Open Science. Gestion et partage des données de la recherche

Windpouire Esther Dzale Yeumo, Dominique L'Hostis

Greg Emmerich
License: CC by SA 3.0

Open Science

Gestion et partage des données de la recherche

Formation URFIST Paris – 22 Janvier 2015
Esther Dzale & Dominique L'Hostis

Plan (1)

- **1- Introduction – Contexte général**
- **2- Qu'est-ce que les données de la recherche ?**
  - ✓ Définitions + exemples
  - ✓ Différents états des données (brutes, élaborées, primaires, secondaires, dérivées, etc..) + exemples
  - ✓ Typologie des données + exemples
- **3- Gestion des données de la recherche**
  - ✓ Cycle de vie des données + questionnements
  - ✓ Plan de gestion + exemples
  - ✓ Documentation
  - ✓ Stockage sécurisé et préservation
  - ✓ DOI
  - ✓ Enjeux d'une bonne gestion des données de la recherche
  - ✓ Difficultés et freins, etc.

Plan (2)

❖ 4- Diffusion des données de la recherche

- ✓ Les différents modes de diffusion
  - Déposer dans des entrepôts + exemples et choix d'entrepôts
  - Publier des données comme matériel supplémentaire à des articles
  - Publier des Data papers
  - Publier des données dans le Web de données
- ✓ La citation des données

❖ 5- Partage des données de la recherche

- ✓ Enjeux du partage (pourquoi partager ? ou ne pas partager ?)
- ✓ Comment partager (Quelles données partager ? sous quelles conditions ? Avec quelles licences ?)
- ✓ Freins et leviers
- ✓ Etat des lieux

Plan (3)

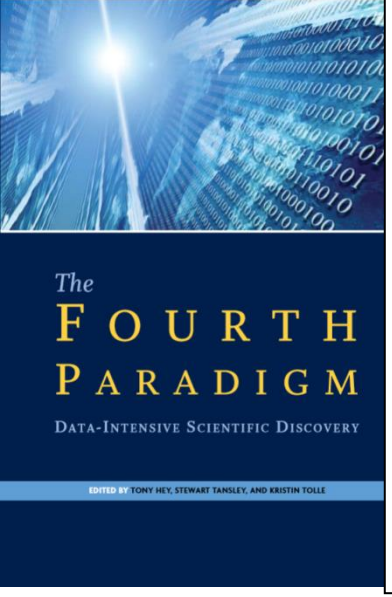
- **6- (Ré)Utilisation des données de la recherche**
  - ✓ Trouver des données (annuaires et entrepôts)
  - ✓ Evaluer la qualité des données (critères, exemples)
  - ✓ Exemples de réutilisation
- **7- Données de la recherche et IST**
  - ✓ Rôles possibles
  - ✓ Compétences mobilisables, à développer ?
  - ✓ Engagement et formation des communautés (en France et ailleurs)

_01

Contexte général

PAYSAGE DES DONNÉES DE LA RECHERCHE

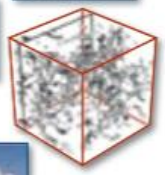
INRA
SCIENCE & IMPACT



Science Paradigms

- Thousand years ago: science was **empirical**
  describing natural phenomena
- Last few hundred years: **theoretical** branch
  using models, generalizations
- Last few decades: a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



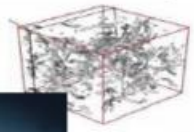
Jim Gray on e-science :
A Transformed Scientific Method
(Hey, Tanslay, & Tolle, 2007)

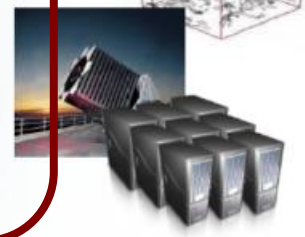
Emergence of a Fourth Research Paradigm

1. Thousand years ago – **Experimental Science**
   - Description of natural phenomena
2. Last few hundred years – **Theoretical Science**
   - Newton's Laws, Maxwell's Equations…
3. Last few decades – **Computational Science**
   - ~~Simulation of complex phenomena~~
4. Today – **Data-Intensive Science**
   - Scientists overwhelmed with data sets from many different sources
     - Data captured by instruments
     - Data generated by simulations
     - Data generated by sensor networks
   - ➢ eScience is the set of tools and technologies to support data federation and collaboration
     - For analysis and data mining
     - For data visualization and exploration
     - For scholarly communication and dissemination

