Exemples d'outils et services pour la gestion et le partage des données à l'Inra

Cyril Pommier, Windpouire Esther Dzale Yeumo

Open Science INRA

Outils et services pour la gestion et le partage des données

E. Dzalé Yeumo & C. Pommier

14/ 12/ 2017

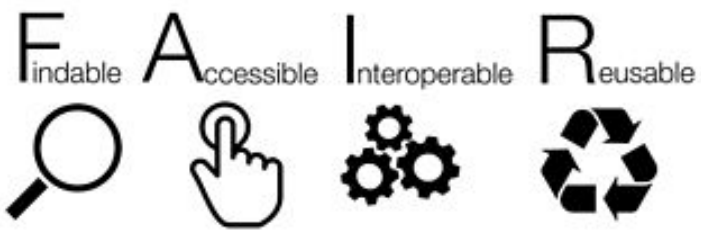
SOMMAIRE

- Cycle de vie de la donnée
- Les services et outils INRA
- Use case séries temporelles
- Use case céréales à pailles

_01
Introduction

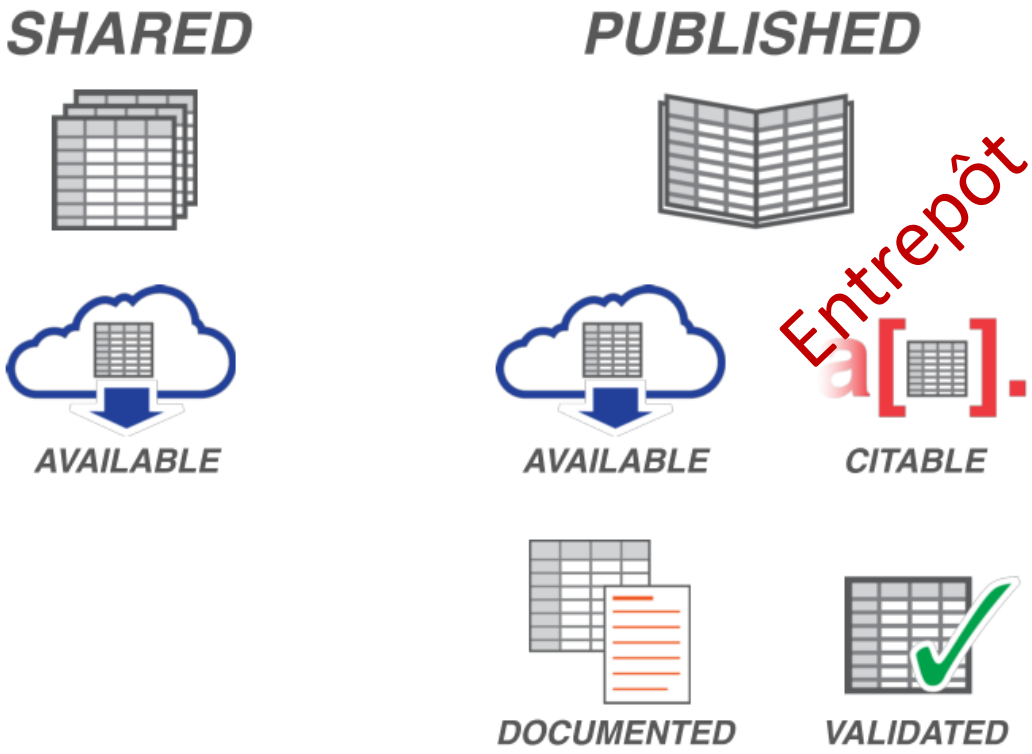
Les enjeux de l'open data

- Bonnes pratiques de gestion : produire des données FAIR



- Accessibilité des données pour une meilleure transparence et reproductibilité de la recherche
- Réutilisation des données
- Traçabilité
- Valorisation des données
- Reconnaissance des auteurs / contributeurs

Données partagées VS données publiées



F1000Research

Kratz J and Strasser C 2014 Data publication consensus and controversies [v2; ref status: indexed, http://f1000r.es/3hi] *F1000Research* 2014, **3**:94
10.12688/f1000research.3979.2

Entrepôt de données

Rôle de stockage, diffusion, archivage des données

Cycle projet scientifique

Pendant

- **Stocker les données** dans des supports différents

Pendant et à la fin

- Déposer les données dans un entrepôt pour **diffusion avec objectif de partage et de publication** → ouverture des données pour le peer review (évaluation et validation par les pairs) ou pour donner accès à un public plus large

Après

- Déposer dans un entrepôt ou un centre d'**archivage** pour archivage (conservation sur le long terme)

- Offre de stockage DSI, poste de travail, serveur collectif, etc.
- Entrepôts et portails :
  - thématiques (Genbank, ICPSR, PANGAEA, ENA, etc.),
  - généralistes (Zenodo, Dryad, etc.),
  - institutionnels : portail Data Inra
- Centres d'archivage (CINES)

_02

Les services et outils INRA

Plan de Gestion des Données

- Document qui décrit la façon dont les données seront obtenues, traitées, organisées, stockées, sécurisées, préservées, partagées,… au cours et à l'issue d'un projet
- Créé au démarrage d'un projet de recherche et **mis à jour tout au long du projet**

Modèle de PGD pour l'Inra, utilisable pour tout type de projet, intégrant les exigences d'H2020 :

1. Informations sur le PGD
2. Informations sur le projet
3. Présentation succincte des données
4. Droits de propriété intellectuelle
5. Confidentialité
6. Partage des données à l'issue du projet
7. Description et organisation des données
8. Stockage et sécurité des données au cours du projet
9. Archivage et conservation des données après la fin du projet

Plan de Gestion des Données

- Un outil de saisie et d'aide à la rédaction

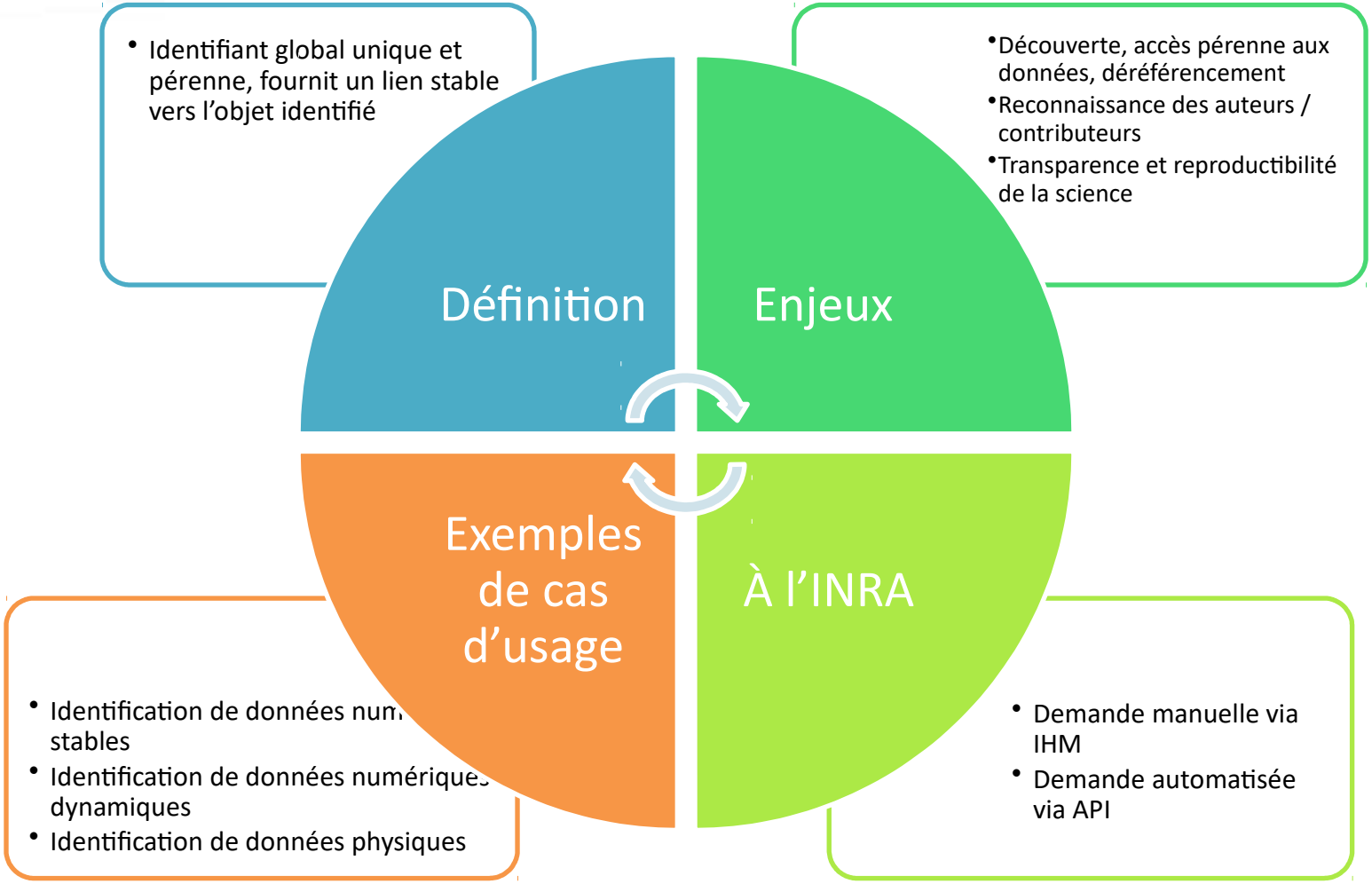
Déployé par l'Inist-CNRS pour l'Enseignement Supérieur et Recherche français https://dmp.opidor.fr/

