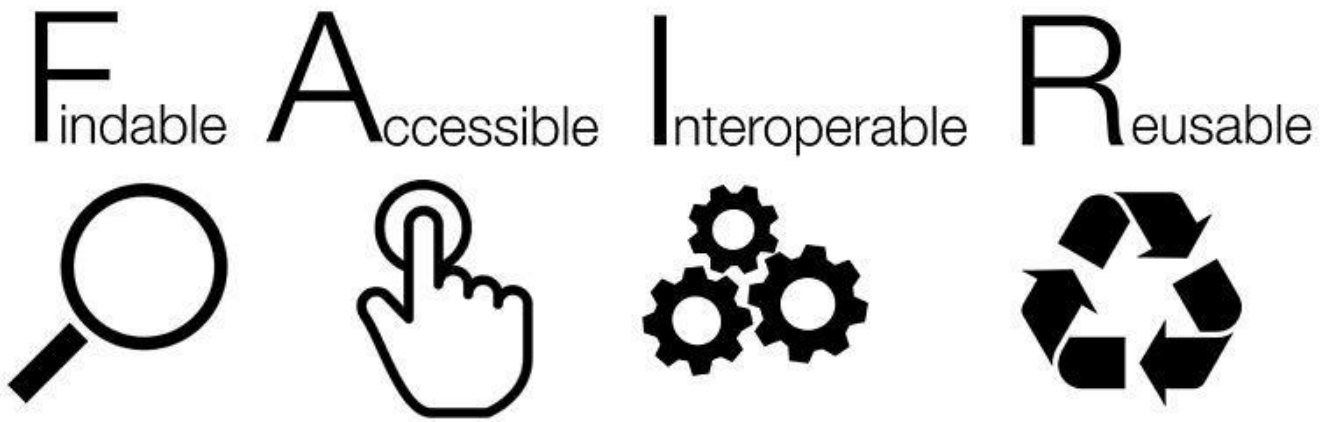
Rendre ses données "FAIR" - Contexte et Infrastructures

Anne-Francoise Adam-Blondon

Rendre ses données « FAIR »- contexte et infrastructures

Anne-Françoise Adam-Blondon
Colloque EPGV, 4 octobre 2018



Contexte actuel des sciences de la vie

Open science … open access, open data

"The days of keeping our research results to ourselves are over. There is far more to gain from sharing data and letting others access and analyse that data.

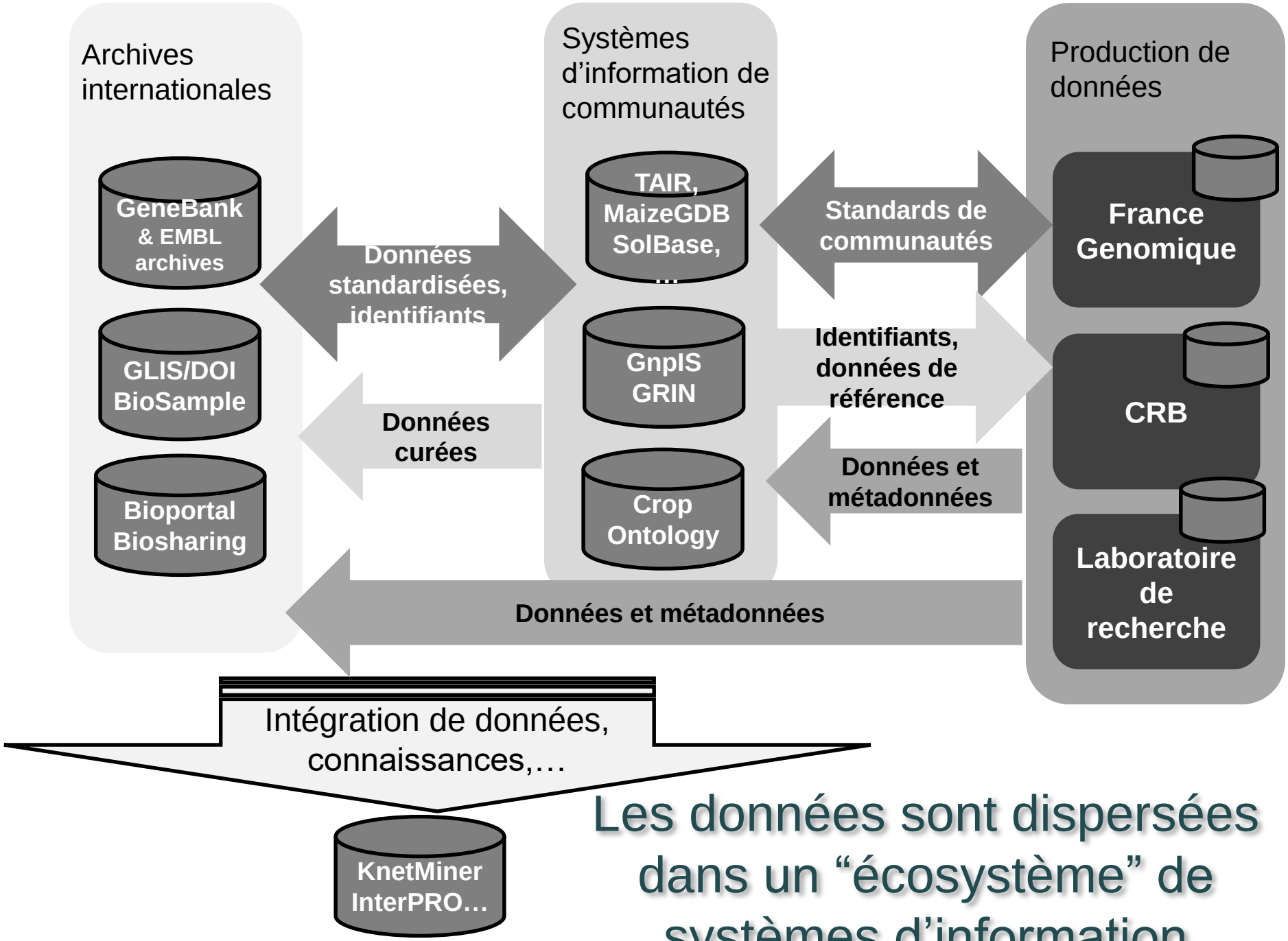
For example, if sharing big data reveals that a certain kind of cancer activates a particular molecular pathway in most cases and it turns out that there is already a drug approved and available to block the activation of that molecular pathway, clinical trials can begin almost immediately. Saving time, money and lives.

Or if scientists want to monitor the effects of climate change on local ecosystems, they can use Open Science to engage citizen reporting, and rapidly multiply the data at their disposal.

To make the most of Open Science opportunities for Europe, I plan to focus on open data, open access and research integrity over the course of my mandate."

