Enjeux des données de la recherche pour l'INRA

C. Gaspin[1] & O. Hologne[2]

[1]Département de Mathématique et Informatique Appliques – INRA Toulouse
[2]Délégation Information Scientifique et Technique – INRA Versailles

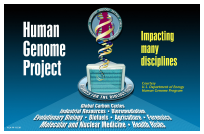
JAO – Paris, 14 Octobre 2015

Plan

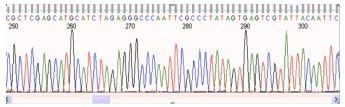
- Un peu d'histoire
- Contexte actuel
- Enjeux des données pour la recherche à l'INRA
- Conclusion

Un peu d'histoire…

L'apport du numérique en Biologie



- Lancement du Projet « Génome humain »
- Naissance du logiciel BLAST (Altschul et al., 1990)
- Entrepôts internationaux



Mise au point du séquençage de l'ADN par Sanger

1990

1988

Séquençage de la première protéine par Sanger

1974

Création du réseau EMBnet et début de l'utilisation d'internet par la recherche au niveau mondial

1970

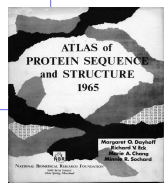
1965

Needleman & Wunsch proposent le premier algorithme d'alignement de séquences

1955

```
RBP:          26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVA 59
                  + K++ + ++  GTW++MA   +    L +   A
glycodelin:   23  QTKQDLELPKLAGTWHSMAMA-TNNISLMATLKA 55
```

M. Dayhoff publie un atlas de séquences protéiques



Le cas des séquences biologiques

Dès les années 90, des entrepôts internationaux

- *Organisation en miroir*
  - Collaboration internationale
  - Mise à jour quotidienne
- *Pour chaque entrepôt*
  - Format de soumission propre
  - Outils de recherche

- *Condition pour publication dans les revues du domaine*
  - Dépôt des séquences dans l'une des bases de données avant publication
  - Attribution d'un identifiant référencé dans la publication

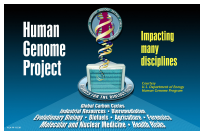
En résumé...

Dès les années 90

- Production collective et partage des données biologiques en vue de l'exploitation dans le cadre de consortiums internationaux
- Séquences biologiques stockées et référencées dans des entrepôts internationaux
- En accès libre pour ré-utilisation pour l'exploitation dans un objectif de revalorisation ou de valorisation de ses propres données

Contexte

Ouverture des données marquée par l'explosion des technologies d'acquisition de données et du numérique



- Lancement du Projet « Génome humain »
- Naissance du logiciel BLAST (Altschul et al., 1990)
- Entrepôts internationaux

2015

2008

1990

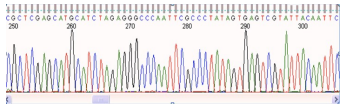
1988

1974

1970

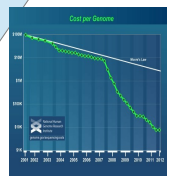
1965

1955

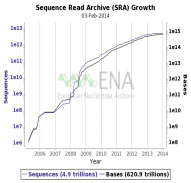


Mise au point du séquençage de l'ADN par Sanger

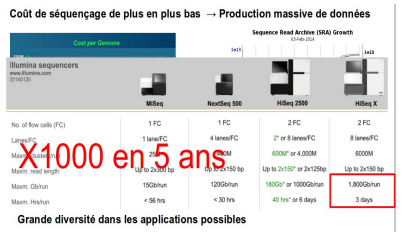
Séquençage de la première protéine par Sanger



Entrée de la génomique dans les « Big Data »



Création du réseau EMBnet et début de l'utilisation d'internet par la recherche au niveau mondial

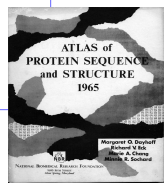


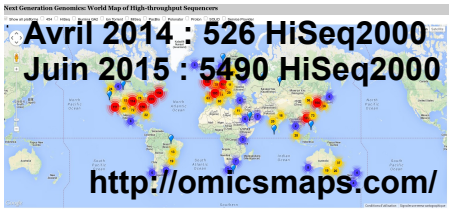
X1000 en 5 ans

Needleman & Wunsch proposent le premier algorithme d'alignement de séquences

```
RBP:          26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVA 59
                  + K++ + ++  GTW++MA   +    L +  A
glycodelin:   23  QTKQDLELPKLAGTWHSMAMA-TNNISLMATLKA 55
```

M. Dayhoff publie un atlas de séquences protéiques





Tout n'est pas qu' « omique » !!!