(With thanks to Jim Gray)

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

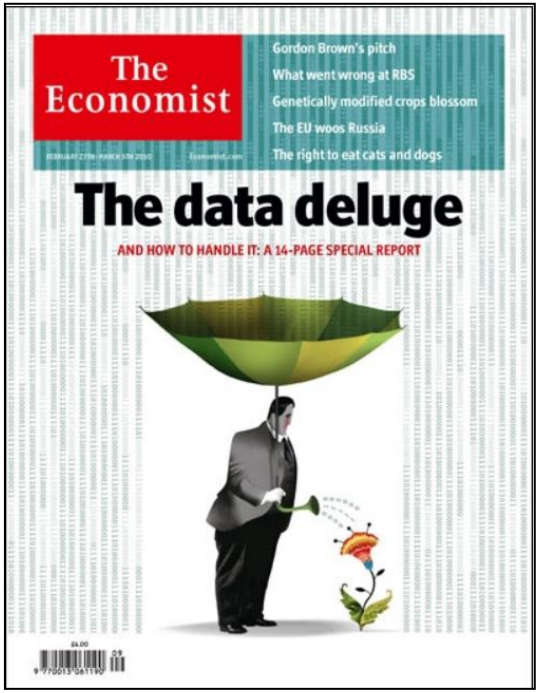




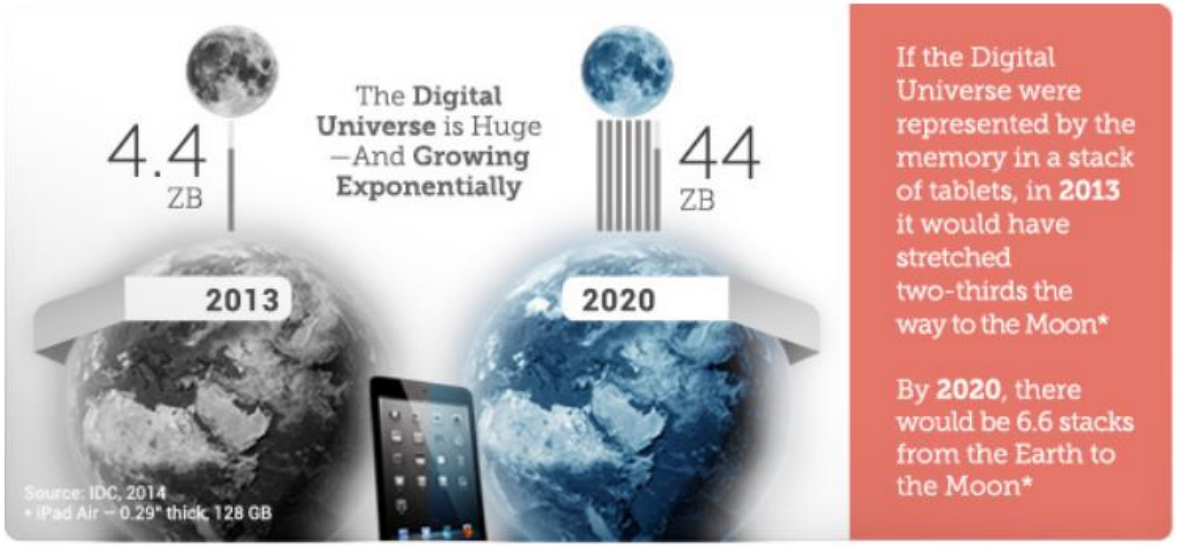
(Andre, 2013)

Explosion du volume de données

- **Explosion du volume** de données numériques produites dans tous les secteurs d'activités