  - Création de PGD en fonction des exigences d'un financeur ou d'une institution (aides en ligne)
  - Partage d'un PGD avec d'autres utilisateurs
  - Export de PGD dans différents formats

- Aide, conseils, relecture : digitalist@versailles.inra.fr

- Tutoriels et guides :
  site datapartage : Gérer > Plan de gestion

DOI


Définition
• Identifiant global unique et pérenne, fournit un lien stable vers l'objet identifié
Enjeux
•Découverte, accès pérenne aux données, déréférencement
•Reconnaissance des auteurs / contributeurs
•Transparence et reproductibilité de la science
Exemples de cas d'usage
• Identification de données num
stables
• Identification de données numériques
dynamiques
• Identification de données physiques
À l'INRA
• Demande manuelle via IHM
• Demande automatisée via API

DOI

Préfixe spécifique INRA

Suffixe unique

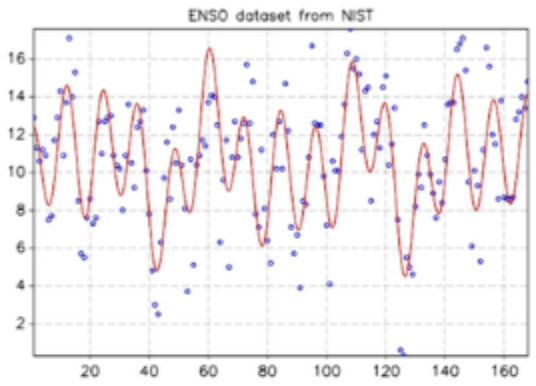
http://doi.org/10.15454/1.4768848104561313E12

Service de résolution de DOI fourni par l'agence d'enregistrement **DataCite** : permet à l'utilisateur d'aller vers la landing page par un simple clic

Nom de DOI : Identifiant unique et pérenne, enregistré auprès de DataCite avec
- l'URL de l'objet ou d'une landing page,
- Un fichier de métadonnées

DOI

1. Take a dataset



2. Describe it

- Title
- Authors
- Year
- Description
- And others…

3. Assign a DOI

10.1234/exampledata

4. Reuse and reference!

ATLAS Collaboration, "Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC: $H \to \gamma\gamma$,"
http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44

☑ Unique ☑ Persistent

5. Enjoy the benefits

Findability	Track citations
Reusability	Measure impact

Cruse, P. (2016). Moving Research Forward with persistent identifiers and services. Paper presented at the Dataverse Community Meeting. http://projects.iq.harvard.edu/files/dcm2016/files/dataverse_and_datacite_july_2016_final_clean.pdf

Data Inra

Stocker

- Déposer directement dans le portail → les données sont stockées dans un entrepôt dédié
- Déposer dans un autre entrepôt interne ou externe

Documenter

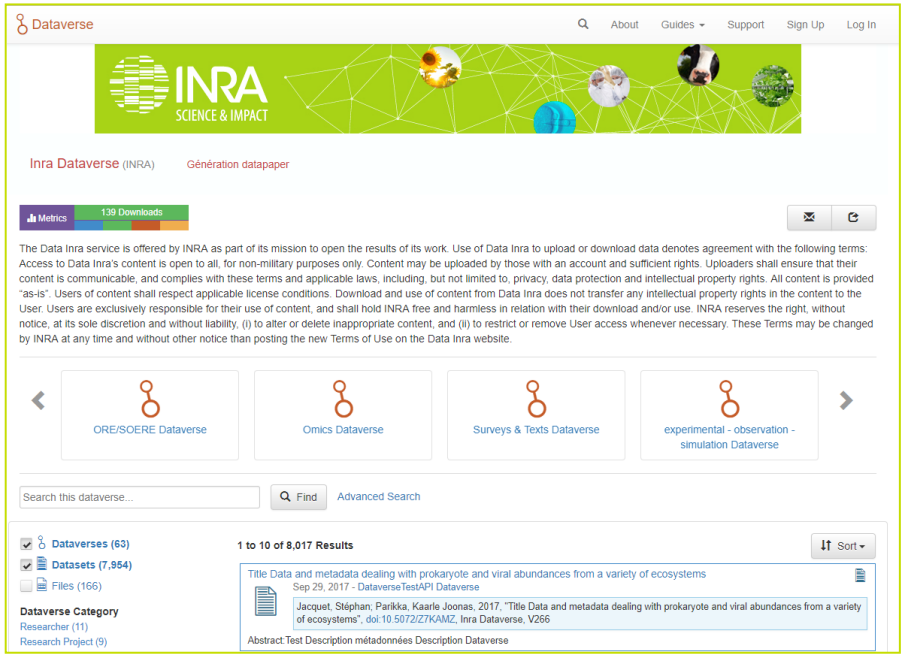
- Métadonnées standards (Dublin Core, DataCite, DDI, ISA-TAB)
- Tout document utile à la compréhension et la réutilisation des données (ex: readme)

Publier

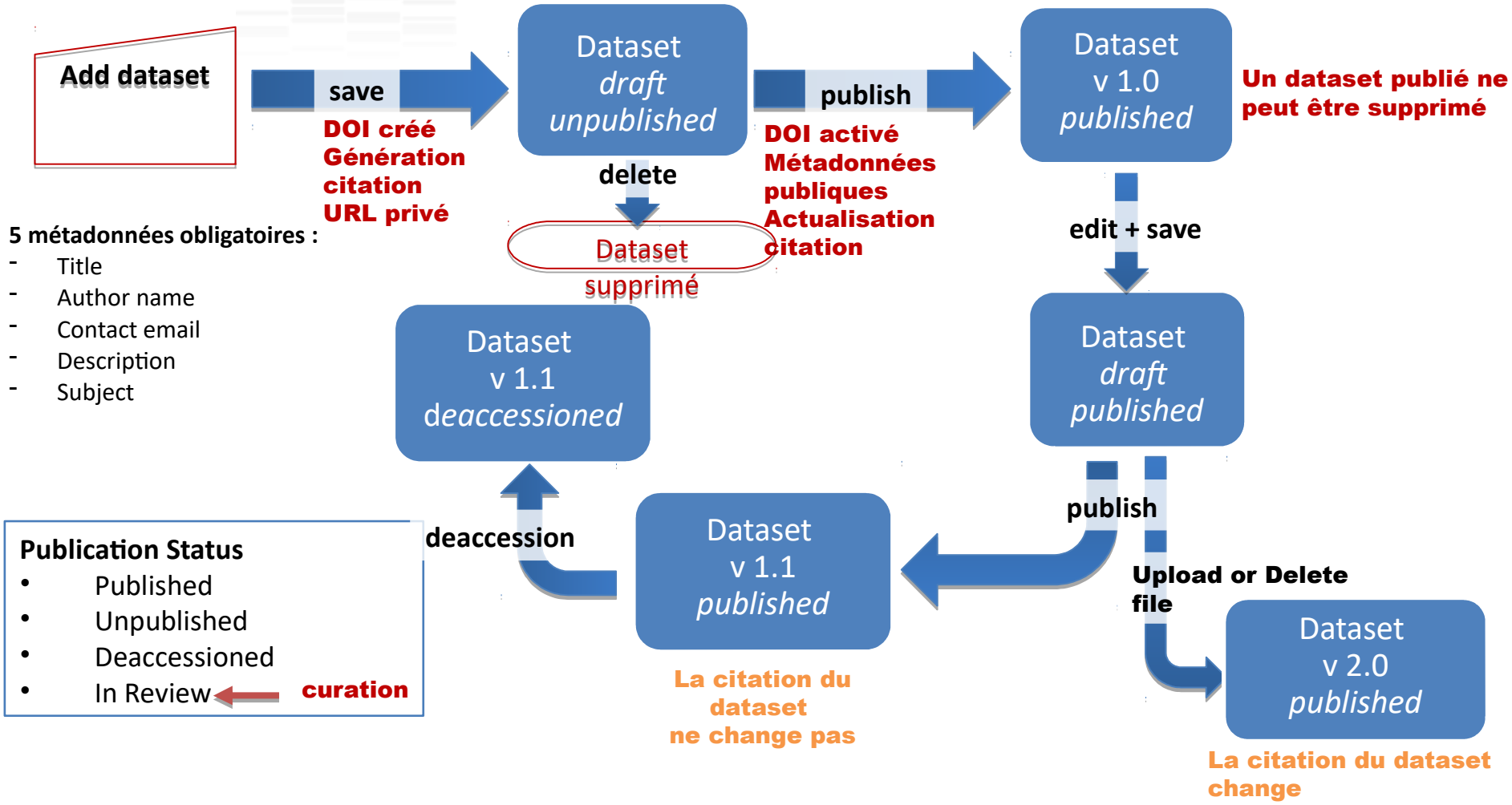
- Attribution DOI unique et pérenne
- Landing page pérenne
- Versionning des données
- Citation

