# Partage des données de la recherche

Dossier suivi par : I. Blanc (Sup SIS), C. Gaspin (MIA) et O. Hologne (DIST)

# Plan

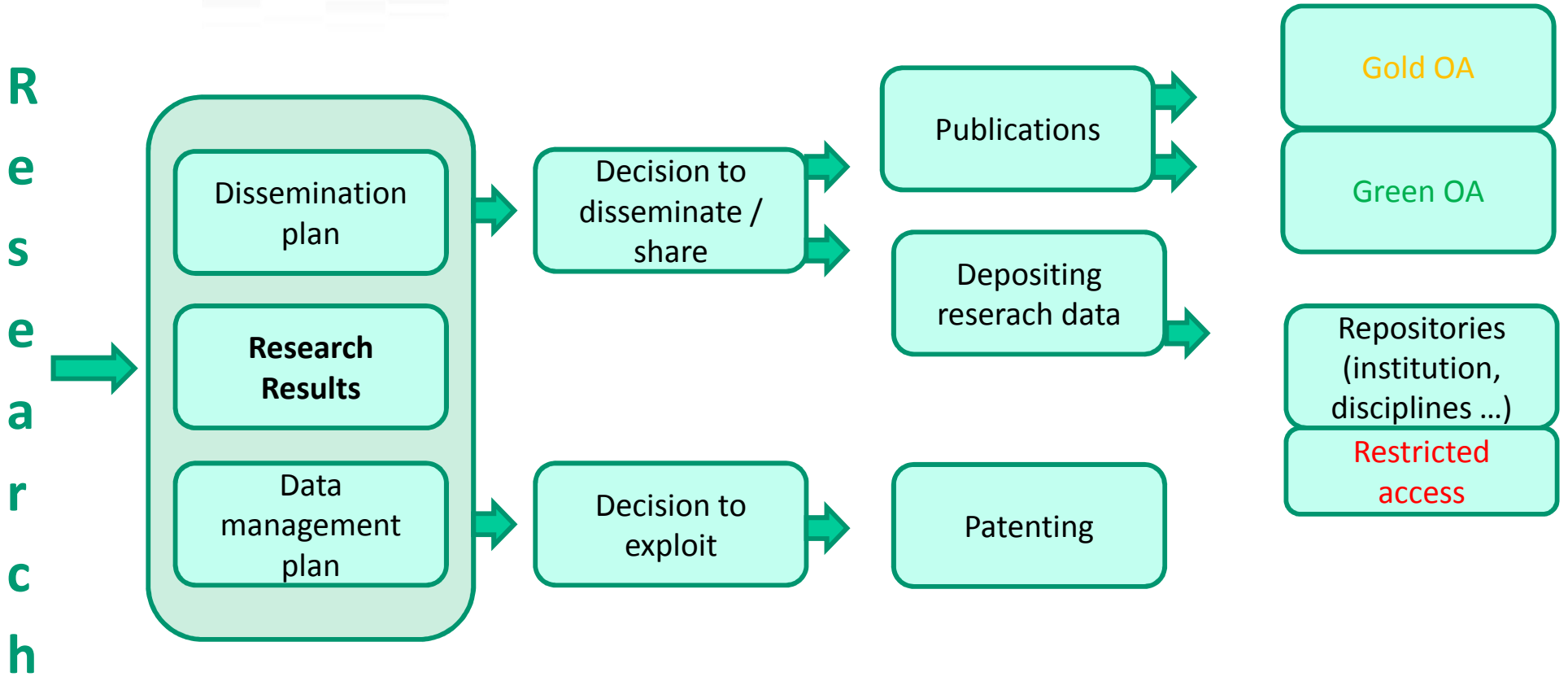
- Contexte international, scientifique et politique
- Approche de l'Inra
- Exemples d'actions ou projets
- Conclusion

# _02

# Contexte

## Science, politique et … publications

# Dissemination dans


# OA to publications mandate in H2020

**Each beneficiary must ensure OA to all peer-reviewed scientific publications relating to its results:**

- **Deposit** a machine-readable copy of the published version or final peer-reviewed manuscript accepted for publication in a repository of the researchers choice (possibly OpenAIRE compliant)
- **Ensure OA** on publication or at the latest within 6 months (12 for SSH)
- **Aim to deposit** at the same time **the research data needed to validate the results ("underlying data")**
- Ensure **OA to the bibliographic metadata** that identify the deposited publication, via the repository

Celina Ramjoue (Head of Sector “Open Access to scientific Publications and Data”, EC DG CNECT)

# Infrastructure européenne Open access, Open data

- publi : OpenAire
- Données : OpenAire +

Zenodo

ZENODO is a repository service that enables researchers, scientists, projects and institutions to share and showcase multidisciplinary research results (data and publications) that are not part of existing institutional or subject-based repositories.

# Code de la recherche modifié par la loi ESR de juillet 2013

Chapitre II : Objectifs et moyens institutionnels de la recherche publique.

**Article L112-1** En savoir plus sur cet article...
Modifié par LOI n°2013-660 du 22 juillet 2013 - art. 16

La recherche publique a pour objectifs :

a) Le développement et le progrès de la recherche dans tous les domaines de la connaissance ;

b) La valorisation des résultats de la recherche au service de la société, qui s'appuie sur l'innovation et le transfert de technologie ;

c) Le partage et la diffusion des connaissances scientifiques en donnant priorité aux formats libres d'accès ;

c bis) Le développement d'une capacité d'expertise et d'appui aux associations et fondations, reconnues d'utilité publique, et aux politiques publiques menées pour répondre aux défis sociétaux, aux besoins sociaux, économiques et du développement durable ;

d) La formation à la recherche et par la recherche ;

e) L'organisation de l'accès libre aux données scientifiques.