Commissioner Carlos Moedas,
"European research and innovation for global challenges", Lund, 4 December 2015

http://bookshop.europa.eu/en/open-innovation-open-science-open-to-the-world-pbKI0416263/



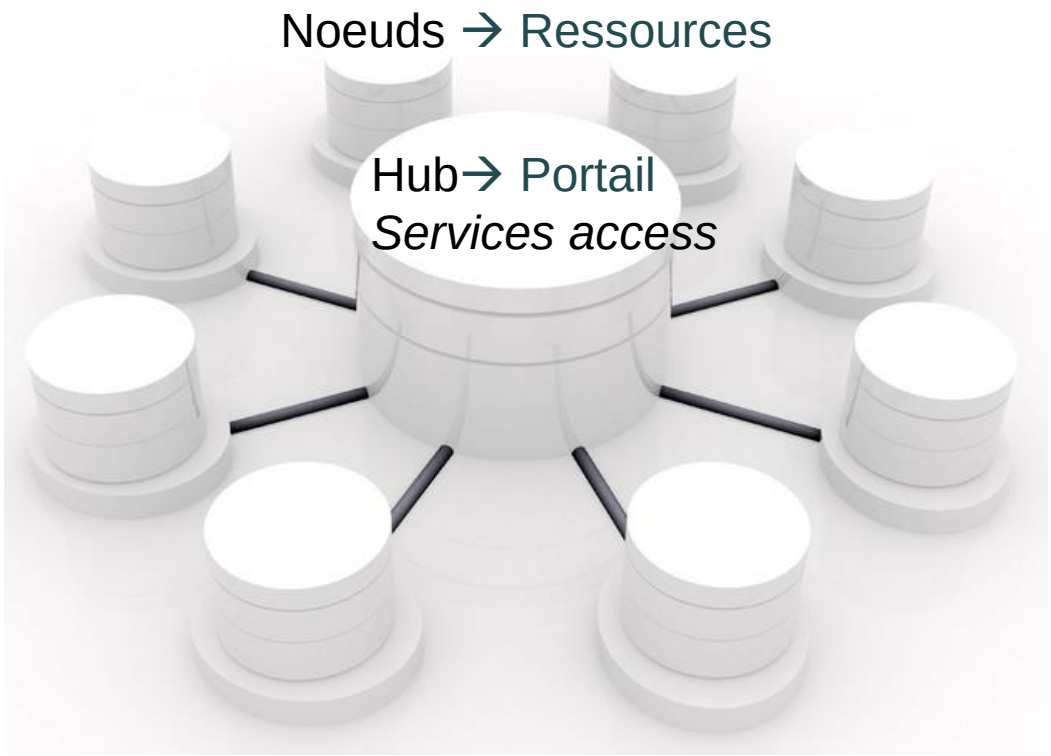
Les données sont dispersées dans un "écosystème" de systèmes d'information

Les données sont dispersées dans de nombreux systèmes d'information

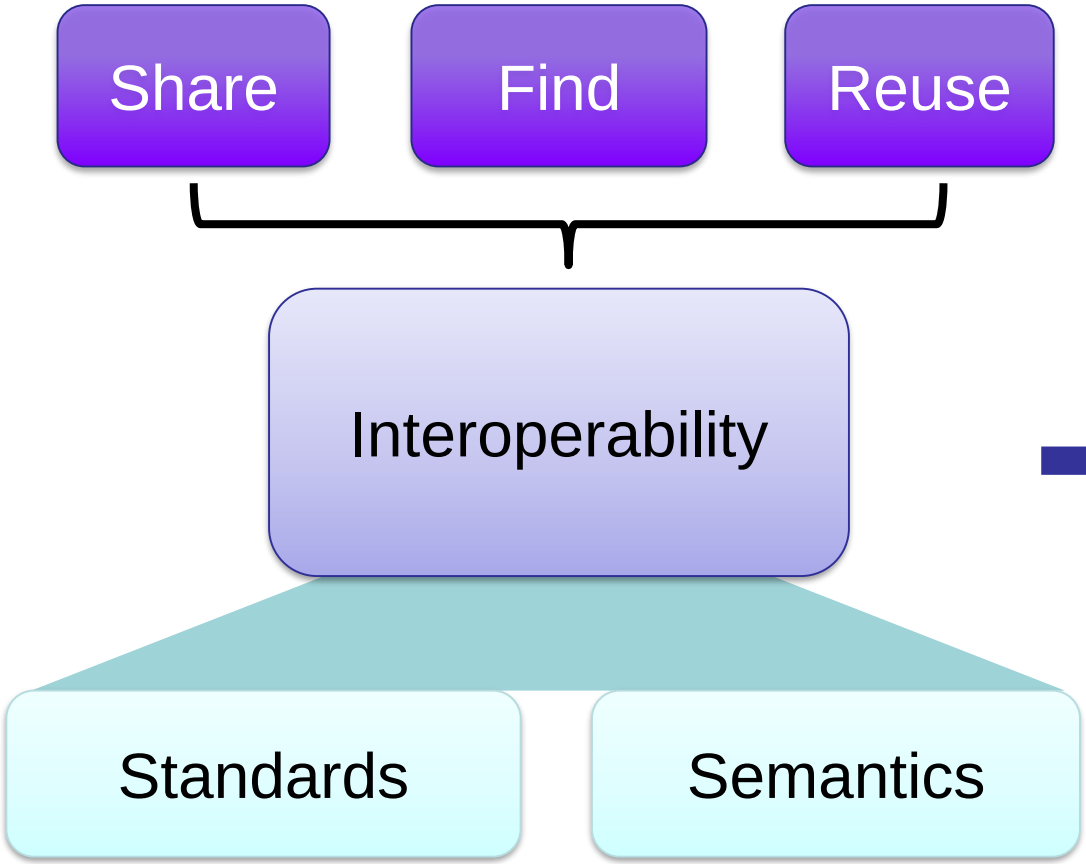
=> Aller vers des fédérations de systèmes d'information interopérables qui permettent aux chercheurs de les trouver

Organisation de ces fédérations

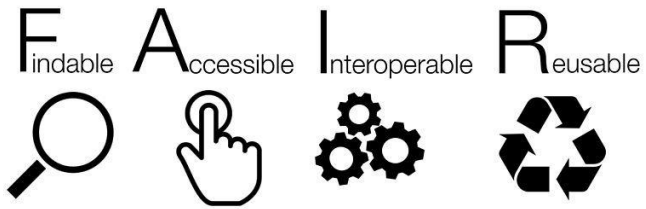
- Un réseau de nœuds (stables)
- Un portail central offrant des services (a minima: trouver et accéder aux données)



Enjeux clefs des E-infrastructures



Recommandations développées par un collectif de bibliothécaires, établissement d'enseignement et de recherche pour l'Open Data en sciences

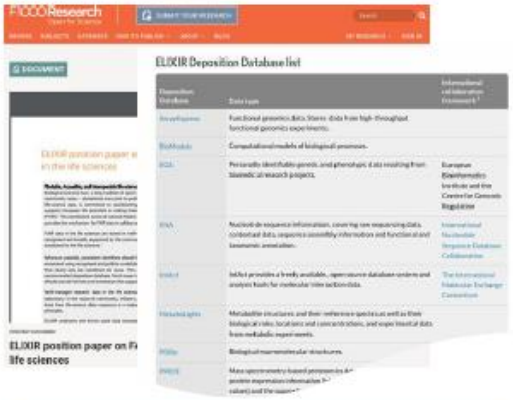


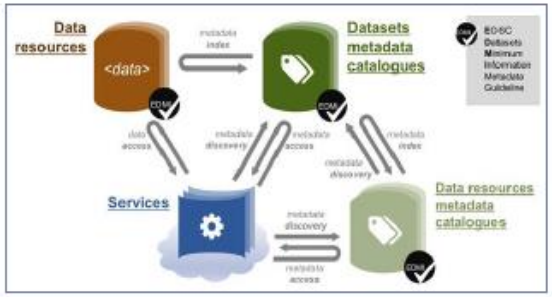
Wilkinson et al (2016) SCIENTIFIC DATA, 3:160018, DOI: 10.1038/sdata.2016.18

