment on the recommendations of national and international infrastructures



Wheat Information System

http://wheatis.org/



About Collaborators Search Data Standards Submit Data Tools Links WheatIS Nodes

WheatIS

@ PRATT J.C. / INRA

About

This project aims at building an International Wheat Information System, called hereafter WheatIS, to support the wheat research community. The main objective is to provide a single-access web base system to access to the available data resources and bioinformatics tools.

This project is based on the principles listed below:

- Collective building of the WheatIS to better respond to the needs of the international wheat community;
- Incremental implementation to offer rapidly an operational information system;
- Emphasis on Quality Assurance to serve as a framework for an approach with incremental implementation;
- Promotion of an open-access model for data exchange;
- Reliance on a distributed system;
- Use of Virtual Machine and Cloud Computing technologies to facilitate sharing data and tools;
- Promotion of the visibility of each participating platform to contribute to their sustainability.

egarding this
tem project, please

Tweets by @WheatIS

WheatIS Retweeted
Hadi Quesneville
@hquesneville
New wheat genome assembly is available wheatgenome.org/News/Latest-ne... twitter.com/michaelalaux/s...
14 Jun

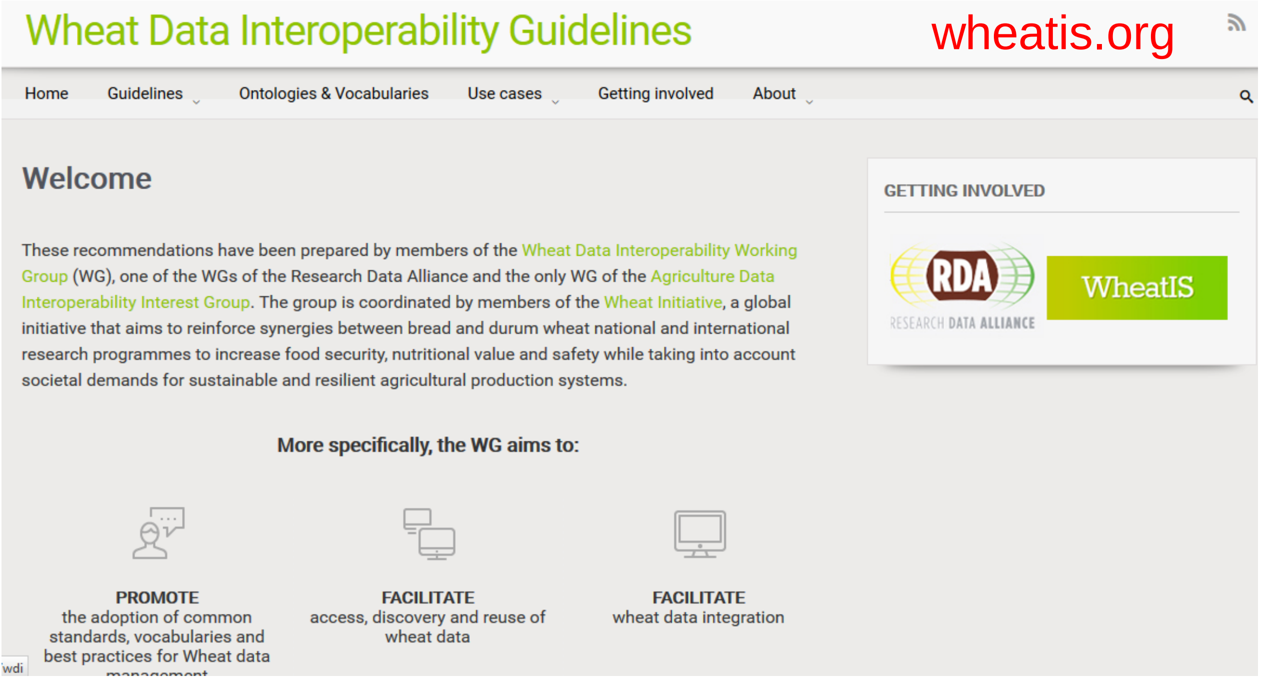
WheatIS Retweeted
Wheat Initiative
@WheatInitiative
We are hiring an admin assistant! Spread the word! bit.ly/25LJsAf
08 Jun

WheatIS
@WheatIS
New small grain cereals genetic resources data

If we have missed a link to your site, please contact the web admin.
The site is supported by funds from the University of Queensland and the Australian Research Council.