Les établissements publics de recherche et les établissements d'enseignement supérieur favorisent le développement des travaux de coopération avec les associations et fondations, reconnues d'utilité publique. Ils participent à la promotion de la recherche participative et au développement des capacités d'innovation technologique et sociale de la Nation. Ces coopérations s'exercent dans le respect de l'indépendance des chercheurs et, en l'absence de clauses contraires, dans un but non lucratif. Les travaux de recherche menés dans le cadre de ces coopérations sont, en l'absence de clauses contraires, rendus publics et accessibles.

# Du côté de l'édition

## Des nouvelles revues :

**BGI Shenzhen and BioMed Central**

GigaScience aims to revolutionize data dissemination, organization, understanding, and use. An online open-access open-data journal, **we publish 'big-data' studies from the entire spectrum of life and biomedical sciences**. To achieve our goals, the journal has a novel publication format: one that **links standard manuscript publication with an extensive database that hosts all associated data and provides data analysis tools and cloud-computing resources.**

**Not just 'omic' type data** … imaging, neuroscience, ecology, cohort data, systems biology and other new types of large-scale sharable data.

## Note aux auteurs – revues classiques

**Data and materials availability** All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*. ]…[

*Science* supports the efforts of databases that aggregate published data for the use of the scientific community. Therefore, appropriate data sets (including microarray data, protein or DNA sequences, atomic coordinates or electron microscopy maps for macromolecular structures, and climate data) must be deposited in an approved database, and an accession number or a specific access address must be included in the published paper. We encourage compliance with MIBBI guidelines (Minimum Information for Biological and Biomedical Investigations).

## De nouveaux entrepôts de données

**Dryad is an international repository of data** underlying peer-reviewed articles in the basic and applied biosciences. Dryad enables scientists to validate published findings, explore new analysis methodologies, repurpose data for research questions unanticipated by the original authors, and perform synthetic studies. Dryad is governed by a consortium of journals that collaboratively promote data archiving and ensure the sustainability of the repository.


DEFINITIONS:

**Data repository**: a database or collection comprising data studies, data sets and/or microcitations which stores and provides access to the raw data. Constituent data studies, and sometimes individual data sets, are marked up with metadata providing a context for the available raw data.

**Data study:** description of studies or experiments held in repositories with the associated data which have been used in the data study. (Includes serial or longitudinal studies over time). Data studies can be a citable object in the literature and may have cited references attached in their metadata, together with information on such aspects as the principal investigators, funding information, subject terms, geographic coverage etc. The level of metadata provided varies between repositories.

**Data set**: a single or coherent set of data or a data file provided by the repository, as part of a collection, data study or experiment. Data sets may present in a number of file formats and media types: they may be number based files such as spreadsheets, images, video, audio, databases etc. Data sets can be a citable object in the literature and may have cited references attached in their metadata, but more commonly they inherit the metadata of the overall study in which they are used.

1. Title: **ESTerases and alpha/beta Hydrolase Enzymes and Relatives.**
Editor(s): Hotelier, Thierry; Renault, Ludovic; Cousin, Xavier; et al.
Source: ESTerases and alpha/beta Hydrolase Enzymes and Relatives
Source URL: **http://bioweb.ensam.inra.fr/ESTHER/general?what=index**
Document Type: Repository Times Cited: 2 (from All Databases)
[ View abstract ]

1. Title: **Enzymatic Activity and Protein Interactions in Alpha/Beta Hydrolase Fold Proteins: Moonlighting Versus Promiscuity**
Author(s): Marchot, Pascale; Chatonnet, Arnaud
Source: PROTEIN AND PEPTIDE LETTERS Volume: **19** Issue: **2** Pages: **132-143** Published: **FEB 2012**
Times Cited: 3 (from All Databases)
S·F·X [ View abstract ]

2. Title: **ESTHER, the database of the alpha/beta-hydrolase fold superfamily of proteins**
Author(s): Hotelier, T; Renault, L; Cousin, X; et al.
Source: NUCLEIC ACIDS RESEARCH Volume: **32** Special Issue: **SI** Pages: **D145-D147** DOI: **10.1093/nar/gkh141** Published: **JAN 1 2004**
Times Cited: 79 (from All Databases)
S·F·X Full Text [ View abstract ]

# Des opportunités au niveau international


Build the social, organisational and technical infrastructure to reduce barriers in data sharing (March 2013)


Share relevant agricultural data available from G-8 countries with African partners and … develop options for the establishment of a global platform to make reliable agricultural and related information available to African partners (April 2013)

GODAN
Global Open Data for Agriculture and Nutrition

seeks to support global efforts to make agricultural and nutritionally relevant data available, accessible, and usable for unrestricted use worldwide. (Oct. 2013)

# _03

# Approche de l'Inra

# Partage des données à l'Inra : étapes clés

**2011-2012 : le CS instruit la question**

- Groupe de travail piloté par D. Pontier
- 9 recommandations pour l'Inra (rapport Juin 2012)


Rapport CS
http://prodinra.inra.fr/record/206746

**Fin 2012 : élaboration d'une politique**

- 11 principes pour mieux gérer et partager les données de la recherche


Note de cadrage :
Principes en matière de partage des données de la recherche

Avril 2013 : lancement des chantiers de mise en œuvre

- 3 chantiers disciplinaires / familles de données
- Chantier juridique
- Chantiers techniques (outils et méthodes)
- Chantier compétences

Chantier
« Data Partage »