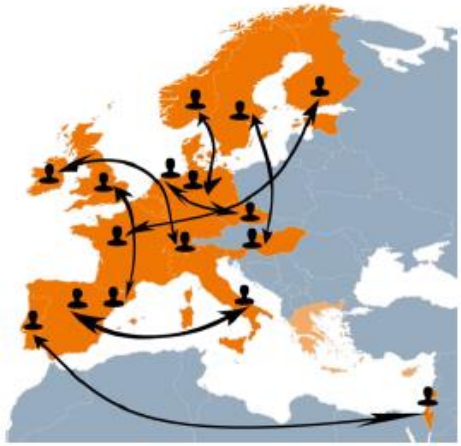
Organisation des e-Infrastructures européennes

- European Open Science Cloud (EOSC)
- European Infrastructure of Bioinformatics for Life-sciences: ELIXIR

ELIXIR mission and strategy aligns with EOSC – implement life-science foundation in EOSC-Life







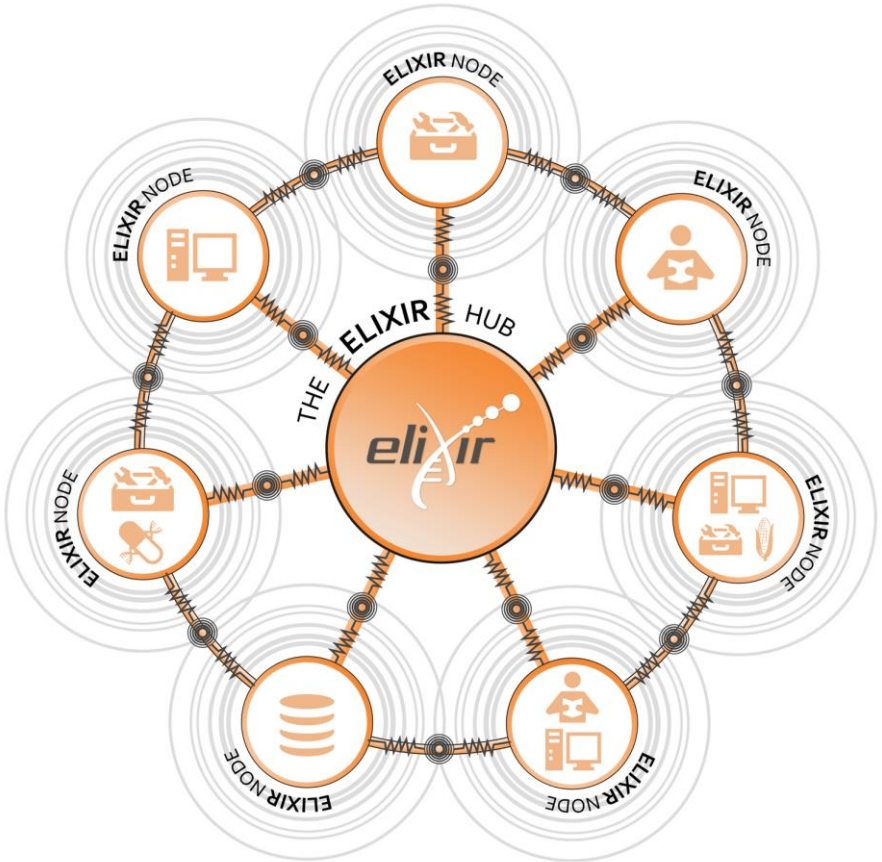
FAIR data management in the life sciences
10.7490/f1000research.1114985.1

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

Compatible Cloud / Workflows / Reference Data Set Distribution Service

D'après N. Blomberg

ELIXIR, a distributed pan-European infrastructure



ELIXIR: The French Node

The French bioinformatics community is currently setting up a national infrastructure of services in Bioinformatics (Institut Français de Bioinformatique, IFB). IFB serves as a unique entry point for requests of service from the Life Science community and is in charge of coordinating and structuring the activities of the regional bioinformatics platforms. IFB is the French ELIXIR node.

Collaborating organisations

IFB consists of a national hub, IFB-core, and more than 20 bioinformatics platforms organized in six regional centres. These PFs belong to the main French research organisations:

- CNRS: National Centre for Scientific Research
- CEA: Alternative Energies and Atomic Energy Commission
- INRA: National Institute for Agriculture Research
- INRIA: National Institute for Computer Science and Control
- INSERM: National Institute for Health and Medical Research
- Universities
- Pasteur Institute (research foundation)
- Curie Institute (research foundation)

IFB comprises more than 110 FTE and 70 FTC researchers and engineers. This represents 25% of the French Bioinformatics community that is involved in provision of service for the Life Sciences.

Services

To provide added-value to various data produced routinely by biological platforms and newly created national infrastructures in sequencing, genotyping, proteomics, metabolomics, etc. the French node will focus on the development of **services for integrative curation** of biological data.

Keywords of IFB contributions are thus **interoperability** (technological and semantic) and **integration**.

IFB will provide access to well-curated databases, tools and services in:

- three biological domains:
  - D-1: **Microbial world**
  - D-2: **Plants**
  - D-3: **Health**
- two transversal fields:
  - A-1: **Phylogeny and classification**
  - A-2: **Protein sequence and structures**
- and two transversal activities:
  - T-1: **Management and analysis of metagenomic data**
  - T-2: **Deployment of an academic cloud dedicated to analyses of Life Science data**

Building on the training and education programs already proposed by the regional PFs, IFB will endeavour to coordinate and compile a **national training and education program** based on **E-learning** technologies.