Accès pérenne

- Accès ouvert aux métadonnées via landing page
- Accès ouvert ou restreint aux données
- Outils complémentaires : data papers, cartographie (worldmap)



Processus simplifié de publication dans Data Inra



BDD dynamiques

Data Inra

Via IHM / API Data INRA

BDD opérationnelle

J'identifie la base de données en tant qu'objet conceptuel pour qu'elle soit citée en tant que tel

- Valorisation de la base de données dans son ensemble
- Reconnaissance des auteurs / contributeurs

Décrire la base de données dans Data Inra
Indiquer la procédure pour accéder aux données en cas de besoin

- Génération d'une **page descriptive** (métadonnées, citation, modalités d'accès, conditions d'utilisation)
- Attribution automatique d'un **DOI** pointant vers la page descriptive

J'analyse une partie des données et publie un article : je veux citer avec précision le sous-ensemble de données (dataset) que j'ai analysé

- **Accès pérenne** à la **version précise** du jeu de données utilisé pour obtenir un résultat (transparence, reproductibilité)
- Lien avec l'article (meilleure visibilité article et données)

- Décrire le dataset dans Data Inra
- Déposer le dataset dans Data Inra ou autre entrepôt de confiance

- Génération d'une page descriptive
- Attribution automatique d'un DOI pointant vers la page descriptive
- Fichiers de données accessibles depuis la page descriptive

J'archive régulièrement la totalité ou des sous parties de base de données et partage les archives (datasets) avec la communauté

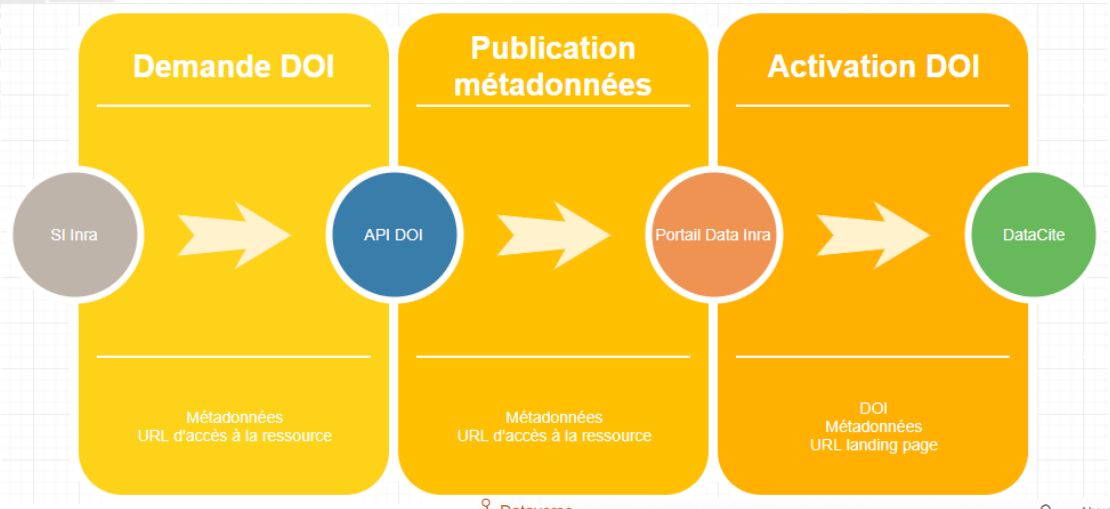
- Valorisation de la base de données
- Reconnaissance des auteurs / contributeurs
- Accès pérenne à la version précise en cas de réutilisation pour produire un article
- Lien avec article en cas de réutilisation pour produire un article

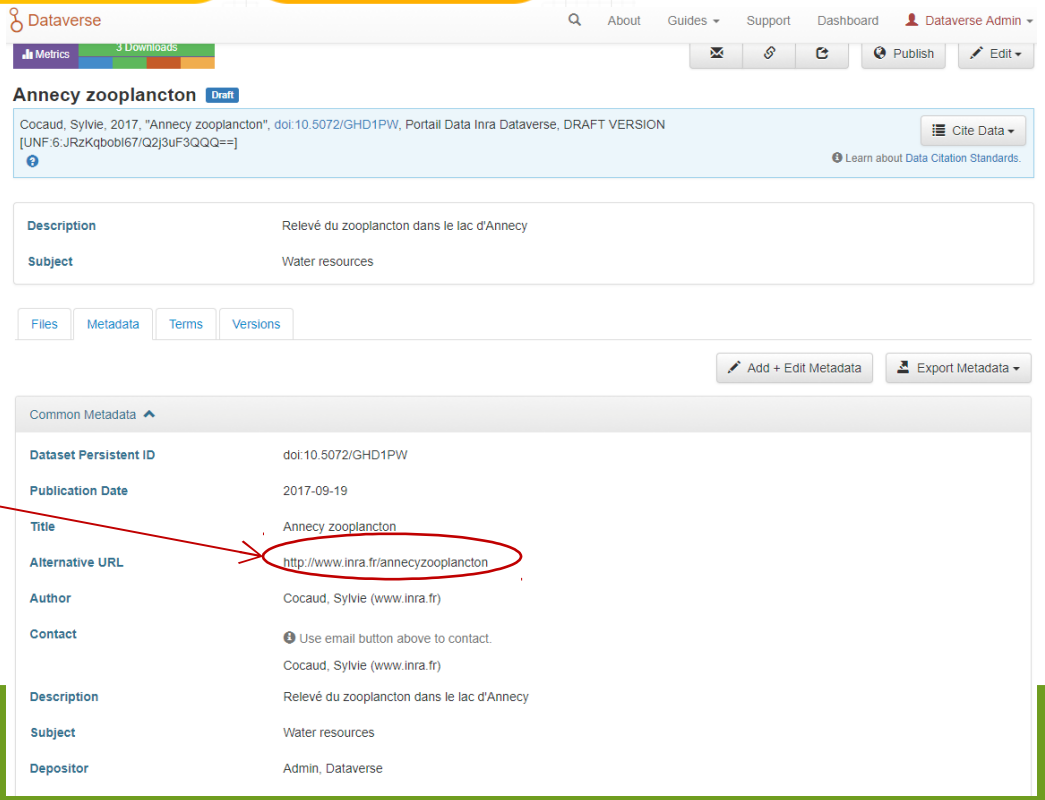
- Décrire le dataset dans Data Inra
- Déposer le dataset dans Data Inra ou autre entrepôt de confiance

- Génération d'une page descriptive à partir de laquelle toutes les versions sont accessibles
- Attribution d'un DOI pointant vers la version la plus récente de la page descriptive
- possibilité de comparer différentes versions
- Fichiers de données accessibles depuis la page descriptive

Ex: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/SUCQ52

Interopérabilité SI INRA





La page descriptive de la ressource dans Data Inra contient : citation + DOI + métadonnées + versions + termes d'usage

→ *Le DOI pointe vers la page descriptive dans Data Inra*

→ *La page descriptive inclut l'URL accès à la ressource dans le SI d'origine*

→ *Les fichiers de données sont disponibles dans le SI d'origine*

Portail d'information Datapartage

http://datapartage.inra.fr

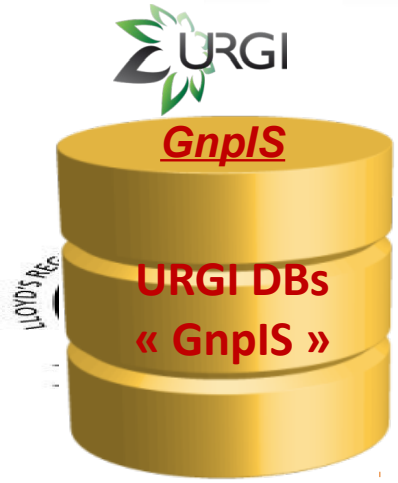
- actualités,
- informations,
- accès direct aux services…