IN *SCIENCE*

INTRODUCTION

Challenges and Opportunities
Science *Staff*

PERSPECTIVES

Climate Data Challenges in the 21st Century
J. T. Overpeck et al.

Challenges and Opportunities of Open Data in Ecology
D. D. Reichman et al.

Changing the Equation on Scientific Data Visualization
P. Fox and J. Hendler

Challenges and Opportunities in Mining Neuroscience Data
H. Akil et al.

The Disappearing Third Dimension
T. Rowe and L. R. Frank

Advancing Global Health Research Through Digital Technology and Sharing Data
T. Lang

More Is Less: Signal Processing and the Data Deluge
R. G. Baraniuk

Ensuring the Data-Rich Future of the Social Sciences
G. King

Metaknowledge
J. A. Evans and J. G. Foster

Access to Stem Cells and Data, Research Property Rights, and Scientific Progress
D. H. Matthews et al.

On the Future of Genomic Data
S. D. Kahn

NEWS FOCUS

Rescue of Old Data Offers Lesson for Particle Physics
A. Curry

Is There an Astronomer in the House?

Ecologie

Santé

Sciences sociales

Génomique

Microsoft Research

Search Microsoft Research

Home | Our Research | Connections | Careers | Hub
About Us | Research in Action | Opportunities | Research Accelerators

> Collaboration > The Fourth Paradigm: Data-Intensive Scientific Discovery > Table of Contents

Table of Contents

The Fourth Paradigm: Data-Intensive Scientific Discovery
Edited by Tony Hey, Stewart Tansley, and Kristin Tolle

This book is currently available as separate PDF files for online reading or downloading. Select chapters or individual essays below or get the full-text version of the book.

Introductions

- Foreword *Gordon Bell*
- Jim Gray on eScience: a transformed scientific method
  Edited by Tony Hey, Stewart Tansley, and Kristin Tolle