Contact

- 23 pays impliqués
- Le Nœud Français est porté par l'IFB

D'après N. Blomberg

Objectifs d'ELIXIR

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

Marine metagenomics

Human data

Plants

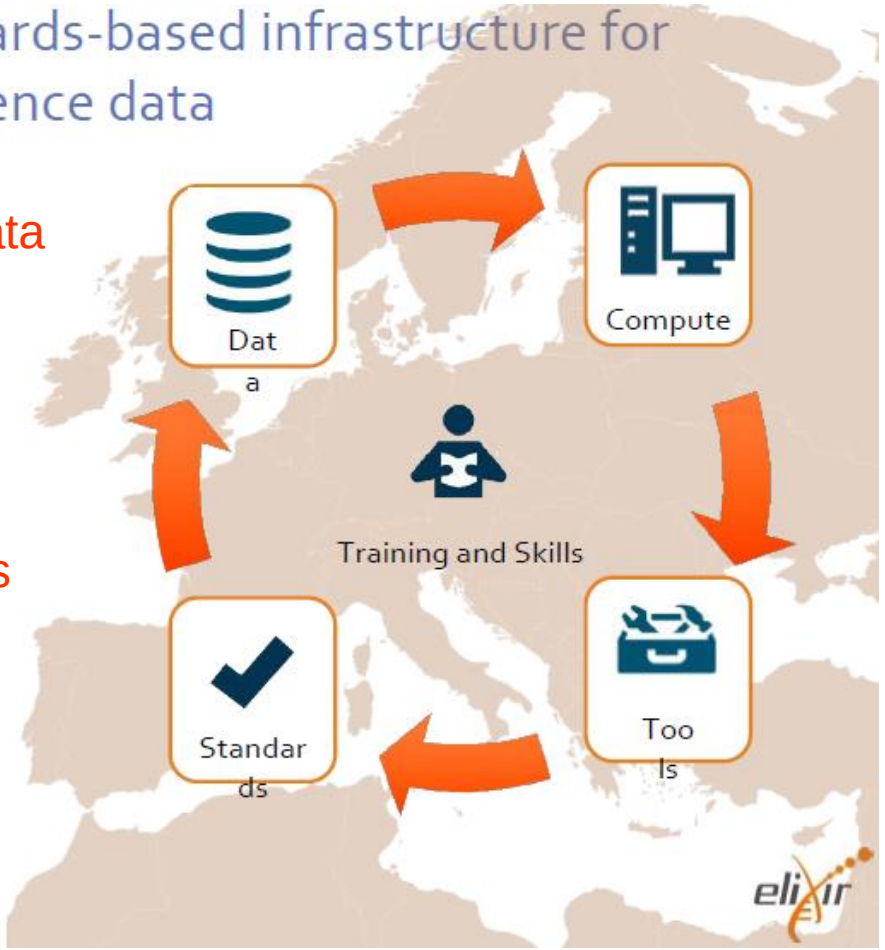
Metabolomics

Proteomics

Galaxy

Rare diseases

...delivered in partnership with research communities

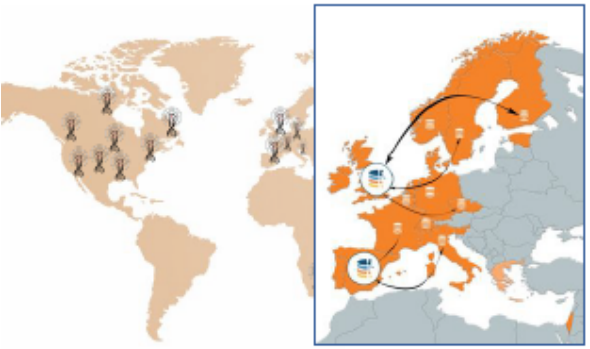


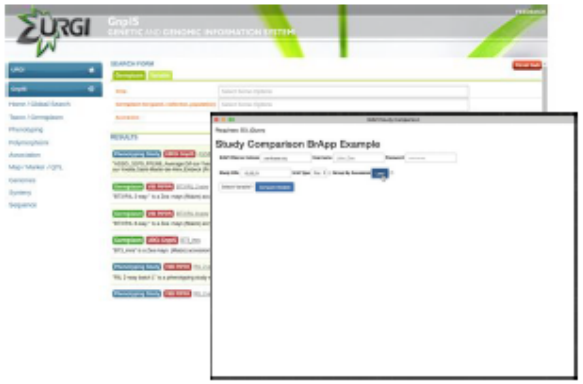
Communauté Plante d'ELIXIR coordonnée par C. Pommier (URGI) et C. Miguel (IBET, PT)

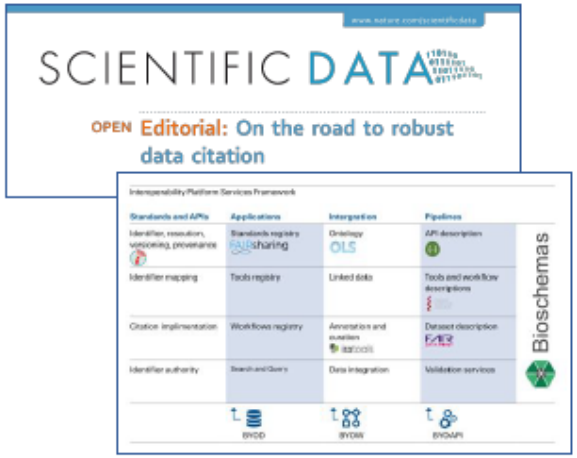
D'après N. Blomberg

Objectifs d'ELIXIR

FAIR publication, sharing and reuse of data require global standards







Beacons/Local EGA – Federated Genomics Discovery & Query services

Standards and services for federated plant phenotyping data

Global collaborations for standards infrastructure

Global Alliance for Genomics & Health

elixir

excelerate

EMPHASIS

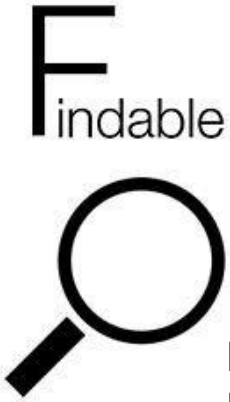
transPLANT

BrAPI

miappe

elixir

D'après N. Blomberg



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.

Producteurs de données:

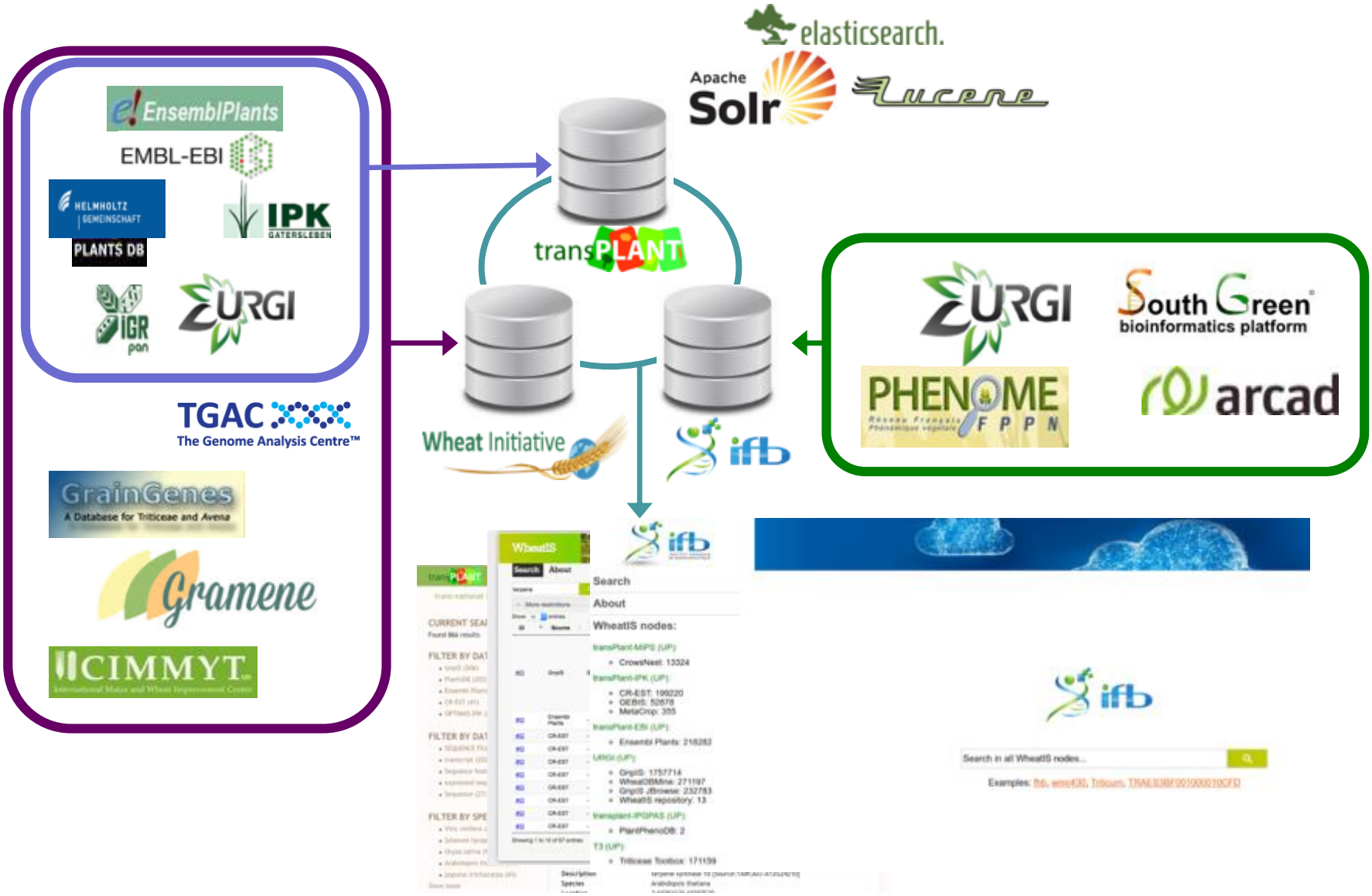
- dépôt de données dans des archives avec des métadonnées riches