_03

Plant Use Case : GnpIS Repository for Linked Data publication

INRA
SCIENCE & IMPACT

INRA information system for crops, forest trees and pathogens



From INRA archive…

Phenotypes

- GxE
- QTL maps
- GWAS, GS

Genetics

- Genetic maps
- Genetic markers
- Genetic resources

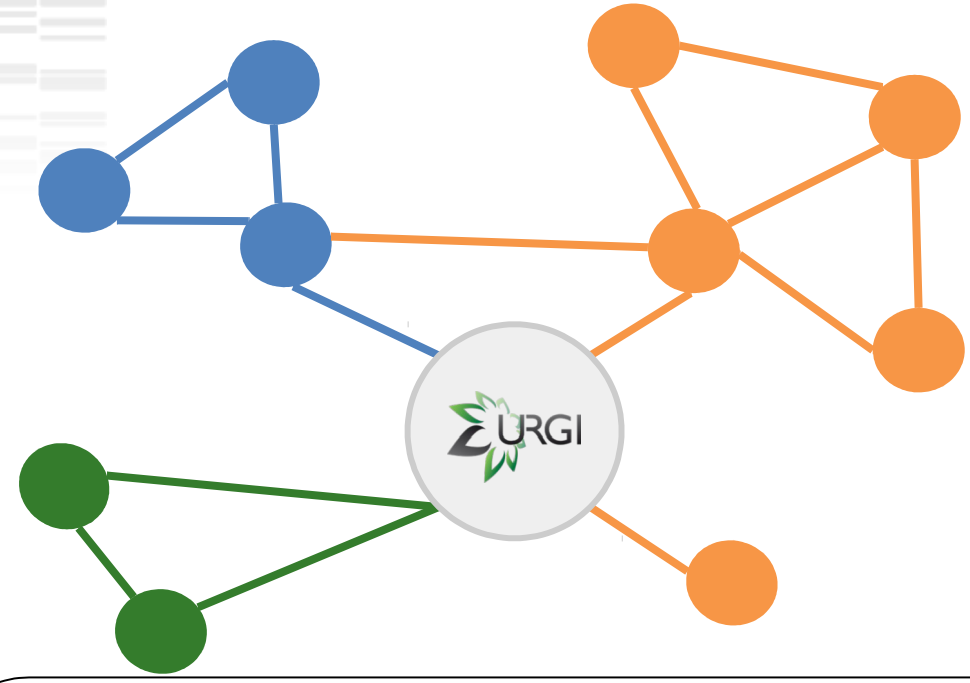
Genome

- Annotation
- Transcriptome
- SNP / Structural variants

International network

French Networks

European Networks



Global Networks

International data standards

BrAPI
Web services

RDA
RESEARCH DATA ALLIANCE

miappe
Minimal information

Crop Ontology
for agricultural data
Controlled vocabularies
Trait dictionaries

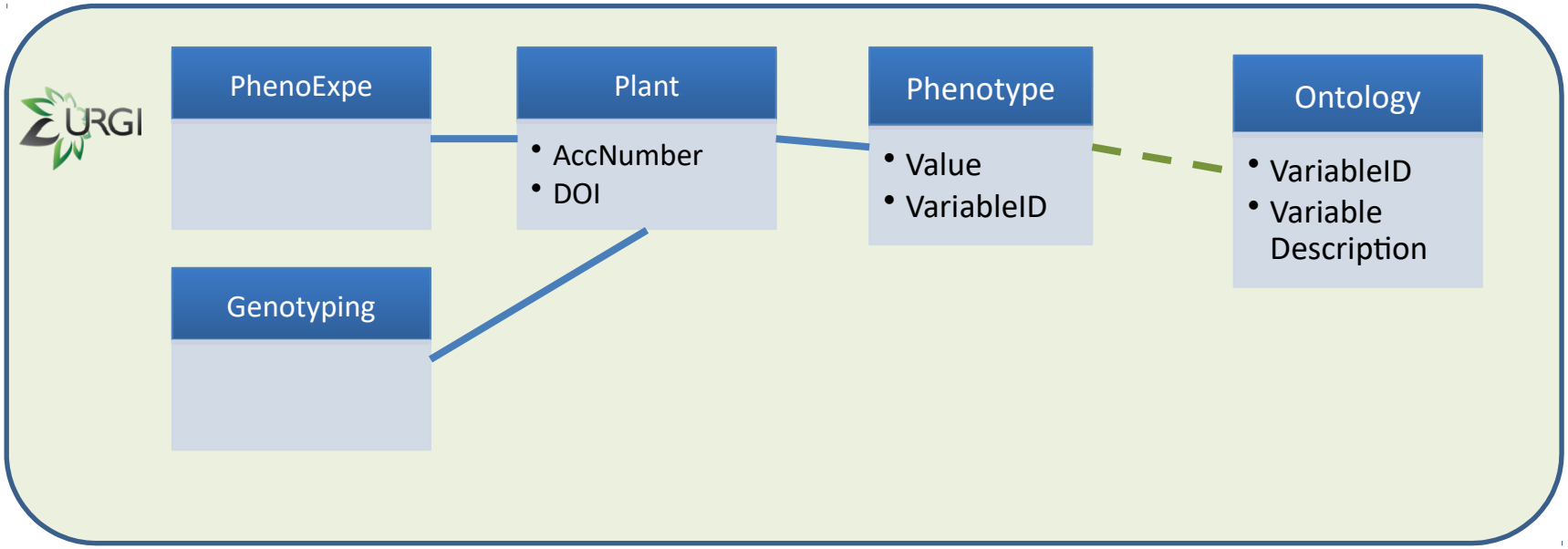
MCPD

GnpIS & Dataportal complementarity

- INRA Dataportal
  - Data discovery
  - Data Publication
  - Minimal Metadata
    - Dublin core
    - Description (text, material & method, etc…)
    - …
  - Data archive
- GnpIS
  - Integration and linked Data
  - Data publication
  - Rich Metadata
  - Rich, browsable Data
  - Findable Metadata & Data
    - Full text
    - By Object types (Plant material, Phenotype Ontology)
    - Distributed (National & international search portal)

GnpIS Linked Data

- Internal
  - Between phenotyping experiment in a data set
  - Between phenotyping datasets
  - From Phenotypes to genetic (genotyping, QTL, GWAS) then to genomic
  - Technical : from PgSQL to ElasticSearch
  - **Pivot Object (shared key resource) definition**



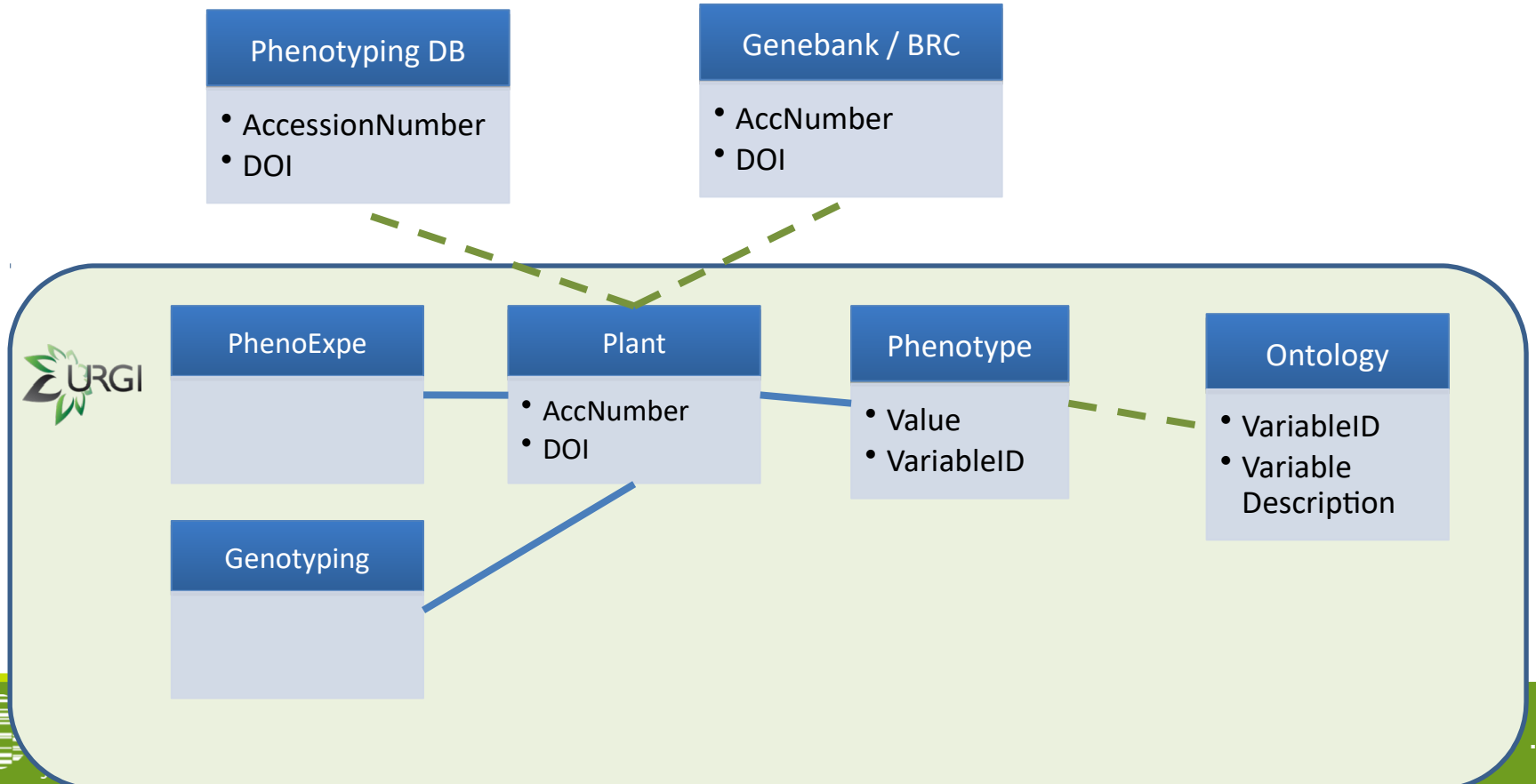
FAIR for plant phenotypic data through GnpIS-Ephesis