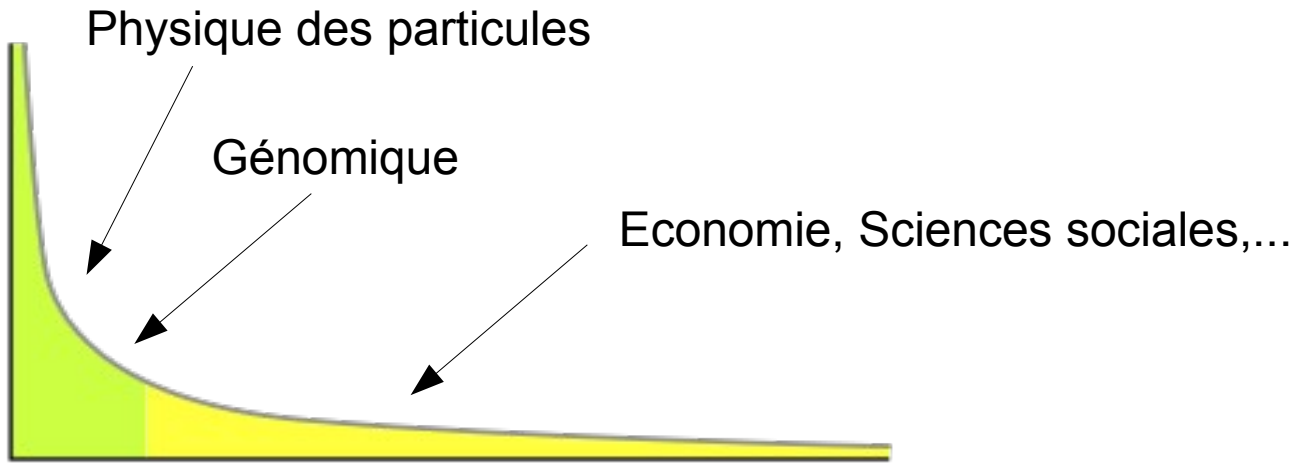
Part 1: Earth and Environment

All Essays in Part 1

- Introduction *Dan Fay*
- Gray's laws: database-c
  Alexander S. Szalay, Jos
- The emerging science of environmental applications
  Jeff Dozier, William B. Gail
- Redefining ecological science using data
  James R. Hunt, Dennis D. Baldocchi, Catharine van Ingen
- A 2020 vision for ocean science
  John R. Delaney, Roger S. Barga
- Bringing the night sky closer: discoveries in the data deluge
  Alyssa A. Goodman, Curtis G. Wong
- Instrumenting the earth: next-generation sensor networks and environmental science
  Michael Lehning, Nicholas Dawes, Mathias Bavay, Marc Parlange, Suman Nath, Feng Zhao

Environment

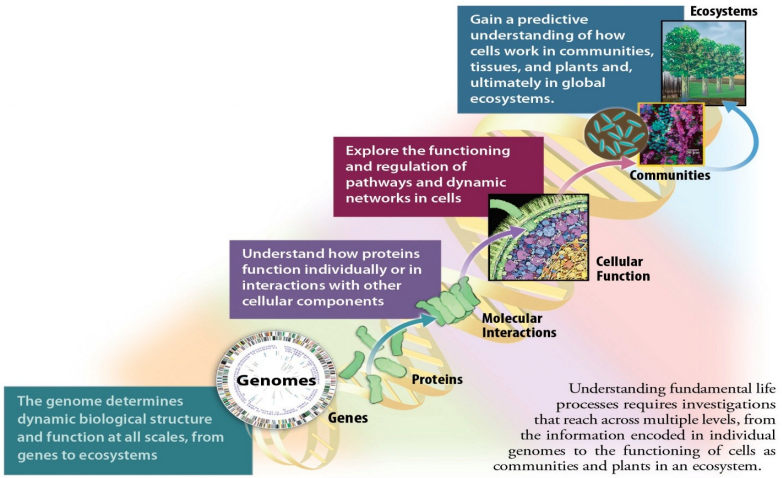
Tout n'est pas que «Big data » !!!

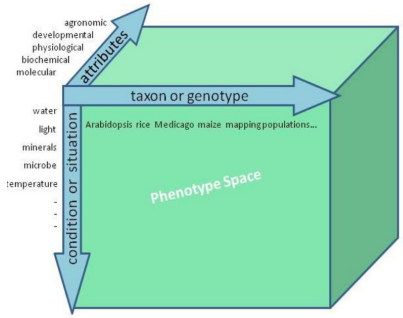


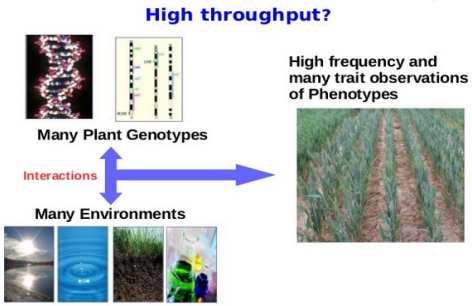
Règle des 20/80 : 20 % des projets génèrent 80 % des données mais pas 80 % de Connaissances !!!

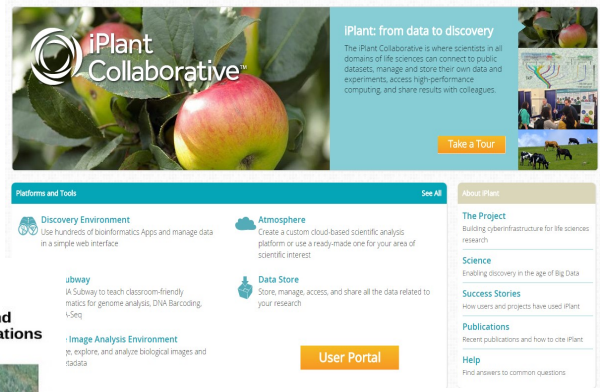
Accessibilité de l'intégration du Vivant ?

- **Intégrer les différents niveaux de complexité**
- **Approches nécessairement pluridisciplinaires et collaboratives !!!**









P. Neveux, Projet Phénome, Paris, 2015

La donnée : un produit de la recherche référençable, citable et évaluable

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]

En résumé...