Informaticiens/curateurs de données:

- développement de services/catalogues de PUID ou persistent Identifiers (ex: DOI pour les accessions, ORCID ID pour les personnes, …)
- développement de portails qui indexent les métadonnées et d'outils de recherche associés

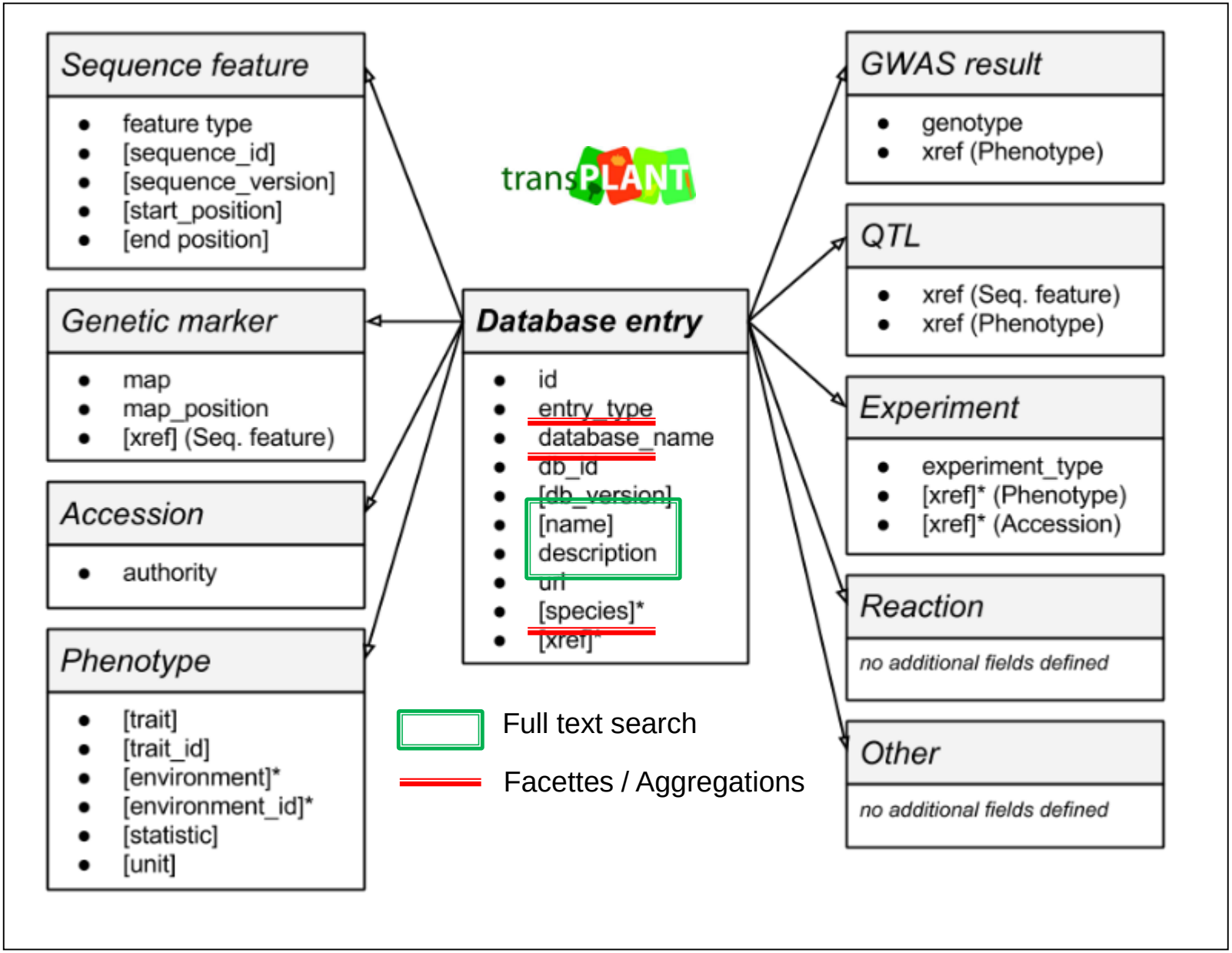
Modèles de données et sémantique pour améliorer la performance des recherches

Plusieurs preuves de concept de portail de fédérations « plantes »