Pommier C[1], Michotey C[1], Lebreton A[1], Cornut G[1], Flores R[1], Alaux M[1], Durand S[1], Kimmel E[1], Letellier T[1],

Writting In progress

GnpIS Linked Data

- External
  - Genetic Resources (accessions)
    - Gene bank to GnpIS
  - Trait Phenotypes from Ontologies



Phenotyping Dataset Definition

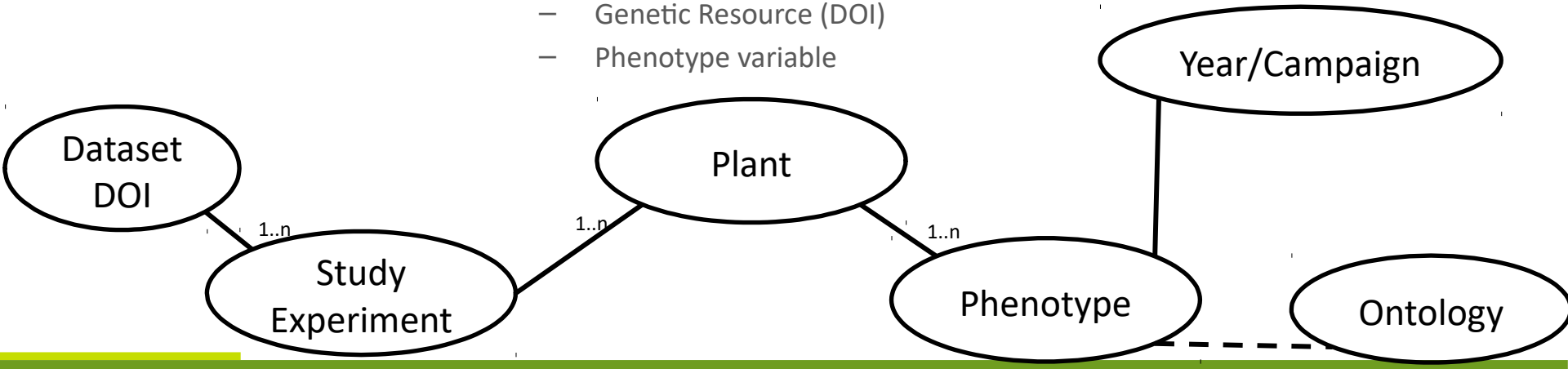
Factors | Design details | Position | Phenotype, Environment, …

lot_Number	itk	bloc	plot	X	Y	X(m)	Y(m)	agd_value	agd_date
ch	fi	1	26210	12	2	36	4	14	
ch	fi	1	26210	12	3	36	6	14	
ch	fi	2	226210	16	2	48	4	25	
ch	fi	2	226210	16	3	48	6	25	

README | Trial | Contacts | Variables | VarietalList | Treatments | StudySubjectLevels | Data | Meteo_Data | Soil_Data | ITK

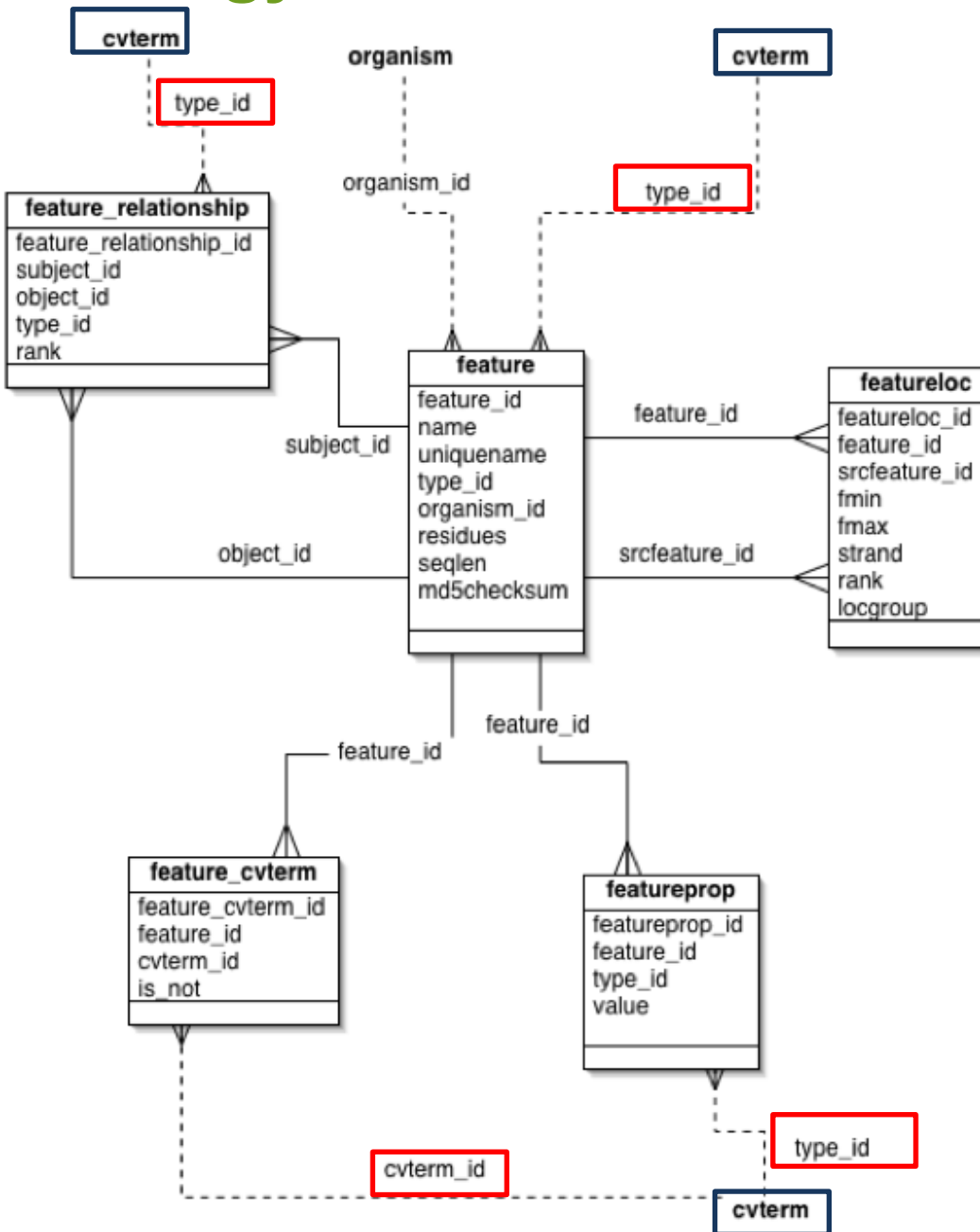
Genetic Resource

- Dataset
  - List of experiment: location
  - DOI, Data Publication
- Pivot Objects
  - Genetic Resource (DOI)
  - Phenotype variable