- Des technologies d'investigation diversifiées et complémentaires permettant des approches globales et intégratives pour appréhender la complexité des mécanismes du vivant
- Approches nécessairement pluridisciplinaires et collaboratives
- Des masses de données considérables et dispersées
- Accélération sans précédent dans l'acquisition de données
- La donnée devient un produit de la recherche référençable, citable et évaluable
- Une accélération de l'accès aux données et aux connaissances par les technologies du numérique

Enjeux des données de la recherche pour l'INRA

- Enjeux en interne
- Enjeux en externe

Enjeux en interne

- Quatre enjeux majeurs en interne à l'INRA
  - Acquisition de données au meilleur niveau de qualité
  - Gestion des données en lien avec la diversité et le cycle de vie des données
  - Analyse des données pour intégrer la complexité des niveaux du Vivant
  - Partage des données

Enjeux en interne

- Quatre enjeux majeurs l'INRA
  - Acquisition de données au meilleur niveau de qualité
  - Gestion des données en lien avec la diversité et le cycle de vie des données
  - Analyse des données pour intégrer la complexité des niveaux du Vivant
  - Partage des données

Partage des données à l'Inra : étapes clés

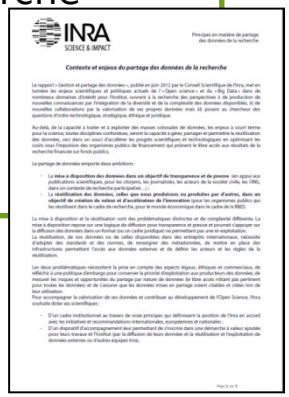
2011-2012 : le CS instruit la question

- 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

Fin 2015 : premiers livrables

Recommandations : fil rouge des actions

RESUME

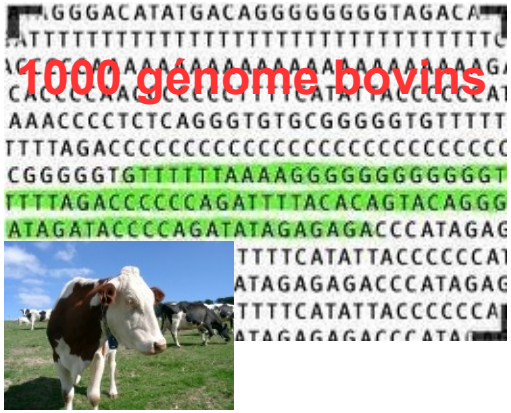
Depuis quelques années, la biologie et les sciences humaines et sociales font face à un accroissement exponentiel des données, provenant de l'adoption en masse des nouvelles technologies, et du développement des sciences et techniques de l'information d'une ampleur et à une échelle sans précédents. Une 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.

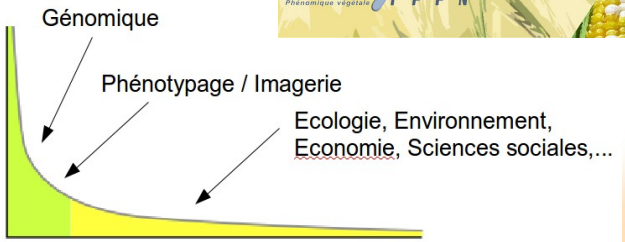
Au terme de son analyse, le groupe de travail propose les recommandations suivantes :

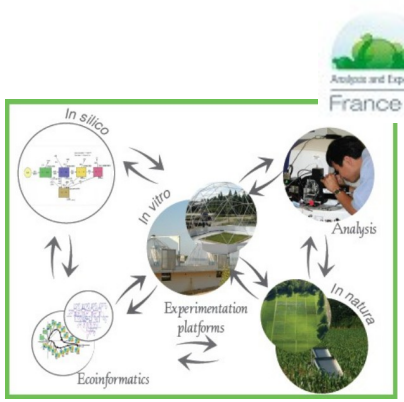
1) Définir la politique de l'établissement et la communiquer.
2) Mettre en place un comité d'évaluation des données produites par l'Inra.
3) S'impliquer dans les comités internationaux de standardisation.
4) Développer un portail d'accès à un ensemble de ressources distribuées.
5) Prendre en compte le 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 ?