Wheat Initiative

Interoperabity, Reusability: standards, identifiers



Wheat Data Interoperability Guidelines

wheatis.org

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

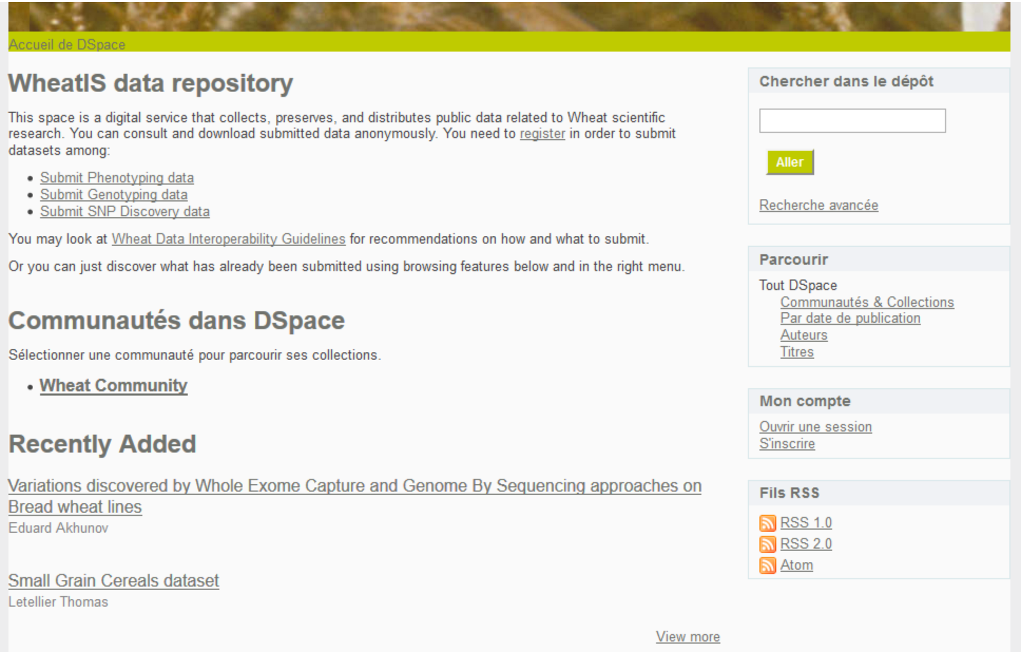
GETTING INVOLVED

RDA RESEARCH DATA ALLIANCE

WheatIS

Globally agreed good practices for data standardization and identifiers choices

Accessibility: central file repository



Accueil de DSpace

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.

Communautés dans DSpace

Sélectionner une communauté pour parcourir ses 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

View more

Chercher dans le dépôt

Aller

Recherche avancée

Parcourir

Tout DSpace
- Communautés & Collections
- Par date de publication
- Auteurs
- Titres

Mon compte

Ouvrir une session
S'inscrire

Fils RSS

RSS 1.0
RSS 2.0
Atom

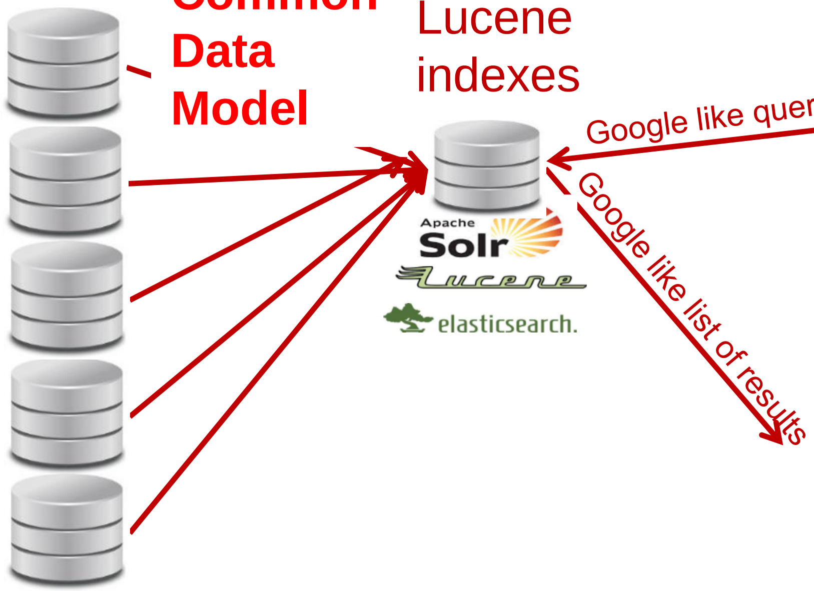
wheatis.org

Raw data + metadata

Metadata indexed and searchable under the central portal

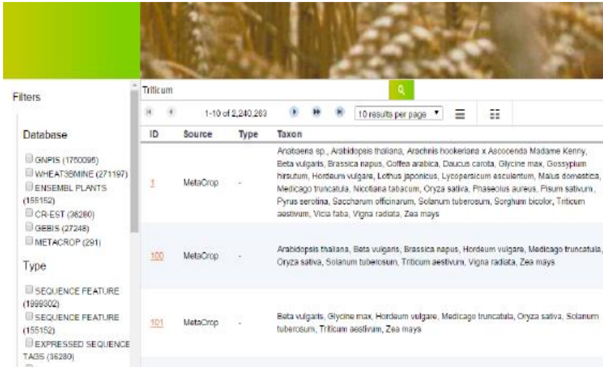
WheatIS: Data discovery through a common portal