# Reco du rapport du CS : fil rouge des actions



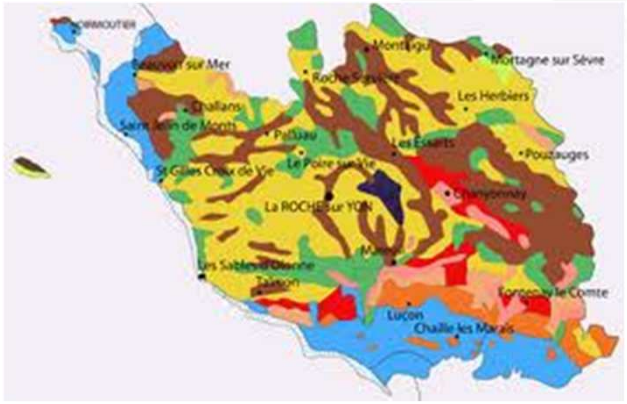
## RESUME

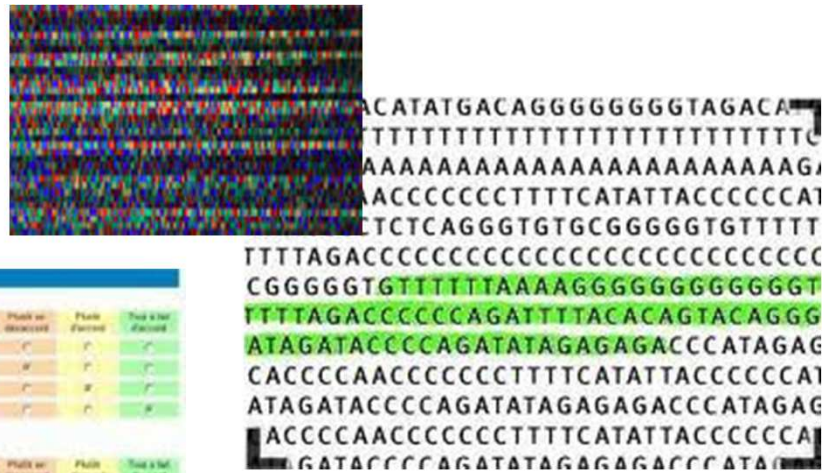
Depuis quelques années, la biologie et les sciences humaines et sociales font face à un accroissement exponentiel des données, provenant de l'adoption en masse des nouvelles technologies, et du développement des sciences et techniques de l'information d'une ampleur et à une échelle sans précédents. Une telle rupture nécessite des transformations stratégiques majeures pour assurer le stockage, la préservation, l'exploitation de ces masses de données, mais aussi leur partage. Elle nécessite également une prise de conscience et une modification des pratiques des ingénieurs et chercheurs de l'institut, pour lesquels ces évolutions constituent un défi culturel.

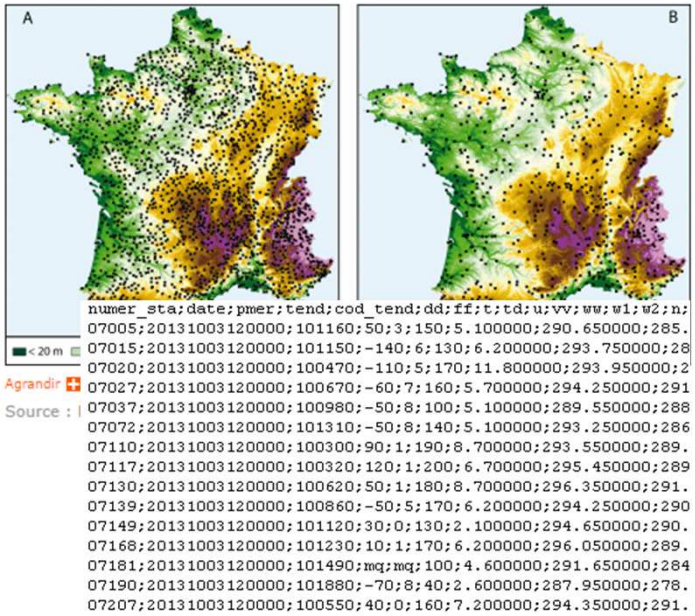
Au terme de son analyse, le groupe de travail propose les recommandations suivantes :

1) Définir la politique de l'établissement et la communiquer.
2) Mettre en place un comité d'évaluation des données produites par l'Inra.
3) S'impliquer dans les comités internationaux de standardisation.
4) Développer un portail d'accès à un ensemble de ressources distribuées.
5) Prendre en compte le cycle de vie des données dès l'élaboration des projets de recherche.
6) Définir un cahier des charges pour les plateformes.
7) Doter l'Inra d'infrastructures dimensionnées pour les stockages et les calculs hautes performances.
8) S'engager dans une politique de gestion des compétences répondant aux besoins en émergence.
9) Conduire une réflexion inter-organismes pour promouvoir une politique nationale et locale en matière de gestion et partage de données.