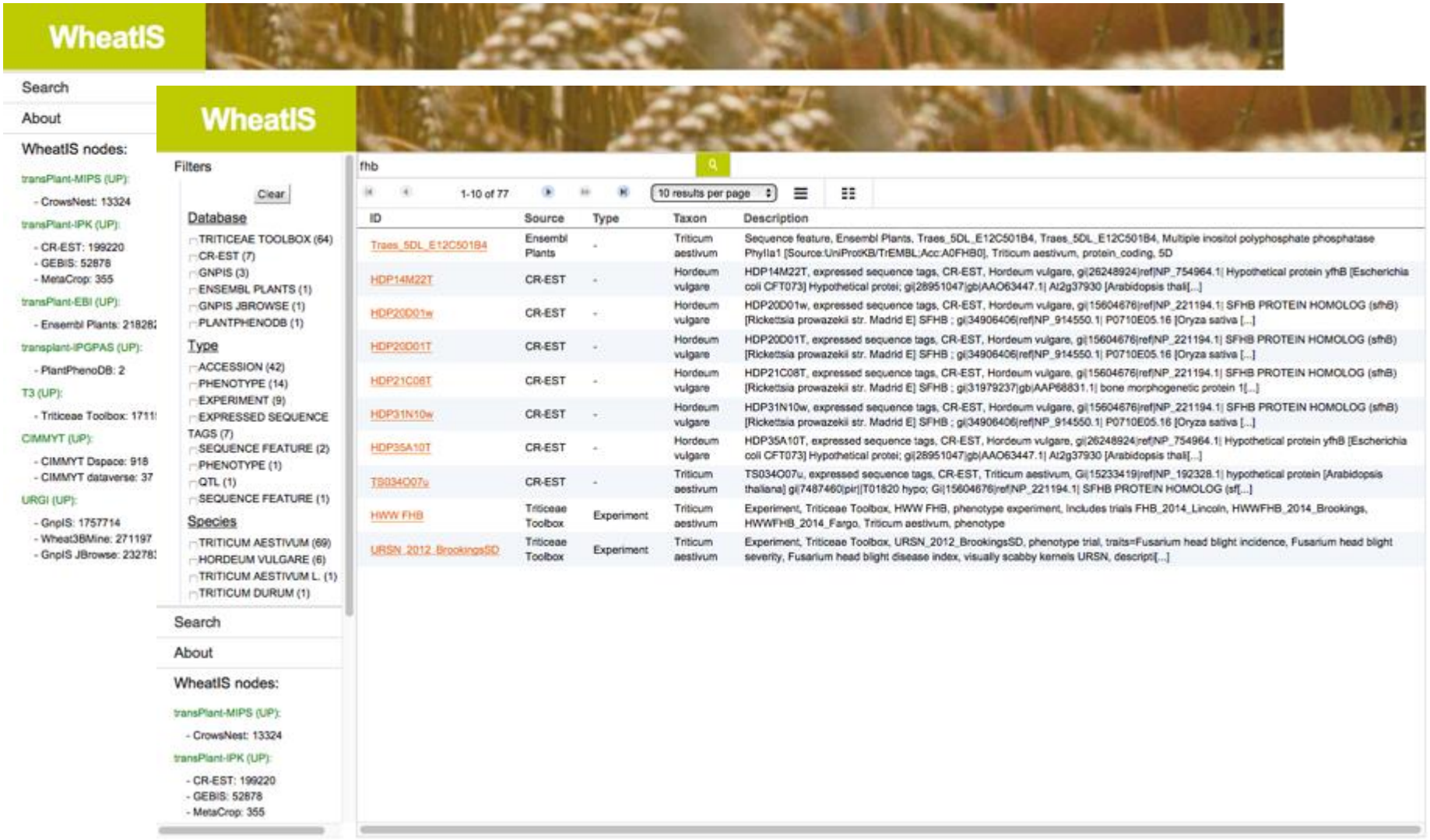


Ex: fédération internationale WheatIS

Basé sur un modèle de données très gros grain



Un portail: wheatis.org



Accessible

A1 (meta)data are **retrievable** by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

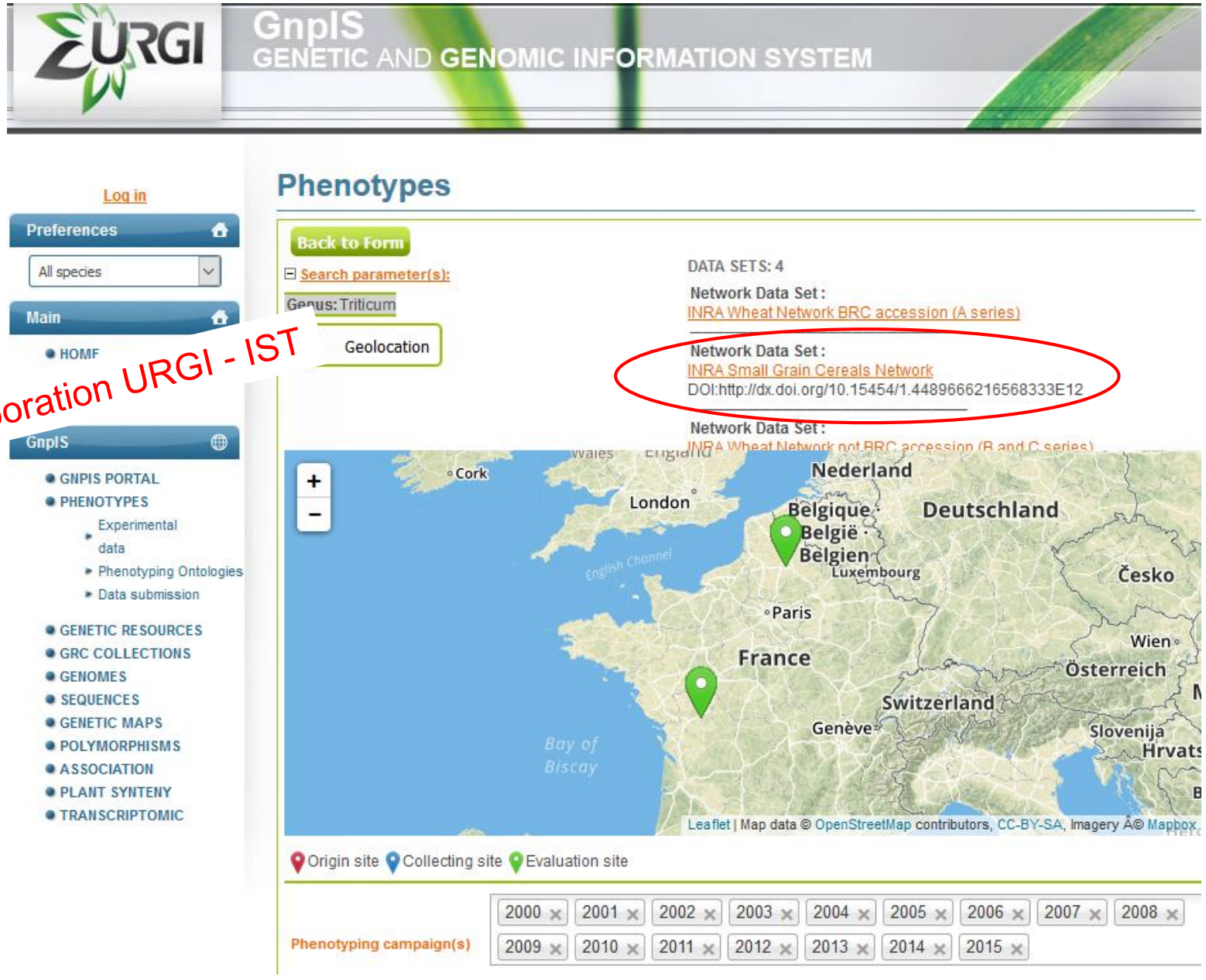
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 **metadata are accessible**, even when the data are no longer available.