Phenotype Ontology model

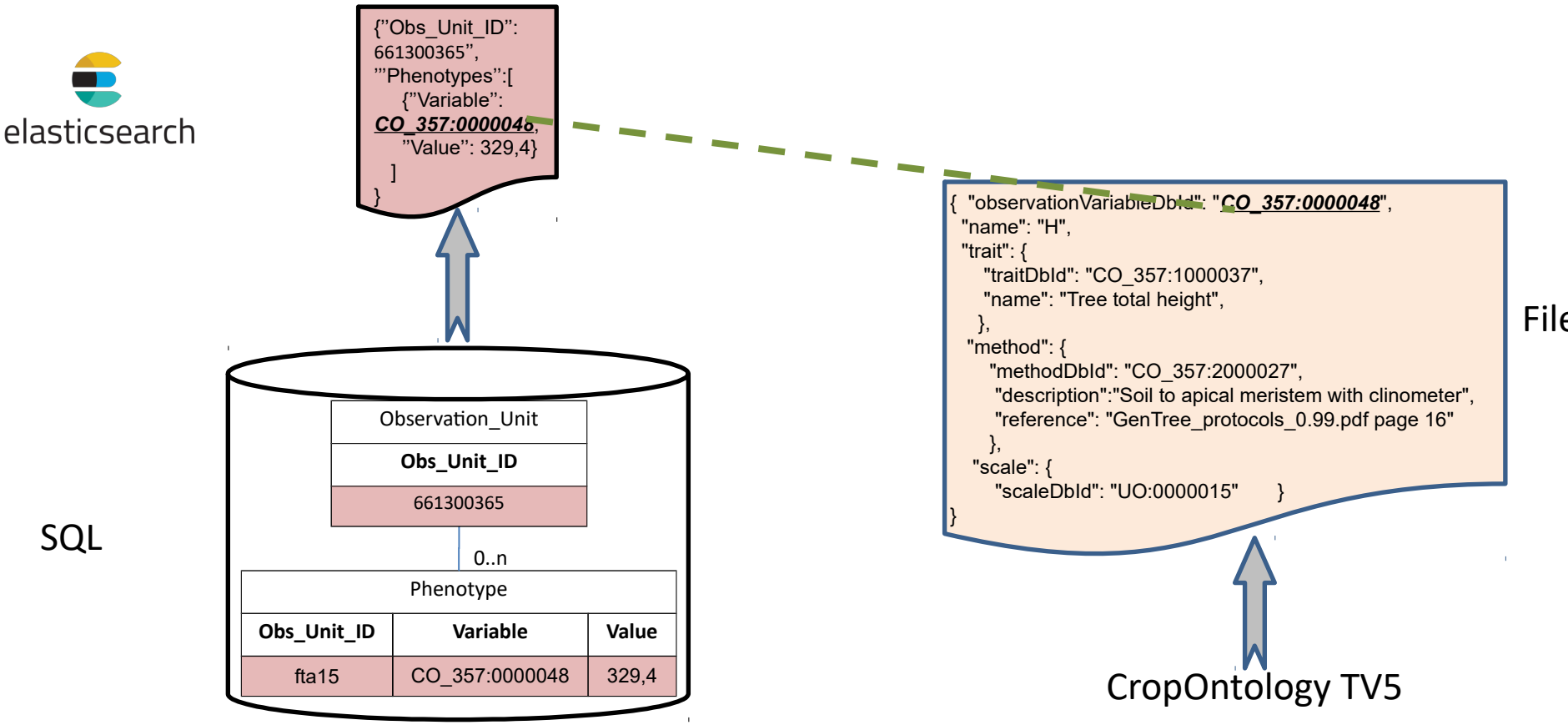
- GMOD Chado Approach
- Ontology Inserted in the database
- Annotation
  - PK/FK
    - FeatureType <-> cvterm
    - Featureprop <-> cvterm
    - FeatureAnnotation <-> cvterm
    - …
- Update Ontology → Update all PK/FK
- Complicated and Error prone



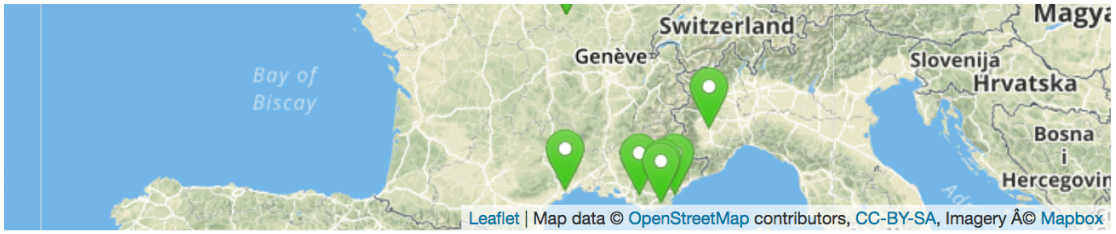
Ontology Managment solution

- Two Datasets
  - Phenotyping Data
  - Ontology
- Linked data
  - Linked by IDs
  - XRef like
- Each dataset (Pheno & Onto) in its own repository

GnpIS Phenotyping Ontology Managment



GnpIS Phenotyping Ontology Managment



Trial list | Phenotypic data

LEVEL: TRIAL

	Trial Site	Campaign	Date	Tree total height at 1 year (Height1)	Tree total height at 2 years (Heigl
ursery	Ardon	2004	2004-12-18		
ursery	Ardon	2004	2004-01-12	329.4	
ursery	Ardon	2004	2004-01-12	306.3	
ursery	Ardon	2004	2004-12-18		614.5
ursery	Ardon	2004	2004-01-12	285.3	
ursery	Ardon	2004	2004-12-18		667.2
ursery	Ardon	2004	2004-01-12	252.4	
ursery	Cavallerma ore	2004	2004-01-01	305.60000000000002	

Total height of the tree (from ground to tallest part of the crown)

Rest WebServices

BrAPI

/phenotype-search

/variables/CO_357%3A0000048

elasticsearch

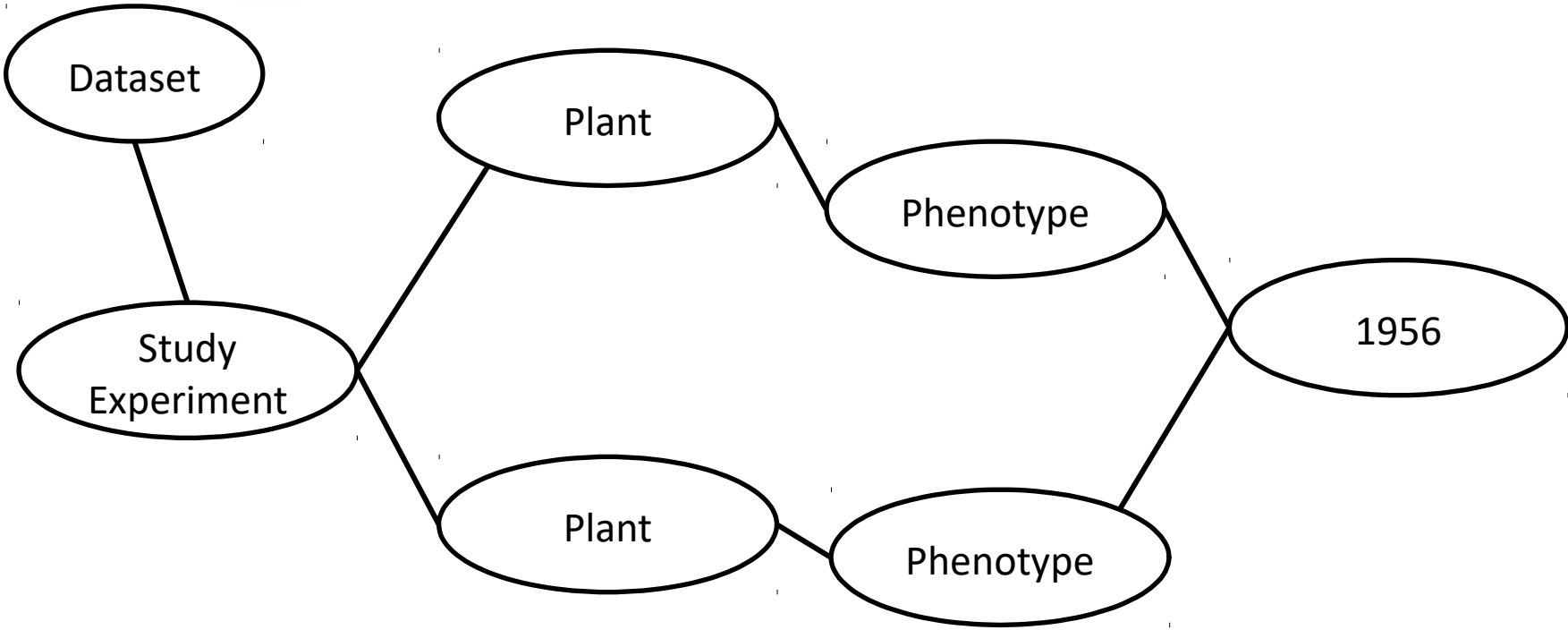
{"Obs_Unit_ID": 661300365",
"Phenotypes":[
{"Variable": ***CO_357:0000048***,
"Value": 329,4}
]
}

{ "observationVariableDbId": "***CO_357:0000048***",
"name": "H",
"trait": {
"traitDbId": "CO_357:1000037",
"name": "Tree total height",
},
"method": {
"methodDbId": "CO_357:2000027",
"description":"Soil to apical meristem with clinometer",
"reference": "GenTree_protocols_0.99.pdf page 16"
},
"scale": {
"scaleDbId": "UO:0000015" }
}