# Les données à l'Inra ?



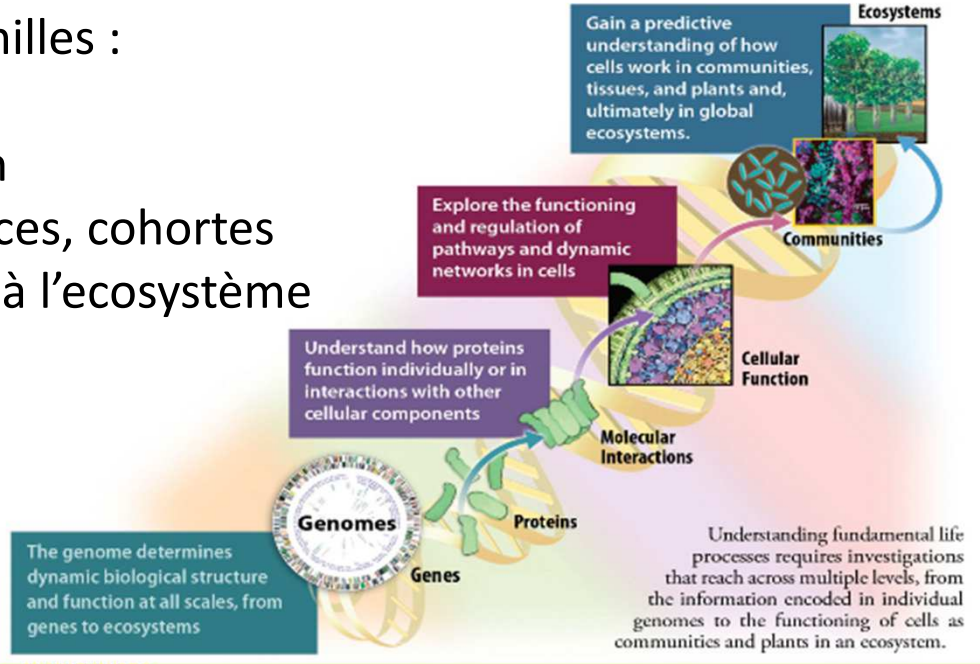



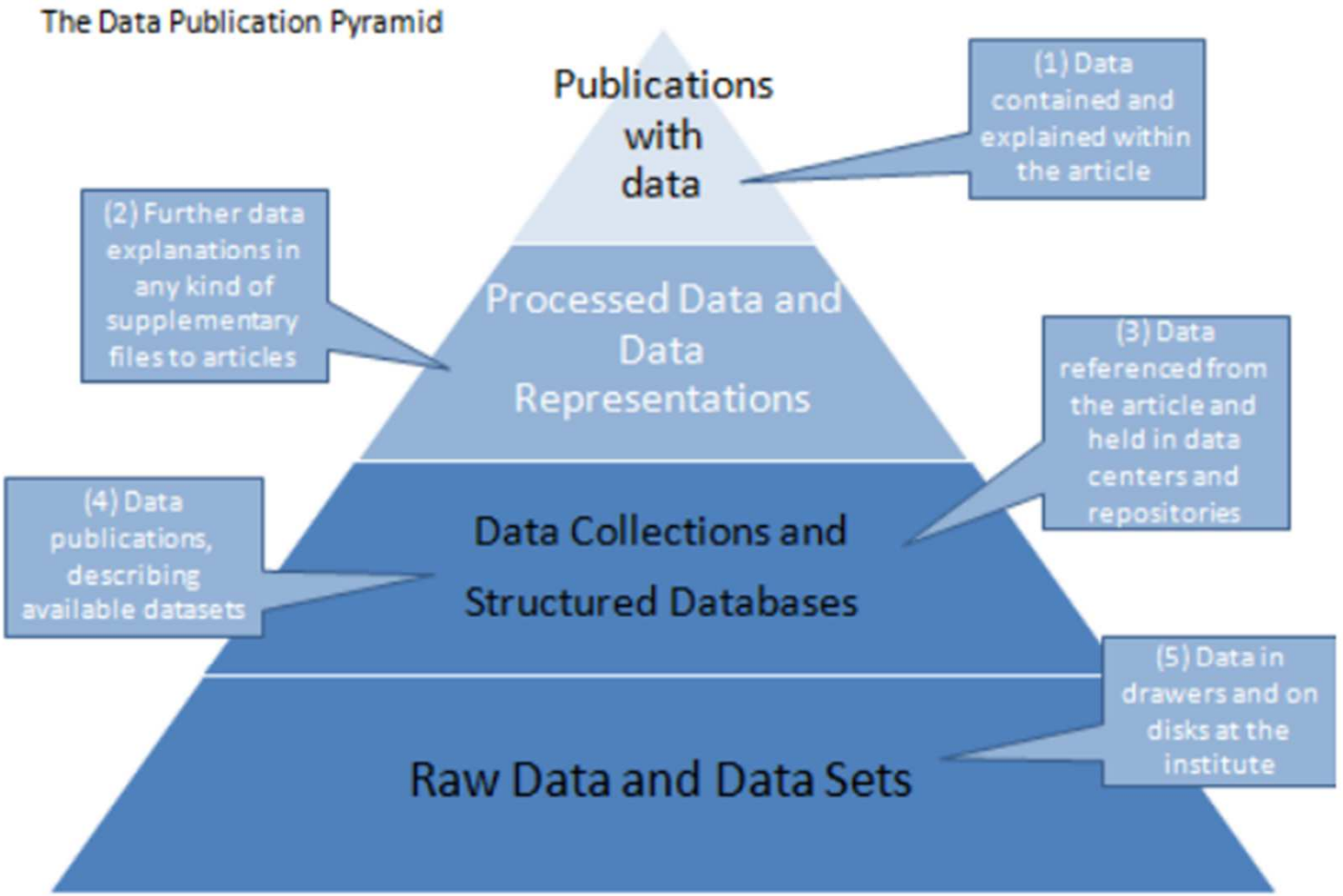





Differentes familles :

- Omics
- Observation
- Social sciences, cohortes
- du génome à l'ecosystème






Graph 1: The Data Publication Pyramid, developed on the basis of the Jim Gray pyramid, to express the different manifestation forms that research data can have in the publication process.