- **Producteurs de données:** dépôt de données et métadonnées sous PUID/DOI dans des archives/bases de données accessibles par le web (stables dans le temps) + licence d'accès définie dans les DOI, base de données

- **Informaticiens:** développement de web services (= Application Programing Interface: API)

Ex: 16 ans d'expérimentations du réseau Céréales à Pailles



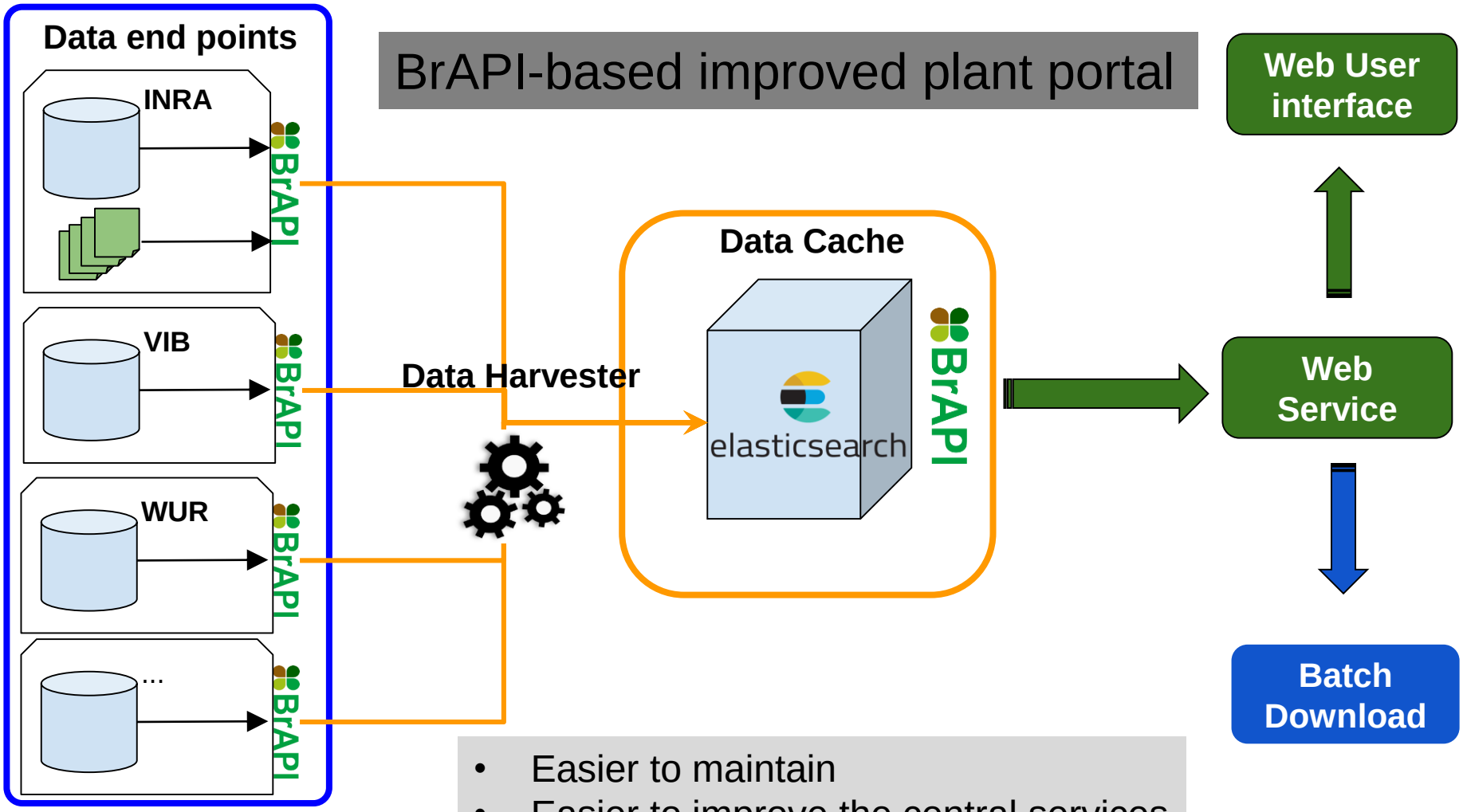
Breeding API initiative

http://www.brapi.org/

- Bill & Melinda Gates Foundation
- CassavaBase
- T3
- IBP
- JHI
- Bioversity
- CIRAD
- INRA
- IRRI
- GOBII
- Wageningen WUR
- CIP
- DaRT
- Cornell
- iPlant
- …

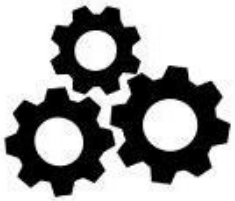
- Développement d'une **API (web service) standard** pour les données dans le champ de la génétique et amélioration des plantes
  - Materiel génétique, expériences de phénotypage, de génotypage, ….
  - Alignement de l'API BrAPI avec les standards pour les données RG (MCPD) et pour les expériences de phénotypage (MIAPPE)

Federation of Plant Information systems



- Easier to maintain
- Easier to improve the central services

Interoperable



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**.

Animation de différentes communautés (et de leurs interactions)

- Développeurs
- Spécialistes des ontologies et standards
- Curateurs de données
- Producteurs de données

Infrastructures (internationales)

Projets focalisés espèces, objets de recherche

En clarifiant

- ce qui est de la responsabilité de chacun dans le processus
- les concepts pivots pour faire de l'interopérabilité
- les standards existants