« The Economist » Février 2010
http://www.economist.com/node/21521549



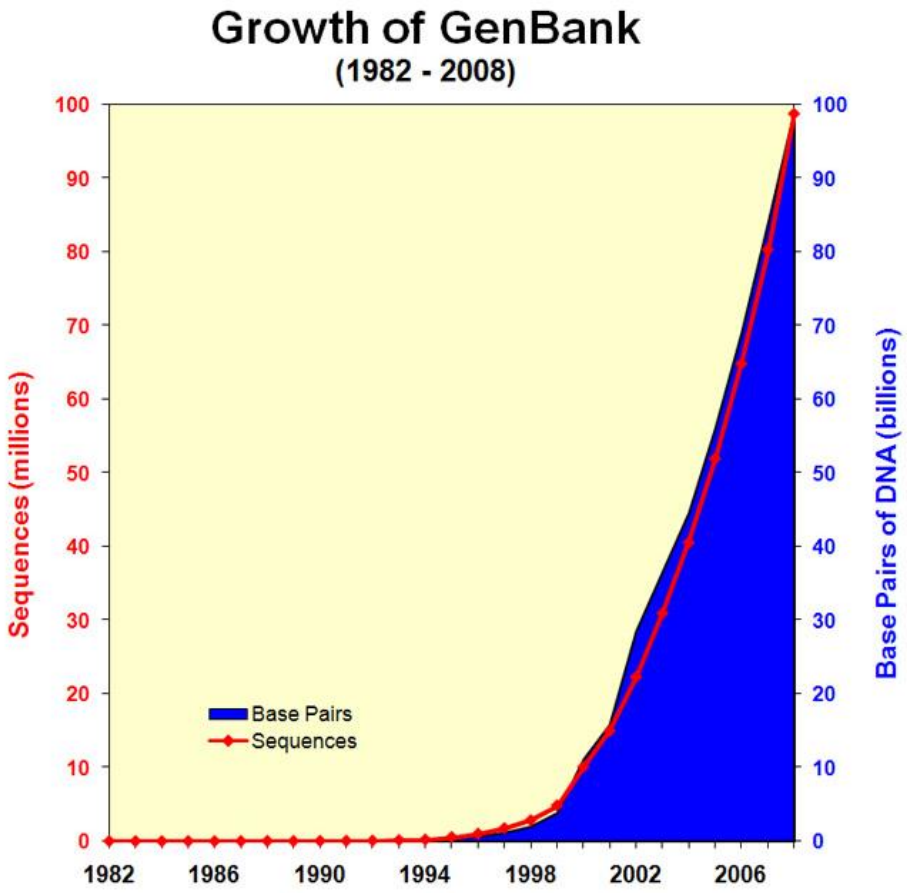
Etude avril 2014 - EMC
http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

L'univers digital double tous les 18 mois

(Abiteboul, 2012)

Inflation des données scientifiques

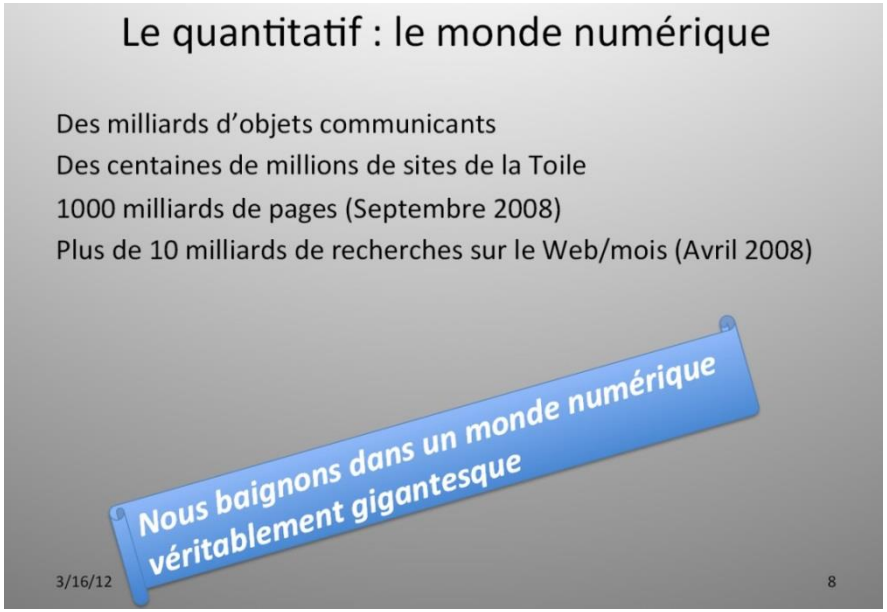
Des disciplines telles que la génomique, l'astronomie, la bioinformatique, la physique par exemple produisent de très grandes quantités de données



Exemple : Genbank
(banque de séquences d'ADN)
La version 205, datée de décembre 2014, contenait plus de 184 milliards de bases de nucléotides dans plus de 179 millions de séquences
http://www.ncbi.nlm.nih.gov/genbank/statistics

Infrastructure technique de plus en plus performante

- qui permet de gérer, distribuer et analyser ces données.