# Partager ?

- Le partage : « déjà un bénéfice interne »
- Différentes modalités
  - ✓ Via des entrepôts : données liées aux publications « underlying data » (H2020) = métadonnées + données
  - ✓ Via des applications (Bases de données et Web services)
- Nécessité d'avoir une vision stratégique du partage pas uniquement obligation des agences de financement) ou technique,
  - ✓ Évaluation du caractère « sensible » de la donnée
  - ✓ Clarification des règles éthiques et juridiques,

# Développement de nouveaux services

- Annuaire comme outil permettant d'identifier le patrimoine numérique Inra sur les données et, plus généralement, sur les SI Inra,
- Entrepôt (métadonnées + jeux de données)
- Attribution de DOI
- Favoriser l'ouverture des SI Inra en place et concevoir leur interopérabilité -> vers des e-infrastructures labellisées « open science »

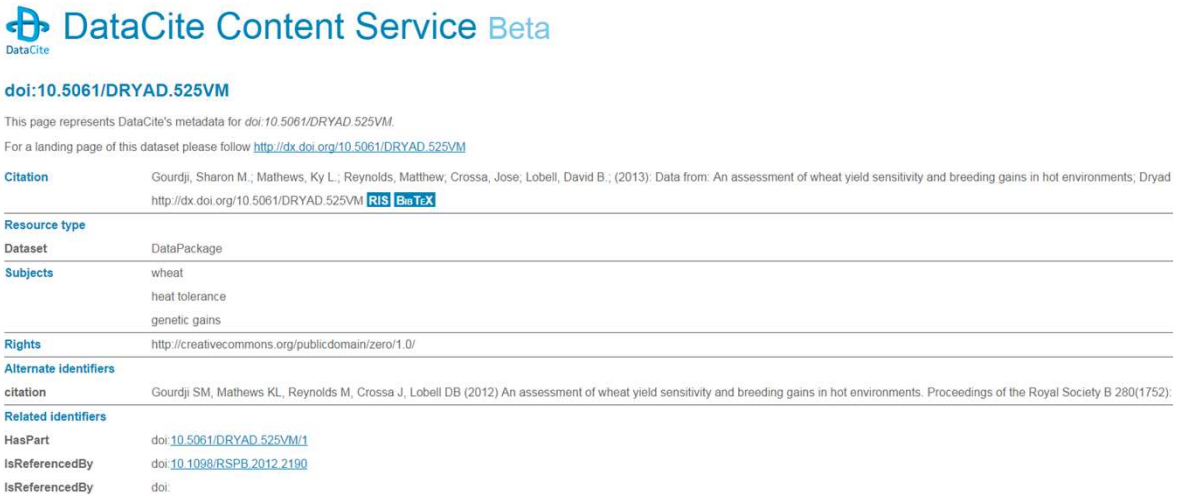
# Exemples de services ou projets

# Portail : annuaire et entrepôt de données

# Attribution des DOI – identifiants numériques

- Objectif : donner des identifiants pérennes aux jeux de données pour qu'ils soient citables, trouvables …
  - ✓ Ex : **10.5061/DRYAD.525VM**

- Etude des besoins et des modalités d'attribution
  - ✓ plateformes / individus
  - ✓ granularité
  - ✓ historisation



DataCite Content Service Beta

**doi:10.5061/DRYAD.525VM**

This page represents DataCite's metadata for *doi:10.5061/DRYAD.525VM*.

For a landing page of this dataset please follow http://dx.doi.org/10.5061/DRYAD.525VM

| | |
|---|---|
| **Citation** | Gourdji, Sharon M.; Mathews, Ky L.; Reynolds, Matthew; Crossa, Jose; Lobell, David B.; (2013): Data from: An assessment of wheat yield sensitivity and breeding gains in hot environments; Dryad http://dx.doi.org/10.5061/DRYAD.525VM RIS BibTeX |
| **Resource type** | |
| **Dataset** | DataPackage |
| **Subjects** | wheat<br>heat tolerance<br>genetic gains |
| **Rights** | http://creativecommons.org/publicdomain/zero/1.0/ |
| **Alternate identifiers** | |
| **citation** | Gourdji SM, Mathews KL, Reynolds M, Crossa J, Lobell DB (2012) An assessment of wheat yield sensitivity and breeding gains in hot environments. Proceedings of the Royal Society B 280(1752): |
| **Related identifiers** | |
| **HasPart** | doi:10.5061/DRYAD.525VM/1 |
| **IsReferencedBy** | doi:10.1098/RSPB.2012.2190 |
| **IsReferencedBy** | doi: |

# Research Data Alliance

- Lancé en mars 2013 par Commission européenne, NSF, Australie
- Contribution à la création de 2 groupes (intérêt, travail)

RDA
RESEARCH DATA ALLIANCE

The Research Data Alliance aims **to accelerate and facilitate research data sharing and exchange**

## Agricultural Data Interoperability IG

Status: Recognised & Endorsed

The Agricultural Data Interest Group is a domain oriented interest group to work on all issues related to data important for the development of global agriculture. The interest group aims to represent all stakeholders producing, managing, aggregating, sharing and consuming data for agricultural research and innovation.

## Wheat Data Interoperability WG

The Wheat Data Interoperability Working Group aims to provide a common framework for describing, representing linking and publishing Wheat data with respect to open standards.

Co chair : E. Dzalé (IST Inra), R. Fulss (CYMMIT)

Wheat Initiative

# Wheat Data Interoperability (WDI) WG

- Status
  - ✓ Recognized and endorsed by the Research Data Alliance (RDA) – March 2014
  - ✓ Part of the Wheat Initiative Information System project
- Focus:
  - ✓ The WG aims to provide a common framework for describing, representing linking and publishing Wheat data with respect to open standards.
  - ✓ The WG will focus first on the following data types: SNP, Genomic annotations, Phenotypes, Genetic Maps, Physical Maps, Germplasm, expression data.