_03

Use case séries temporelles

Dataset


Dataset
Plant
Phenotype
1956
Study
Experiment
Plant
Phenotype

Laver 1 : 1956

Origin site | Collecting site | Evaluation site

Phenotyping campaign(s) 1956 ×

remove all add all

Trial list | **Phenotypic data**

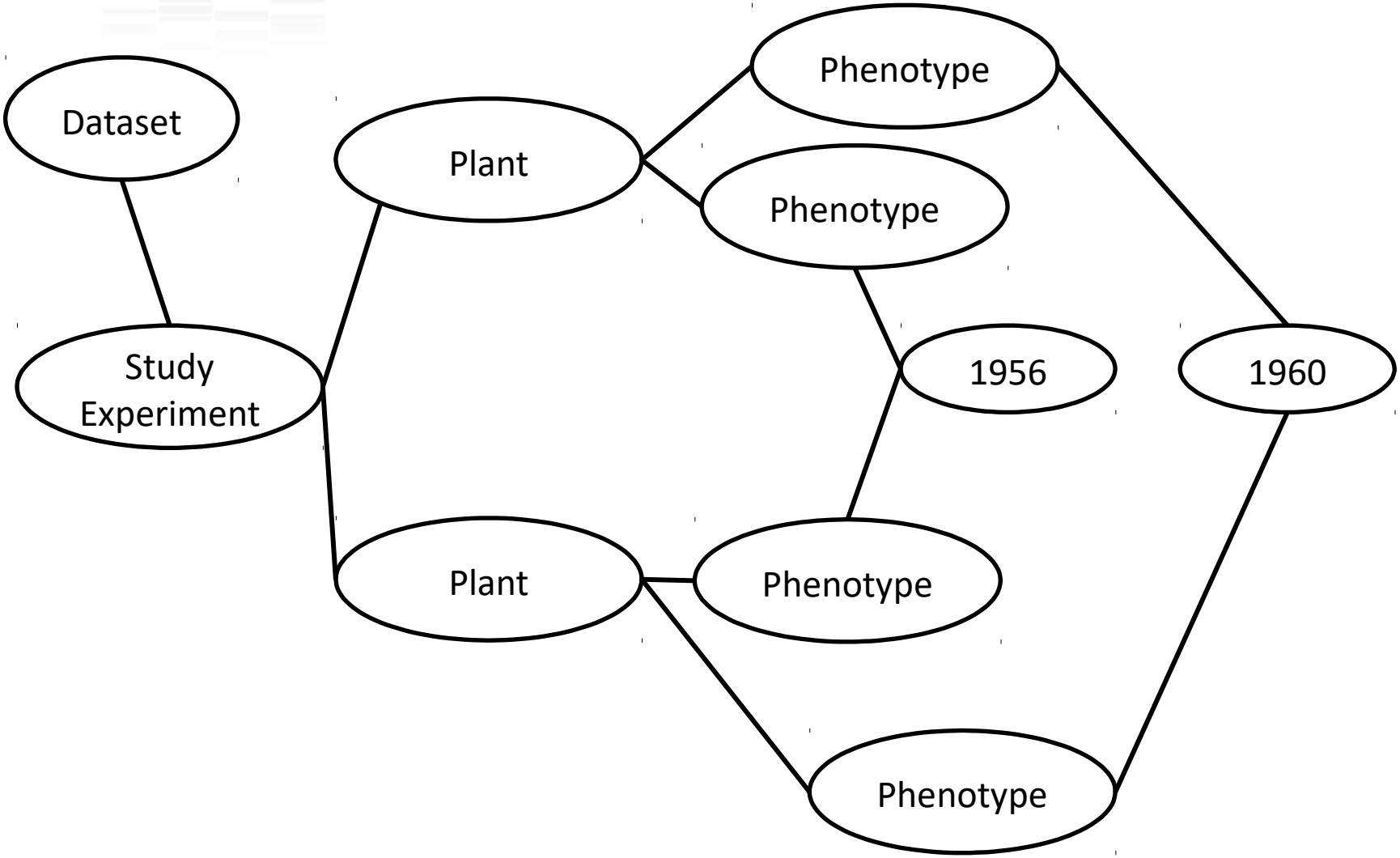
LEVEL: TRIAL

play 10 results per page

Accession Name	Trial Name	Trial Site	Campaign	Budbreak date (50%) (B
Inconnu = Carignan	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-03-29
Pinot Renevey amélioré	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-03-28
Pinot Crépet	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-03-26
Pinot Crépet	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-03-28
Plant gris de Ludes = Pinot noir	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-03-30
Vert noir = Pinot noir	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-03-26
Sauvignon blanc N°13	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-04-03
Cabernet Sauvignon	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-04-03

INRA SCIENCE & IMPACT

Dataset


Dataset
Study Experiment
Plant
Plant
Phenotype
Phenotype
Phenotype
Phenotype
1956
1960

Layer 2 : 1960

- Dataset Extension
- New layer on existing phenotype variable
- No Update of existing data

Origin site | Collecting site | Evaluation site

Phenotyping campaign(s) 1956 × 1960 ×

remove all add all

Trial list | **Phenotypic data**

LEVEL: TRIAL

play 10 results per page

Accession Name	Trial Name	Trial Site	Campaign	Budbreak date (50%)
Inconnu = Carignan	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-03-29
Inconnu = Carignan	Données phénologiques brutes Vassal témoins	Vassal-UE	1960	1960-03-16
Pinot Renevey amélioré	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-03-28
Pinot Crépet	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-03-26
Pinot Crépet	Données phénologiques brutes Vassal témoins	Vassal-UE	1960	1960-03-18
Vert noir = Pinot noir	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-03-26
Cabernet Sauvignon	Données phénologiques brutes Vassal témoins	Vassal-UE	1956	1956-04-03

INRA SCIENCE & IMPACT

Data Elaboration

- New Variable

BUD_DATE: Budbreak date (50%) VARIABLE

Ontology name	Vitis inra ontology
Identifier	CO_356:1000001
Name	BUD_DATE
Synonyms	Budbreak date (50%)
Institution	INRA
Scientist	Eric Duchene
Crop	VITIS

Budbreak TRAIT

Identifier	CO_356:2000000
Name	Budbreak
Entity	Plant
Attribute	Budbreak
Class	Phenological

Bud_date method METHOD

Identifier	CO_356:3000146
Name	BUD_DATE Method
Class	Estimation

Calendar date SCALE

Identifier	CO_356:4000003
Name	Calendar date
Data type	Time

MI-BUD-relativ: Budbreak date (50%) relative to Chasselas VARIABLE

Ontology name	Vitis inra ontology
Identifier	CO_356:1000002
Name	MI-BUD-relativ
Synonyms	Budbreak date (50%) relative to Chasselas
Institution	INRA
Scientist	Eric Duchene
Crop	VITIS

Budbreak TRAIT

Identifier	CO_356:2000000
Name	Budbreak
Entity	Plant
Attribute	Budbreak
Class	Phenological

Relative to chasselas METHOD

Identifier	CO_356:3000001
Name	Relative to chasselas
Description	Number of day before (-) or after (+) the Chasselas cultivar for this stage. This reference cultivar is often used as phenology control in grape germplasm collection (ex. INRA Vassal)
Class	Computation

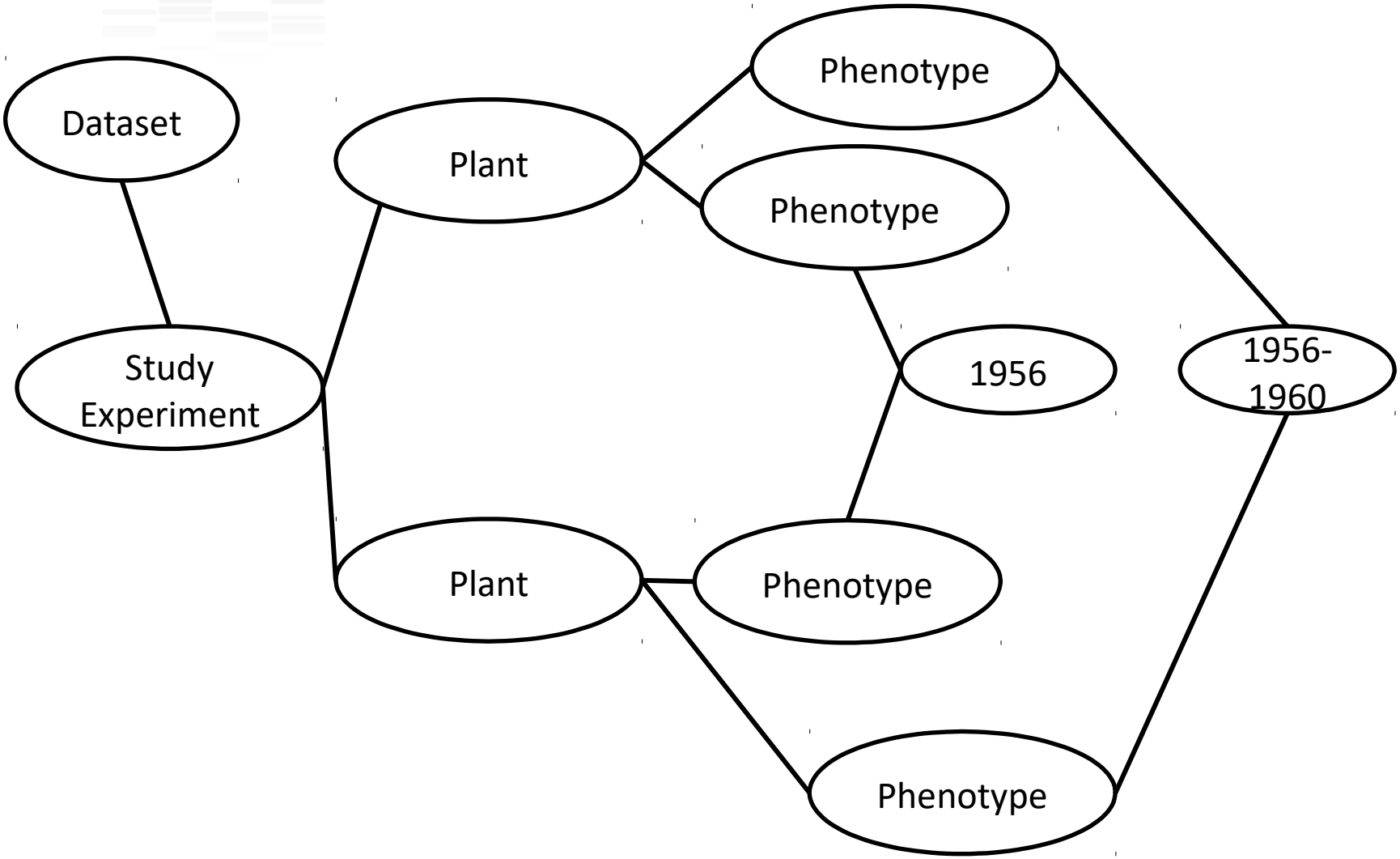
Phenotyping campaign(s)