- Trois grands ensembles de données à l'INRA pour l'agriculture, l'alimentation, l'environnement :
  - Données 'omiques ' et ressources génétiques
  - Données d'observation et d'expérimentation
  - Données textuelles et d'enquêtes
- Du génome à l'écosystème







Questions étudiées par groupe disciplinaire

- Ecosystème informationnel (entrepôts internes ou externes, data journals ...., vocabulaires ...) : description et analyse
- Identification du Patrimoine historique à pérenniser
- Créneaux de données originales pour se positionner à l'international – une vision stratégique du partage
- Attentes en terme d'offre Inra
- Projets pilotes ?
  - Test de l'offre de service
  - Exploitation des données (dont aspects juridiques ...)

Besoin de nouveaux services

- Annuaire : identifier le patrimoine numérique Inra sur les données (plus généralement, sur les SI Inra)
- Entrepôt : stocker métadonnées & jeux de données
- Plan de gestion de données
- Attribution de DOI
- Gérer l'ouverture des SI Inra en place aux plans technique et juridique
  - Interopérabilité
  - Droit d'accès
  - API
  - ....

Exemple de livrable

Plan de gestion des données

Table des matières

- Pourquoi rédiger un plan de gestion ?
- Comment rédiger un PGD ?
  - Éléments types d'un PGD
    - Informations générales
    - Présentation succincte du Projet de recherche
    - Description des données, métadonnées
    - Stockage et sauvegarde des données au cours du projet
    - Archivage et conservation des données à long terme (après la fin du projet)
    - Droits de propriété intellectuelle
    - Éthique et confidentialité
    - Accès et partage des données à l'issue du projet
  - Guides et outils d'aide à la rédaction des plans
    - Guides
    - Formulaires en ligne
  - Exemples de plans rédigés

https://wiki.inra.fr/wiki/donneesrechercheist/Main/PlanGestion

Exemple de livrable
Plan de gestion des données

Table des matières

Pourquoi rédiger un plan de gestion ?

"L'ouverture des données ne peut constituer un point de départ, elle ne doit être envisagée que comme le résultat nécessaire d'une bonne politique de gestion de données, qui en constitue le préalable indispensable".[1]

Un plan de gestion des données est un document officiel, à établir au démarrage d'un projet de recherche et qui décrit la façon dont les données seront gérées pendant la phase de recherche et une fois le projet terminé. L'objectif est de décrire finement la nature des données, les méthodes d'obtention, et d'examiner les différents aspects de gestion des données, création de métadonnées, conservation des données immédiate et à plus long terme dans le but de faciliter leur compréhension et leur éventuelle ré-exploitation.

Les plans de gestion (data management plan ou DMP) "constituent le pendant concret d'une politique de données" [2]. Ils peuvent prendre des formes très variées selon les disciplines scientifiques, le type de données produites...

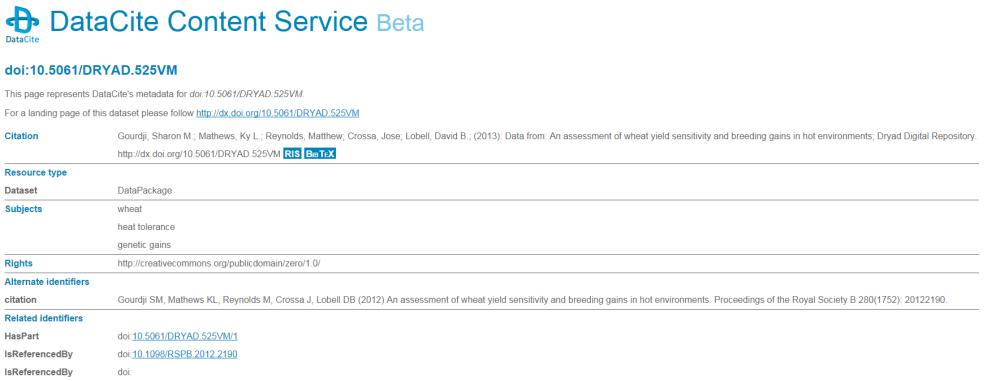
Les principaux financeurs de la recherche souhaitent aujourd'hui la mise à disposition des publications et des données de recherche associées aux travaux qu'ils financent. Les dossiers de candidatures pour l'obtention de financement doivent de plus en plus souvent intégrer des plans de gestion et de partage des données qui comprennent les exigences des financeurs, complétées éventuellement de celles des institutions de recherche.