# Réalisation d'une enquête

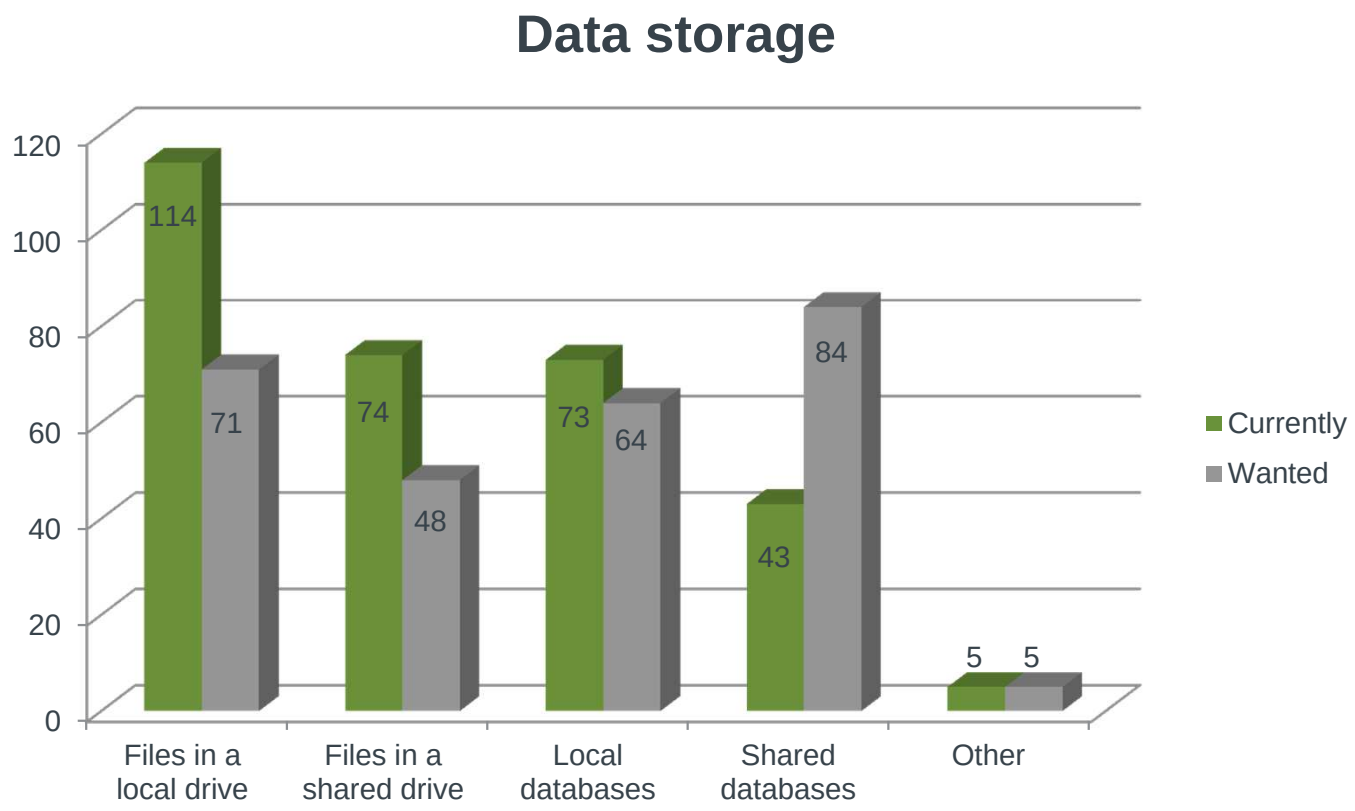
- "Data management and data standards in the wheat research community" Wheat data interoperability WG" (RDA, 2014)

  Objectives:

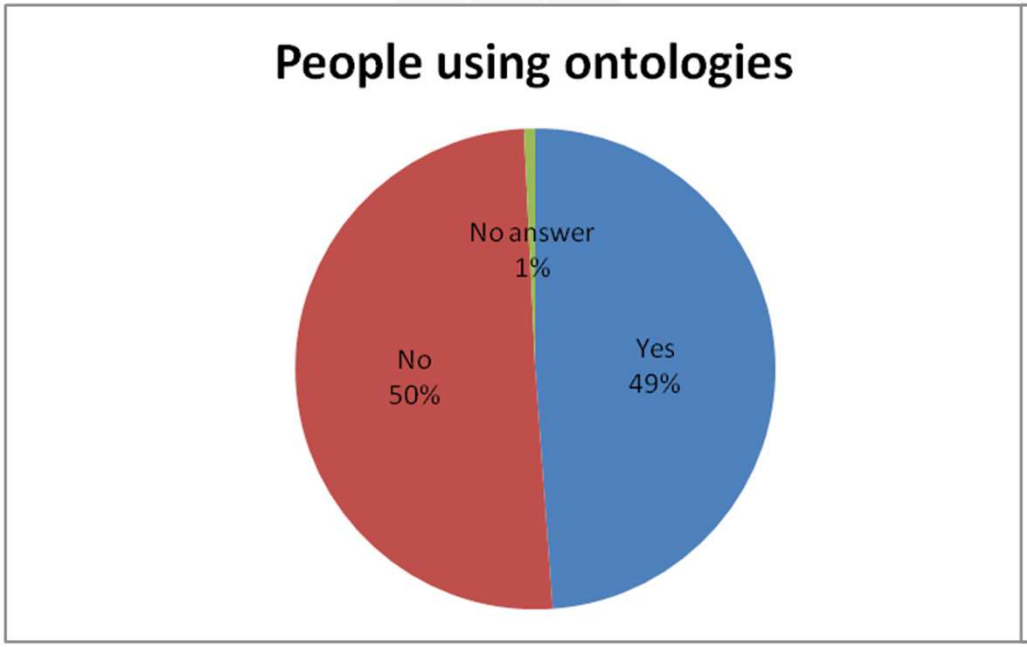
  - To focus on two main pillars of the semantic interoperability :
    - Data Structures
    - Controlled vocabularies, ontologies
  - To identify:
    - Use of common metadata and ontologies
    - Use of standards and formats
    - Level of accessibility
    - Level of interoperability or data exchange
    - Case study