Interopérabilité des données

Crop Ontology
Variable=trait + method + scale

Identification : MultiCrop
Passport Data
Standard

Phenotype 1 = measurement on a cultivar in an environment-GPS1-time1

Phenotype 2 = measurement on a cultivar in an environment-GPS2-time2

Genotype = observed marker's alleles on a cultivar

Climate 1 = climatic data at GPS1-time1

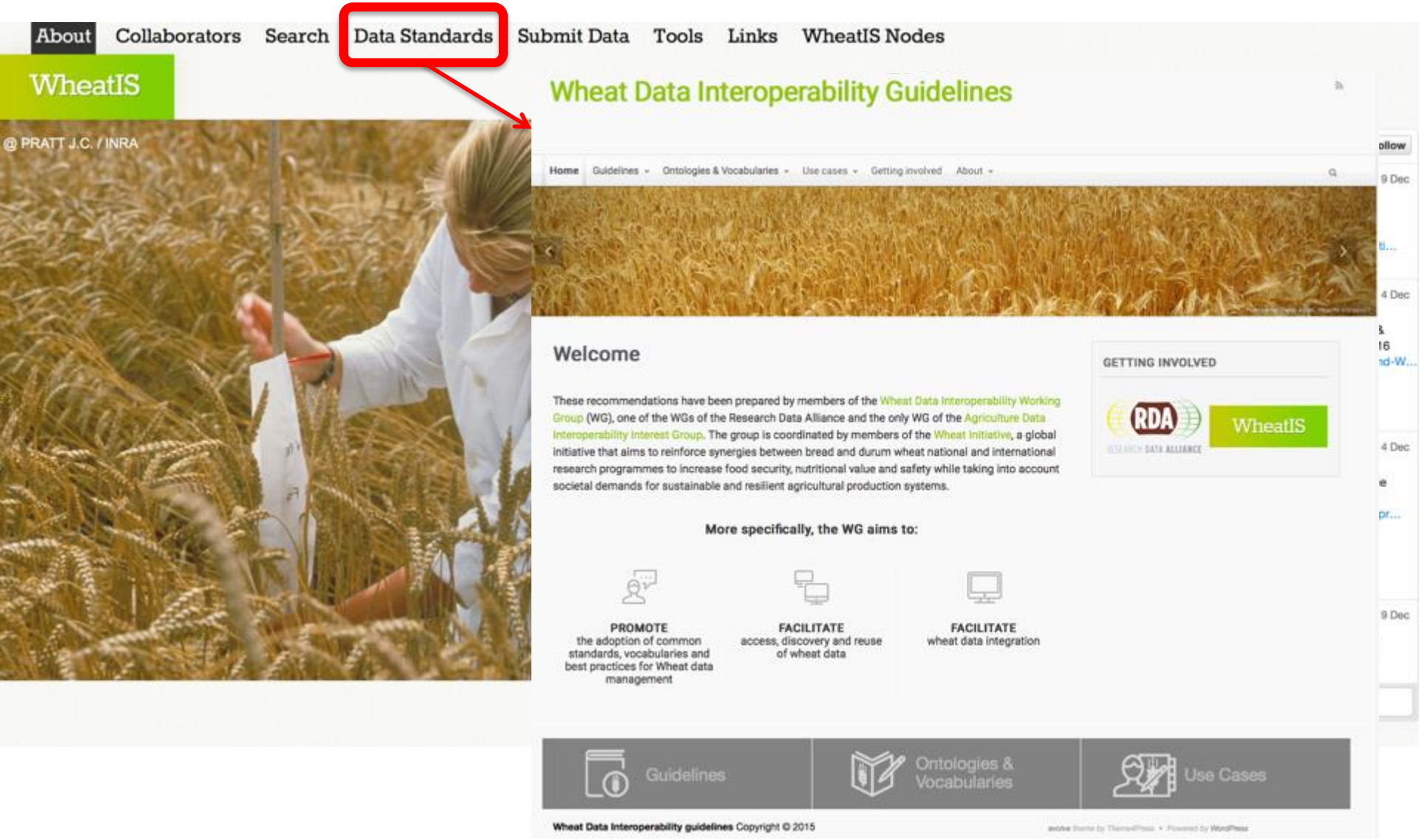
Inspire EU directive

Différents niveaux de standardisation possible

- *Check list* d'informations décrivant le comment, par qui et pourquoi du jeu de données : standard de métadonnées
- Dictionnaires de vocabulaires contrôlés, ontologies, échelles standards: Crop Ontology, Gene Ontology, BBCH scales…
- Identifiant uniques (idéalement persistents): gene ID, accessions ID, Trait ID, pubmed ID, DOI,…
- Formats de fichiers: VCF, IsaTab, GnpIS excel submission format, …

Domaine foisonnant et complexe: nécessité de développer des guides, catalogues et formations ciblées sur les différentes communautés

wheatis.org


About
Collaborators
Search
Data Standards
Submit Data
Tools
Links
WheatIS Nodes
WheatIS
@ 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
RDA
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 Copyright © 2015

Reusable

Reusable

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

Producteurs de données:

- Contribution à la définition des standards de métadonnées cohérents avec le domaine
- Publication des jeux de données sous DOI permet de spécifier l'ensemble des contributeurs et leurs rôles et les droits et conditions de réutilisation des données

Informaticiens/ bibliothécaires:

- Développement d'outils d'aide à la publication des jeux de données

Developpement d'un standard de métadonnées pour les données de phénotypage

- MIAPPE: Minimum Information About Phenotyping Experiment
- Developpé et maintenu par une communauté internationale: maintained by an international community interested in plant phenotyping: large community of breeders and biologists, European infrastructure for Plant Phenotyping (EPPN/EMPHASIS), European infrastructure of Bioinformatics (ELIXIR), Planteome, Excellence in Breeding Plateform…
- www.miappe.org
- Steering committee avec des représentants de Emphasis, Elixir, CGIARs

Remerciements

H. Quesneville
C. Pommier
M. Alaux

E. Dzale-Yeumo
S. Aubin

Financial supports

International infrastructures/ initiatives

National and international crop projects

Pas que les plantes cultivées!

Merci!

Catalogue (FAIR et durable) de recommandations et standards

Find

Recommendations

Standards and/or databases recommended by journal or funder data policies.

Discover

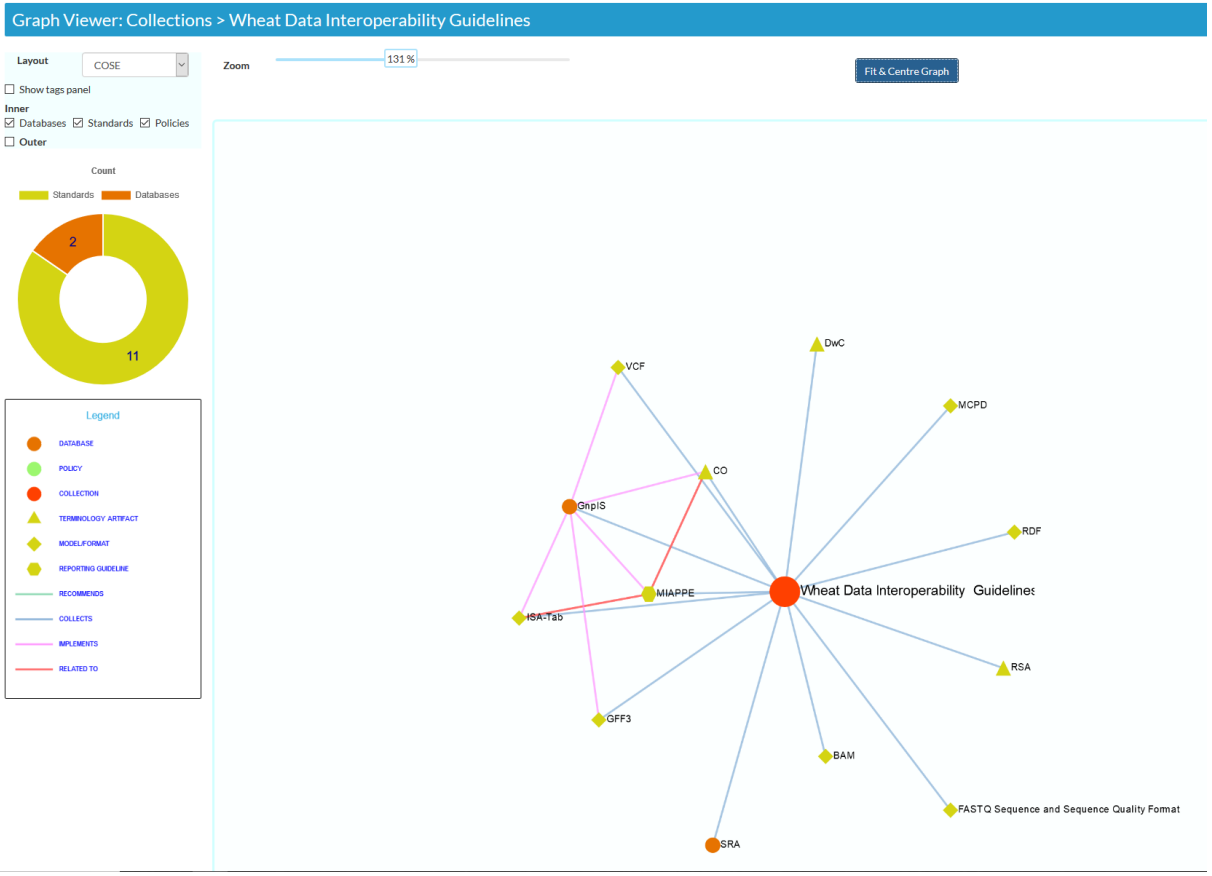
Collections

Standards and/or databases grouped by domain, species or organization.

Learn

Educational

About standards, their use in databases and policies, and how we can help you.



elixir UNITED KINGDOM

EMBL Australia Bioinformatics Resource

RDA RESEARCH DATA ALLIANCE

FORCE11

F1000 FACULTY OF 1000

DRYAD

PLOS

nature publishing group npg

EMBOpress

FAIRsharing.org standards, databases, policies