Trial list | Phenotypic data

LEVEL: TRIAL

	Trial Site	Campaign	Budbreak date (50%) relative to Chasselas (MI-BUD-relativ)	Budbreak date (50%) (BUD_DATE)
	Vassal-UE	1960		1960-03-18
var	Vassal-UE	1956-2012	7	
var	Vassal-UE	1956-2012	9	
var	Vassal-UE	1956-2012	12	

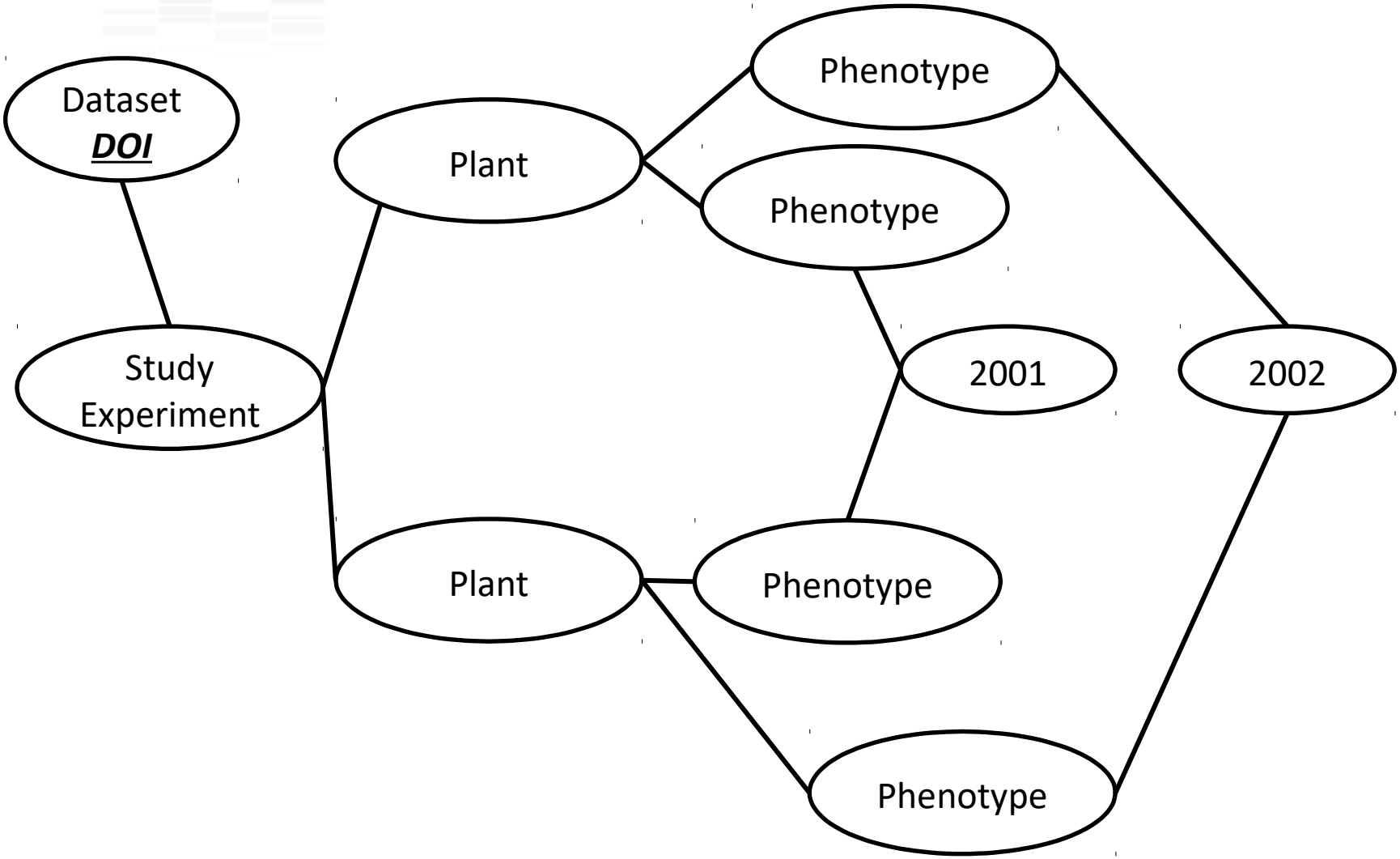
Dataset


Dataset
Study Experiment
Plant
Plant
Phenotype
Phenotype
Phenotype
Phenotype
1956
1956-1960

_03

Use case céréales à paille

Dataset


Dataset
DOI
Study
Experiment
Plant
Plant
Phenotype
Phenotype
Phenotype
Phenotype
2001
2002

DOI.Org Web services

- 2014 & 2015 addition
- Stable DOI

Doi:10.15454/1.4489666216568333E12

- No new version
- Perspectives :
  - New Versions of the same DOI

Winter wheat (Triticum aestivum L) phenotypic data from the multiannual, multilocal field trials of the INRA Small Grain Cereals Network.

François-Xavier Oury, Emmanuel Heumez, Bernard Rolland, Jérôme Auzanneau, Pierre Bérard, Maryse Brancourt-Hulmel, Xavier Charrier, Hubert Chiron, Camille Depatureaux, Laurent Falchetto, Olivier Gardet, Stéphane Gilles, Alex Giraud, Christophe Lecomte, Jean-Yves Morlais, Pierre Pluchard, Didier Tropée, Maxime Trottet, Patrice Walczak, Gérard Doussinault, Michel Rousset, Gilles Charmet

Query dataset as a semantic graph.

Or download the dataset as RDF archive.

⊞ Abstract

Published 2015 by INRA

Back to Form

⊟ Search parameter(s):

Geolocation

DATA SETS: 4

Network Data Set :
INRA Wheat Network BRC accession (A series)

Network Data Set :
INRA Small Grain Cereals Network
DOI:http://dx.doi.org/10.15454/1.4489666216568333E12

Network Data Set :
INRA Wheat Network not BRC accession (B and C series)

Origin site | Collecting site | Evaluation site

Phenotyping campaign(s): 2000 × 2001 × 2002 × 2003 × 2004 × 2005 × 2006 × 2007 × 2008 × 2009 × 2010 × 2011 × 2012 × 2013 × 2014 × 2015 ×

remove all add all

Trial list | Phenotypic data

URGI GnpIS 813 trials

INRA SCIENCE & IMPACT

Dataset Format

- Increasing complexity, ie data quality and documentation
- Archive
  - MIAPPE Compliant zip
- Web Pages
- Web services
  - Breeding API
- RDF
  - http://wheatis.org/DataStandards.php
  - http://ist.blogs.inra.fr/wdi/phenotypes-as-rdf/
  - Dataset Ids
  - External :
    - DOI : Plant material
    - URI : Ontology
  - Internal : URI
    - Non dereferencable, ID only
    - Perspectives : PURL or INRA URI dereferencer
      - www.datapartage.inra.fr

Dynamic Dataset Updates

- Addition of new layers on time series
  - Version incrementing every 2 years if necessary
    - Update Database Object in GnpIS to handle Version
  - Dynamic dataset/DB
- Reproducibility
  - Re run an analysis on a given dataset
    - → expect the same result
  - Data Analysis dataset
    - Archive
    - Extraction from Dynamic Dataset