(Abiteboul, 2012)

Partager les données : une nécessité

- Des défis scientifiques complexes , enjeux de société, qui supposent : transdisciplinarité, collaboration des équipes, ouverture et mutualisation des informations, des données et des compétences.
  - ✓ Alimentation : Nourrir 9 milliards d'hommes…
  - ✓ Changement climatique…
  - ✓ Agriculture durable, éco-responsable…
  - ✓ Santé et nutrition…
  - ✓ Biodiversité

Numéro spécial – février 2010
http://www.sciencemag.org/content/331/6018.toc#SpecialIssue

- Logique de production massive de données
  - ✓ rendue possible par la baisse des coûts et l'accroissement des performances technologiques,
  - ✓ mais capacités de stockage potentiellement limitées par les coûts de mise en place des infrastructures de prise en charge…

Partager les données : une nécessité (2)

Le qualitatif : données, informations et connaissances		
Données	Description élémentaire d'une réalité	*Mesures de températures dans une station météo*
Informations	Données avec un sens (pour construire une représentation de la réalité)	*Une courbe donnant l'évolution des minimas & maximas moyens en un lieu suivant le mois de l'année*
Connaissances	Informations avec une vérité, plus généralement une loi qui est considérée comme vraie	*Le fait que la température sur terre augmente du fait de l'activité humaine*

(Abiteboul, 2012)

Impact sur les méthodes de la recherche…

- Changement de perspectives, avec des approches à grande échelle lié aux masses de données disponibles : de « Hypothese Driven » à « Data Driven »

  Exemple en génomique :

  > *« Mode de fonctionnement « resource driven » : résultats obtenus à partir d'une masse de données, sans avoir fait d'hypothèse au départ. C'est le matériau qui fait émerger les hypothèses »* (Piétu, 2010)

- Des **formes de publications spécifiques** aux données apparaissent associées à de nouvelles **politiques éditoriales**.

- La plupart des scientifiques sont sensibles et **favorables** au **principe du partage des données** mais les pratiques réelles sont encore en décalage.

Soutien large de l'ouverture des données Vers l'« Open Science »

- par la plupart des organismes de recherche,
  - De l'Open Access à l'Open Data…
- par leurs financeurs,
  - Union Européenne, NSF, RCUK …
- par les gouvernements
  - Prise de position du G8…
- par des collectifs transversaux ou spécialisés dans une discipline
  - Science Europe,
  - Research Data Alliance
  - Global Open Data for Agriculture and Nutrition (GODAN)

G8+5 International Conference on Open Data for Agriculture April 29-30, 2013

G8 Science Ministers Statement London UK, 12 June 2013

Introduction

We, the G8 Science Ministers met in London on Wednesday 12 June with Presidents of our respective national science academies, as part of the UK's G8 Presidency. At this unique meeting we discussed how our nations could lead efforts to improve the transparency, coherence and coordination of the global scientific research enterprise in order to address global challenges and maximise the social and economic benefits of research.

So today, recognising the role that science has to play in securing present and future sustainable growth, we approved a statement which proposes to the G8 for consideration new areas for collaboration and agreement on global challenges, global research infrastructure, open scientific research data, and increasing access to the peer-reviewed, published results of scientific research.

Sommet Aquila 2012
Engagement politique G8 collaboration sur la sécurité alimentaire

- Working Groups
- Research Data
- Research Infrastructures
- Research Integrity
- Cross-border Collaboration
- Research Careers
- Open Access to Scientific Publications
- Horizon 2020
- Research Policy and Programme Evaluation
- Gender and Diversity Working Group

Working Group on Research Data

The Research Data Working Group brings together experts on research data management from Science Europe Member Organisations. During its initial two-year mandate (2013-2015), the Working Group will act as Science Europe's platform to reflect upon shared challenges related to research data. It will specifically focus on developing policies related to data collection, quality assurance of data, data preservation, and sharing and re-use of research data.