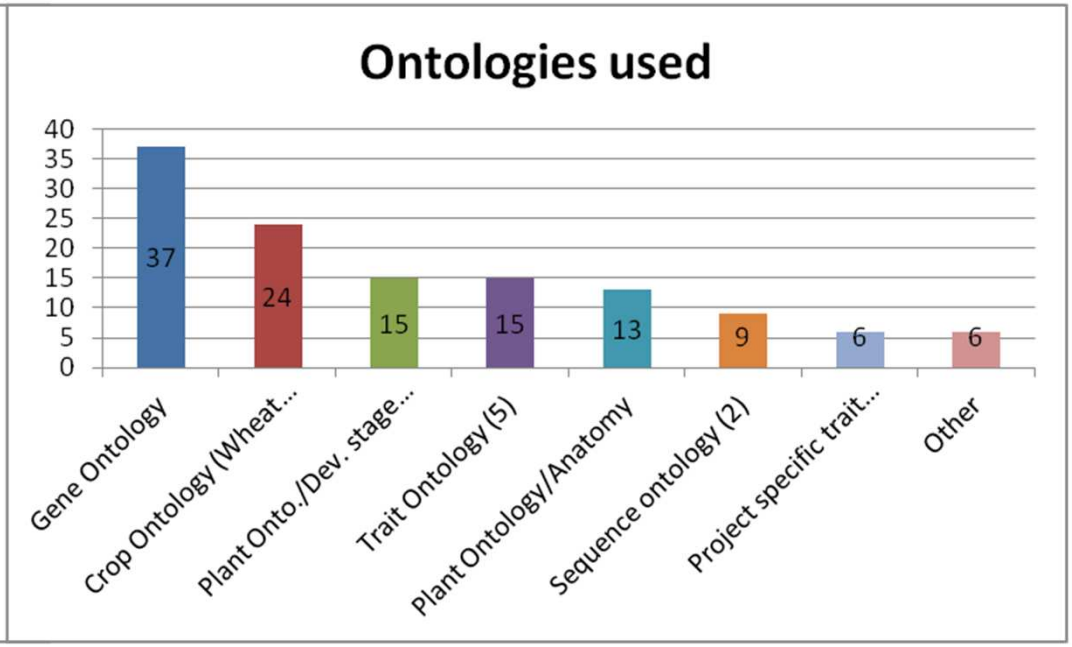
# Survey results – Data storage practice

**114 of the196 respondents currently store their data on local drives; 84 are willing to use shared databases and repositories.**



# Survey results – Ontologies




**Why not ?**

- Lack of knowledge (don't know, too difficult, etc.)
- Lack of trust (lots of talk about their development, but little/no implementation, no agreement, standards, incomplete)
- Lack of interest (not useful)
- In progress.
- No need/required/relevant
- Too complicated
- Not available

**Other ontologies mentioned are:**

- ECPGR
- Ontologies to develop conceptual ABM
- PATO, XEML
- Plant Environmental conditions ontology
- plant pathogens:: http://www.pathoplant.de/; PLEXdb;
- QUDT

# Programme National de Développement Agricole et Rural, objectif 3 // GIS Relance Agronomique

Trois objectifs opérationnels sont fixés :

1. Favoriser le repérage, la production et la diffusion d'innovations sur les systèmes et les modes d'organisation
2. Construire des dynamiques territoriales innovantes en multipartenariat
3. Optimiser la production, la diffusion et l'usage des données (références), des méthodes, des outils et des résultats

- ***chantier n°1 :*** *définir et mettre en œuvre une stratégie de capitalisation des données et résultats produits dans le cadre des financements attribués au titre du PNDAR. Ceci comprend la mutualisation des bases de données de référence, l'harmonisation des méthodes d'évaluation multicritères de la durabilité, l'interopérabilité des systèmes d'information dans un souci de facilitation des échanges et d'accessibilité au plus grand nombre possible d'utilisateurs ;*

- ***chantier n°2 :*** *définir et mettre en œuvre une stratégie pour la validation des résultats issus des expérimentations, projets et expériences de terrain et leur diffusion, par des dispositifs existants ou à créer, vers les utilisateurs finaux (agriculteurs , acteurs économiques dans les territoires), les opérateurs intermédiaires (acteurs du développement agricole, Etat, …) ainsi que l'enseignement agricole en intégrant à la fois les aspects techniques, économiques et sociologiques.*

# Créer un data journal

Open Data Journal for Agricultural Research

- Open access, basé sur OJS
- Partenariat international, sans éditeur privé

AgMIP cimsans WAGENINGEN UR For quality of life MACSUR Global Yield Gap Atlas

Agricultural research uses and produces many relevant data sets in studying agricultural systems across the globe, through its efforts in investigating conditions of global food (in)security at different spatial scales (from regional to national to continental. These data sets have a value to the specific research as these are analysed and investigated, leading to results and conclusions, that are published in peer-reviewed scientific journals or presented at scientific conferences. These data have a longer term value as a resource for the future than the specific research in which they are collected. Other researchers or experts can use these data in new analysis, meta-analysis, or different applications of modelling or statistical tools, leading to new insights for the future. The Open Data Journal for Agriculture Research (ODjAR) acts as a central hub for storing, curating and publishing the data sets as a resource for the future where publications and their authors get appropriate credit through citations and digital object identifiers for future reference.