Établis au démarrage du projet, les plans doivent intégrer un cycle de vie des données qui déborde la durée du projet puisque l'enjeu est de permettre la valorisation et l'accessibilité aux données produites une fois le projet terminé.

Attribution des DOI

Faciliter l'archivage et l'accès aux ressources 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 Digital Repository. http://dx.doi.org/10.5061/DRYAD.525VM RIS BibTeX
Resource type	
Dataset	DataPackage
Subjects	wheat heat tolerance 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): 20122190.
Related identifiers	
HasPart	doi:10.5061/DRYAD.525VM/1
IsReferencedBy	doi:10.1098/RSPB.2012.2190
IsReferencedBy	doi:

Guide juridique

Faciliter l'archivage et l'accès aux ressources numériques

- Groupe inter-organismes

Préambule

…
« A défaut d'un cadre légal clair sur la question de l'Open Data, ce guide a pour vocation d'**accompagner les agents des établissements concernés dans une démarche d'ouverture raisonnée des données de recherche** en tentant de répondre aux questions pratiques les plus courantes auxquelles ils pourront être confrontés. »

Ouverture des données de recherche

Guide pratique du chercheur

Le présent guide est issue des réflexions d'un groupe de travail inter-organismes. Il ne prétend pas à l'exhaustivité et ne peut engager la responsabilité de leurs auteurs. Il est fourni uniquement à titre d'information. Il ne saurait en tout état de cause se substituer à la lecture des dispositions législatives, règlementaires et de la jurisprudence applicables en la matière. Ce guide peut être amené à évoluer.

Membres du groupe de travail : AMZIANE Mehdi (Inria), BOURCIER Danièle (CNRS), CASTETS-RENARD Céline (UT1), CHASSANG Gauthier (Inserm), COURTOIS Marie-Audrey (Inra), DANTANT Martin (CNRS), GALLON Claire (Libertic), GANDON Nathalie (**co-animatrice**, Inra), GARDIN Timothée (Inra), MARTIN Caroline (Irstea), MARTELLETTI Andrea (stagiaire Inra, M2 droit et informatique), MENDOZA-CAMINADE Alexandra (UT1), MORCRETTE Nathalie (**co-animatrice**, Inra), NEIRAC claire (Cirad), STAMBOLIYSKA Rayna (UPD), VERTOT Manuelle (Anses).

Guide retravaillé par MARTELLETTI Andrea dans le cadre de son stage de fin d'études de M.2 Droit et informatique.

Portail / Annuaire / Entrepôt institutionnel ?

- En cours d'instruction en lien avec les acteurs du groupe + DSI
  - Besoins exprimés par les groupes “données”
  - Fonctionnalités prioritaires ?
    - Stocker / Gérer / Préserver les données associées aux publications
    - Rendre visible et citables les données

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 / Services Web)
- 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
  - A quelle « maille » doit se faire la réflexion stratégique ?

Se mettre en posture d'aller chercher les données ailleurs !

Enjeux en externe

- S'inscrire dans le partage à l'international
  - Rejoindre les comités éditoriaux des revues


Open Data Journal for Agricultural Research

AgMIP cimsans 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.

These recent developments of Open Access to data and the acknowledge value of data archiving lead to four global networks in agricultural and food security research to come together to support the Open Data journal for Agricultural Research:

+ Agricultural Model Improvement and Intercomparison Project (AgMIP)
+ Global Yield Gap Atlas (GYGA)
+ Center for Integrated Modeling of Sustainable Agriculture and Nutrition Security (CIMSANS)
+ Modelling European Agriculture with Climate Change for Food Security (MACSUR)

Financial support and in-kind contributions towards realizing ODJAR is acknowledge from:

+ Center for Integrated Modeling of Sustainable Agriculture and Nutrition Security (CIMSANS), part of ILSI
+ Wageningen University and Research Centre Library
+ United Kingdom Department of International Development (UK DFID)

INRA CGIAR WAGENINGEN UR

  - Participer aux initiatives internationales de mise en place des standards
    - Consortiums
    - Alliances