The activities of the group will touch upon topics such as 'Text and Data Mining', a set of techniques which can enhance the potential for analysing research data and enabling new findings, and the concept of 'Big Data', related to the collection of data sets too large and complex for traditional data management tools to process them.

The work of the group is guided by the principle that openness of data should be the standard rule, but can be postponed and limited for legitimate reasons, such as privacy or disciplinary-specific issues.

The Working Group aims to guide Science Europe members, along with other stakeholders in Europe and beyond, in a process to foster a cultural change towards sharing scientific data for the advancement of research and ultimately for the benefit of society at large. In doing so, it will take into account issues related to other policy fields such as: research infrastructures and the necessary development of appropriate data infrastructures for instance; research

- Login Area
- News

Federation of European Academies of Medicine, the Wellcome Trust and the British Heart Foundation to create a European Data in Health Research Alliance. See the website here.

(7 November 2014) Just published: Position Statement

France	French National Research Agency	ANR	Martine Garnier
France	National Institute for Agricultural Research	INRA	Odile Hologne
France	National Centre for Scientific Research	CNRS	Francis André
France	French Alternative Energies and Atomic Energy Commission	CEA	Delphine Vidart-Dufort
France	French Research Institute for Exploitation of the Sea	Ifremer	Jean-François Masset
France	French National Institute of Health and Medical Research	Inserm	Anita Burgun
France	National Institute for Development	IRD	Jean-Pierre Finance

http://www.scienceeurope.org/policy/working-groups/Research-Data

Research Data Alliance

https://rd-alliance.org/

Research Data Sharing without barriers

- Lancée en mars 2013 par Commission européenne, les USA (NSF), l'Australie (ANDS)
- Vise à accélérer et faciliter le partage et l'échange des données scientifiques
- Contribution INRA à la création de 2 groupes

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.

E. Dzalé (IST Inra), R. Fulss (CYMMIT)

Données de la recherche agricole

GODAN
Global Open Data for Agriculture and Nutrition

http://godan.info/

Home | Statement of Purpose | Partners | Success Stories | 2015 Meeting | APPLY NOW - Secretariat

Statement of Purpose

The Global Open Data for Agriculture and Nutrition (GODAN) initiative seeks to support global efforts to make agricultural and nutritionally relevant data available, accessible, and usable for unrestricted use worldwide. The initiative focuses on building high-level policy and public and private institutional support for open data. The initiative encourages collaboration and cooperation among existing agriculture and open data activities, without duplication, and brings together all stakeholders to solve long-standing global problems.

Open access to research, and open publication of data, are vital resources for food security and nutrition, driven by farmers, farmer organizations, researchers, extension experts, policy makers, governments, and other private sector and civil society stakeholders participating in "innovation systems" and along value chains. Lack of institutional, national, and international policies and openness of data limit the effectiveness of agricultural and nutritional data from research and innovation. Making open data work for agriculture and nutrition requires a shared agenda to increase the supply, quality, and interoperability of data, alongside action to build capacity for the use of data by holders.

DAN initiative is a voluntary association brought together around a shared purpose. Launched er 2013, the initiative welcomes all those who share this purpose to join as members and to te in shaping coordinated activities that can deliver on the potential of open data for re and nutrition. Together, initiative partners seek to support this initiative through the guidelines and principles.

ith global movements for open data and open access, the initiative seeks to:

vocate for open data and open access policies by default, in both public and private sectors, ilst respecting and working to balance openness with legitimate concerns in relation to vacy, security, community rights and commercial interests;
vocate for the release and re-usability of data in support of Innovation and Economic Growth, proved Service Delivery and Effective Governance, and Improved Environmental and Social tcomes;

Tweets

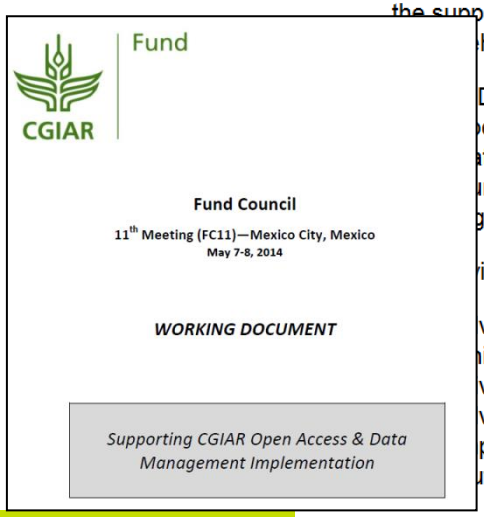
AgroKnow @AgroKnow 17 Sep
Open #AGRIgate Competition: @agINFRA #OpenData Campaign aims.fao.org/community/open #GODAN #SemaGrow pic.twitter.com/SgpVP9fk2g
Retweeted by Laurent Lefort