Many different data sets exist, that are of value and deserve accreditation: experimental data, surveys, model inputs, model outputs, derived indicators and statistics, data assimilation and mark-ups, maps, measured data points. Unlike journal articles describing the main new insights and the most important lessons learned, these data sets are often lost when the funding period ends or the research is published, leading to a situation where these are difficult to reuse for other purposes, or difficult to re-use in reproducing the results described. With the advance of Open Access, Linked Open Data and Open data portals of governments, there is increasing awareness of the value of sharing data with others for further investigation, increased innovation, creation of jobs and better services. Also, governments and science funders are increasing their pressure for science to open up its data, as it is paid with tax-payer financial resources, and should thus have a public benefit.

F1000Research
Open for Science

Articles Collections For Authors For Referees About Blog

F1000Research »

## Article Collections

An *F1000Research* Article Collection relates to a specific community, institution, academic society or conference and can be personalized for the relevant community it serves, and cited as a whole.

Following publication, all articles undergo open peer review by invited referees (see our publishing model), and those that pass peer review are indexed in PubMed and other significant bibliographic databases.

- Open Knowledge for Agricultural development
- Collection F1000

# Ethique du partage des données

- Commande des présidents Inra et Cirad au comité d'éthique
- calendrier : 2014-2015 (lancement 4/7/14)
- Quelques facettes :
  - ✓ anonymisation des données personnelles
  - ✓ éthique du partage
  - ✓ conséquences d'une « data-driven » science
  - ✓ crowdsourcing, citizen science : retour vers le citoyen

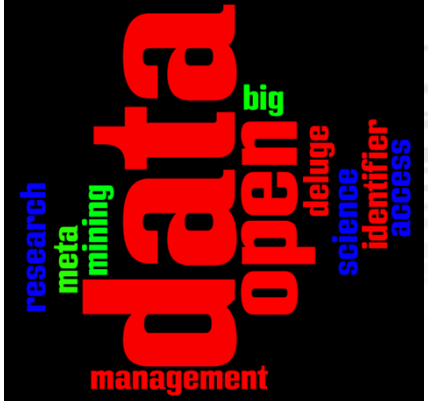
Les directions des deux établissements s'interrogent aussi sur la démarche éthique et déontologique à mettre en place pour assumer cette nouvelle forme indicible de « pouvoir ».

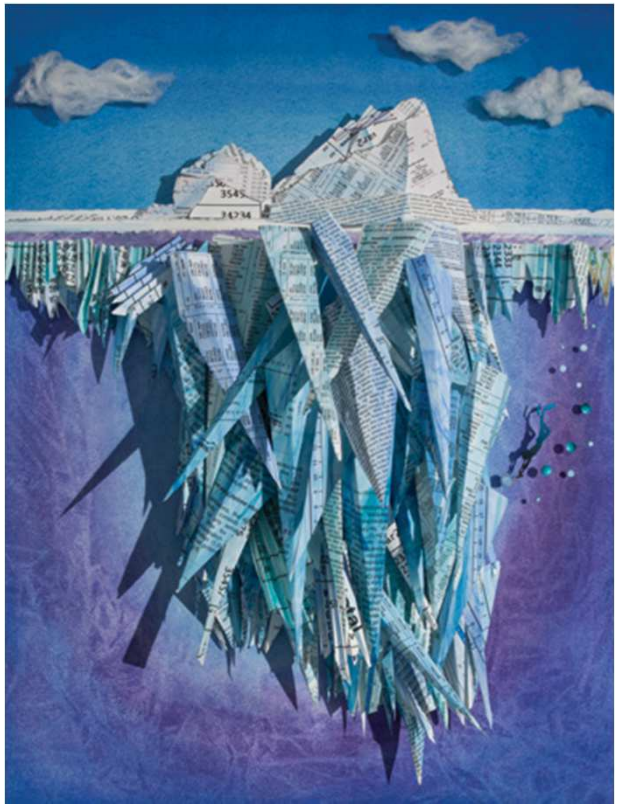
# Pour conclure

- De nombreuses incitations au partage ou à l'ouverture :
  - ✓ H2020
  - ✓ Publications
  - ✓ Demande des « usagers »
- Ne pas le faire n'importe comment
  - ✓ Règles éthiques et juridique
  - ✓ Aspects techniques
  - ✓ Moyens nécessaires
- Nécessité d'inclure la dimension « partage » dans les SI pour être capable de la gérer correctement :
  - ✓ identifiant, métadonnées, protocoles
  - ✓ interopérabilité : normes, standards
  - ✓ historisation
  - ✓ droit d'accès
  - ✓ embargo
- Situer notre action à différentes échelles :
  - ✓ Inra et ses partenaires
  - ✓ National
  - ✓ International

# Colloque dans le cadre de l'IAVFF

- open data et agriculture : quels leviers de croissance ?
  - ✓ impacts sur : Recherche, formation, développement, innovation
  - ✓ création de valeur, modèles économiques :
    - Données de la recherche vers l'entreprise
    - Données de l'entreprise vers la recherche
- semaine du 8 juin



# Merci pour votre attention…

odile.hologne@versailles.inra.fr
@Holo_08

twitter