WHEAT INITIATIVE

AnaEE Analysis and Experimentation on Ecosystems

RDA RESEARCH DATA ALLIANCE

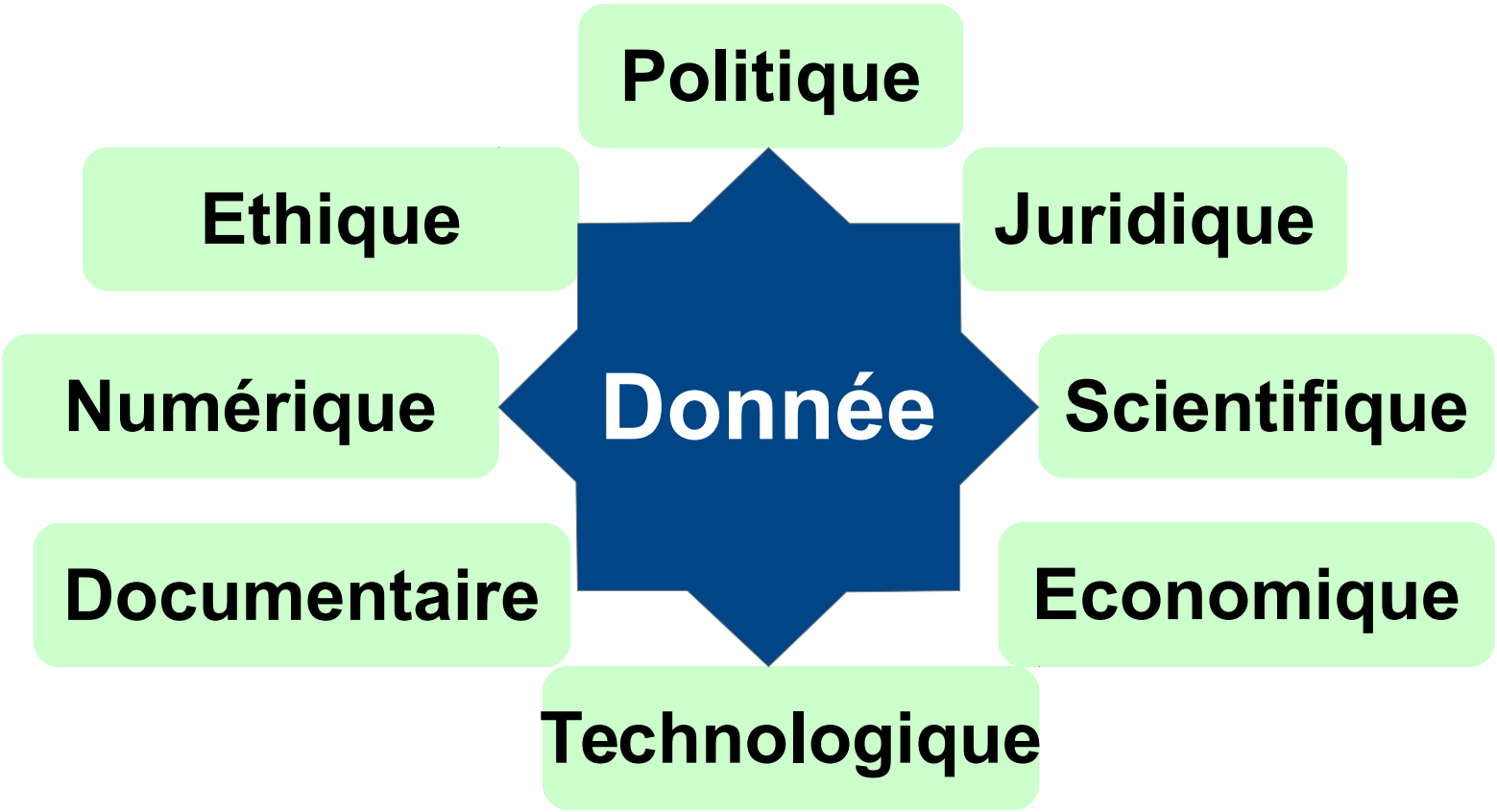
The Research Data Alliance aims **to accelerate and facilitate research data sharing and exchange**

Conclusion

Points non abordés

- Quatre enjeux majeurs l'INRA
  - Acquisition de données au meilleur niveau de qualité
  - Gestion des données en lien avec la diversité et le cycle de vie des données
  - Analyse des données pour intégrer la complexité des niveaux du Vivant
  - Partage des données

Diversité/complexité des acteurs impliqués



Merci de votre attention