Open AGRIgate

Expand

AgroKnow @AgroKnow 17 Sep
Open #AGRIgate Competition: @agINFRA #OpenData Campaign aims.fao.org/community/open #GODAN #SemaGrow pic.twitter.com/SgpVP9fk2g
Retweeted by Andreas Drakos

Open



BOUCHOUT DECLARATION

The Bouchout Declaration for Open Biodiversity Knowledge Management

The purpose of the Bouchout Declaration is to help make digital data about our biodiversity openly available. It offers members of the biodiversity community a way to demonstrate their commitment to open science.

Declaration

As signatories, we encourage an overarching approach to Open

News

The latest news and announcements covering the

Sign

Your signature is your commitment to an Open

Les signataires s'engagent à promouvoir un libre accès aux données et aux informations sur la biodiversité par les utilisateurs et les systèmes informatiques. Ils s'engagent également à aboutir à une infrastructure de connaissances large et partagée, afin de permettre à notre société de réagir plus efficacement aux défis présents et futurs.

http://www.bouchoutdeclaration.org/

Bibliographie – Introduction

- Abiteboul, S. (2012). [Sciences des données : de la logique du premier ordre à la toile. Leçon inaugurale au Collège de France (08/03/2012)]. http://www.college-de-france.fr/media/serge-abiteboul/UPL4129881692607880347_lecon_inaugurale.pdf
- Andre, F. (2013). *Les enjeux autour des données de la recherche. Journées FRéDoc, 2013/10/07-10, Aussois - France.* *http://renatis.cnrs.fr/IMG/pdf/fANDRE_AUSSOIS_FREDOC2013.pdf*
- Bartling, S., & Sascha Friesike, S. (2013). *Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*: Springer. http://link.springer.com/book/10.1007/978-3-319-00026-8
- Gantz, J., & Reinsel, D. (2011). Extracting Value from Chaos from http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf
- Hey, A. J. G., Tanslay, S., & Tolle, K. (2007). Jim Gray on e-science : A Transformed Scientific Method (Based on the transcript of a talk given by Jim Gray to the NRC-CSTB1 in Mountain View, CA, on January 11, 2007). from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf
- Tolle, K. M., Tansley, D. S. W., & Hey, A. J. G. (2011). *The Fourth Paradigm: Data-Intensive Scientific Discovery (Vol. 99): Microsoft Corporation. http://research.microsoft.com/en-us/collaboration/fourthparadigm/*
- Piétu, G. (2010). Le projet « Génome humain et l'Open Source » Farchy, J.; Froissard, P.; Méadel, C.- 2010. Sciences.com libre accès et science ouverte. p. 151. In J. Farchy, P. Froissard & C. Méadel (Eds.), *Sciences.com libre accès et science ouverte (pp. 151-152): Hermès.*
- Vincey, C. (2012). [Opendata benchmark - FR vs UK vs US]. http://fr.slideshare.net/cvincey/opendata-benchmark-fr-vs-uk-vs-us

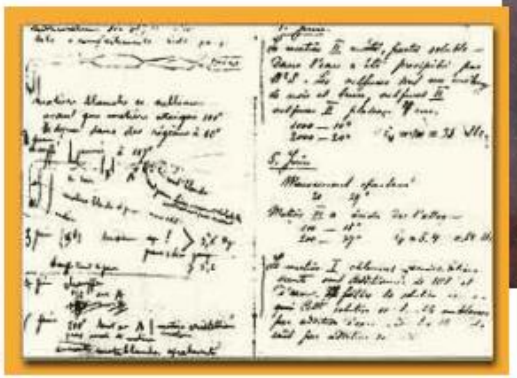
_02

Qu’est ce que les données de la recherche ?