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



User web interface

http://www.wheatis.org/



Triticum

1-10 of 2,240,263 10 results per page

Database

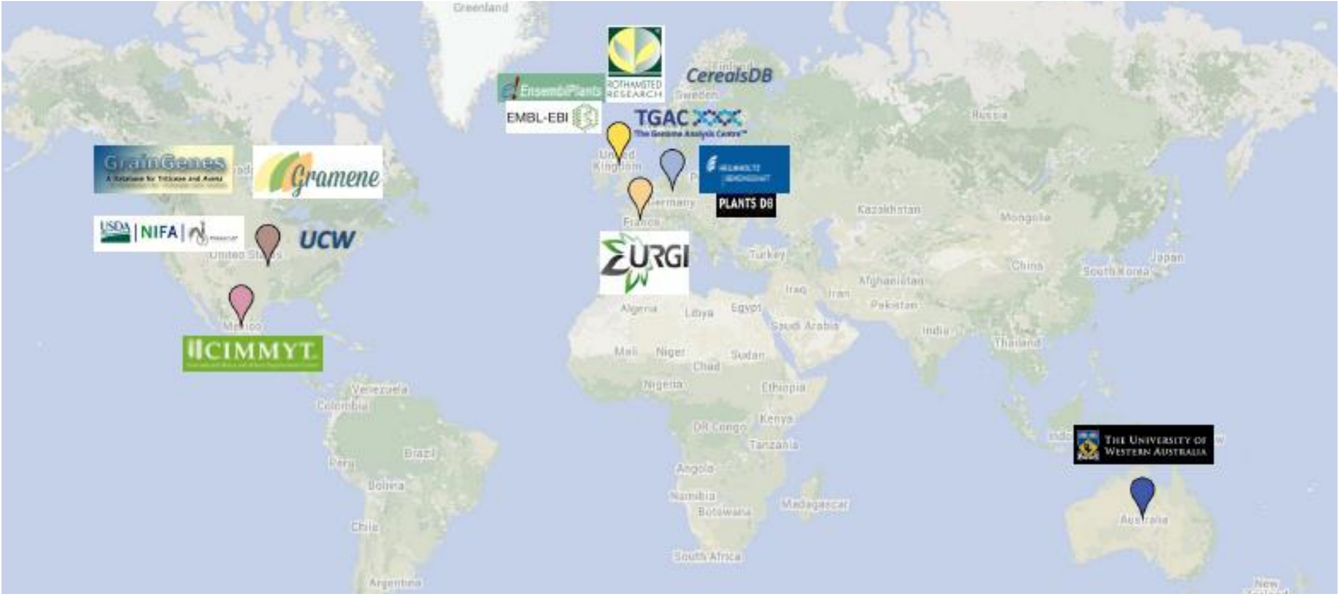
□ GNPIS (1760096)
□ WHEAT3BMINE (271197)
□ ENSEMBL PLANTS (155152)
□ CR-EST (36280)
□ GEBIS (27248)
□ METACROP (291)

Type

□ SEQUENCE FEATURE (1999302)
□ SEQUENCE FEATURE (155152)
□ EXPRESSED SEQUENCE TAGS (36280)

ID	Source	Type	Taxon
1	MetaCrop	-	Anabaena sp., Arabidopsis thaliana, Arachnis hookeriana x Ascocenda Madame Kenny, Beta vulgaris, Brassica napus, Coffea arabica, Daucus carota, Glycine max, Gossypium hirsutum, Hordeum vulgare, Lothus japonicus, Lycopersicum esculentum, Malus domestica, Medicago truncatula, Nicotiana tabacum, Oryza sativa, Phaseolus aureus, Pisum sativum, Pyrus serotina, Saccharum officinarum, Solanum tuberosum, Sorghum bicolor, Triticum aestivum, Vicia faba, Vigna radiata, Zea mays
100	MetaCrop	-	Arabidopsis thaliana, Beta vulgaris, Brassica napus, Hordeum vulgare, Medicago truncatula, Oryza sativa, Solanum tuberosum, Triticum aestivum, Vigna radiata, Zea mays
101	MetaCrop	-	Beta vulgaris, Glycine max, Hordeum vulgare, Medicago truncatula, Oryza sativa, Solanum tuberosum, Triticum aestivum, Zea mays

WheatIS nodes #12



Challenges:

- Synchronize technical updates of the infrastructure
- Synchronize improvements of the data model
- Searching with increasingly natural language (e.g. for traits)

Conclusion

Making data FAIR is a lot about community management (within and between):

- Developpers
- Specialists of ontologies and standards
- Data managers
- Biologists (data producers)

(Global) Infrastructure projects

Crop projects

Need for identification and long term maintainance of:

- Searchable central repositories of standards and ontologies for agriculture (e.g. agroportal.lirmm.fr, biosharing.org)
- FAIR tools for data managers/developers for automatic formatting or format validation (BioSchemas, …)

Aknowledgements

transPLANT

excelerate

openMINTED
Open Mining Infrastructure for Text & Data

RDA
RESEARCH DATA ALLIANCE

Wheat Initiative

Bioversity International

BrAPI

ANR

INVESTISSEMENTS D'AVENIR

INRA
SCIENCE & IMPACT