INRA
SCIENCE & IMPACT

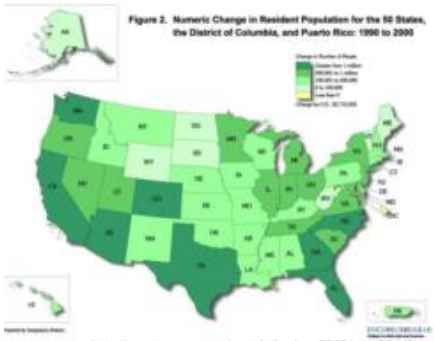
What are data?

NASA Astronomy Picture of the Day

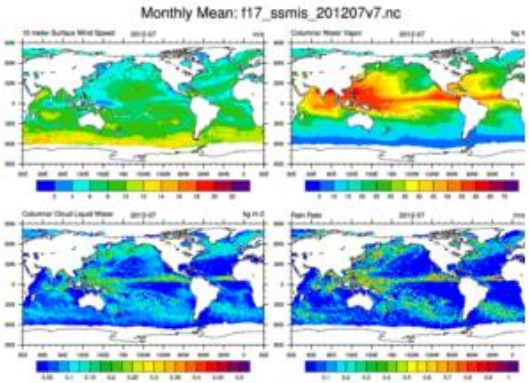


Marie Curie's notebook aip.org

hudsonalpha.org



http://www.census.gov/population/cen2000/map02.gif



ncl.ucar.edu

Date:1/2.07.75 Place:Sakaltutan
Zafor
He will grow old in his present house; new house is for sons - 5 sons. Not sure they want to live in village. He will only build another if they want him to. eS came from Germany and did the plastering. He arranged the carpentry in Kayseri. Çok para gitti. (much money went) Has a tractor.

Date:July1980 Place:Sakaltutan
Zafor:
Household now Zafor and wife; Nazif Unal and wife and youngest son, still a boy. They run two dolmuß; one with a driver from Süleymanli. Goes in and out once a day. He gets 8,000 a month. Zafor then said, keskin deOil. (not sharp - i.e.? not profitable) I said he did very well on 8,000 TL with only two journeys a day. Nazif Unal has "bought" a Durak (dolmuß stop) from Beledive and works all day in Kayseri.

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php

(Borgman, 2014)

Plusieurs définitions des données de la recherche

« Providing an authoritative definition of research data is challenging, as any definition is likely to depend on the context in which the question is asked"

http://ands.org.au/guides/what-is-research-data.html

- La notion de « données de la recherche » est différente selon les disciplines scientifiques
- Le périmètre des données de la recherche varie selon les politiques :
  - *Extrait – Rapport pour le conseil scientifique de l'Inra*
    « En l'absence d'une définition juridique, nous appellerons une donnée scientifique ou donnée de la recherche : information qui représente la matériau de base d'une activité de recherche ayant bénéficié d'un financement sur fonds publics. » (Gaspin & Pontier, 2012)
  - Université de Bristol : tout objet numérique qui résulte d'un travail de recherche. Les documents administratifs sont exclus.
  - Université de Melbourne : aussi bien des objets numériques que d'autres types d'objets, y compris les documents administratifs.

- Le contexte est important : quand, quelle question scientifique, pour quoi, etc… ?

  Exemples (Fayet, 2013)

  - les images d'une ville préhistorique deviennent des données pour un chercheur qui étudie l'histoire de cette ville.
  - les « données » d'un linguiste peuvent être des écrits ou des discours, des enregistrements de locuteurs ;
  - les « données » d'un médiéviste sont des sources archivistiques, archéologiques, épigraphiques, iconographiques, littéraires ;
  - les « données » d'un géologue rassemblent des coupes et observations de terrain consignées sur un carnet, des résultats de carottage, des analyses d'échantillons, des données sismographiques…

What is Data?

"Data are facts, observations or experiences on which an argument, theory or test is based. Data may be numerical, descriptive or visual. Data may be raw or analysed, experimental or observational."

http://research.unimelb.edu.au/integrity/conduct/data/review

May originate from various sources:
Primary and/or secondary

May contain different content:
Quantitative and/or qualitative

May be expressed in different forms:
Datasets, still images, audio-video, audio recordings, interactive resources

May be held in a number of variations:
Raw, cleaned, anonymised/pseudomised, analysed

May be encoded in different formats:
MS Excel, TIFF, MPEG2, STATA, FoxPro

LONDON SCHOOL of HYGIENE &TROPICAL

What type of data do you have at home?

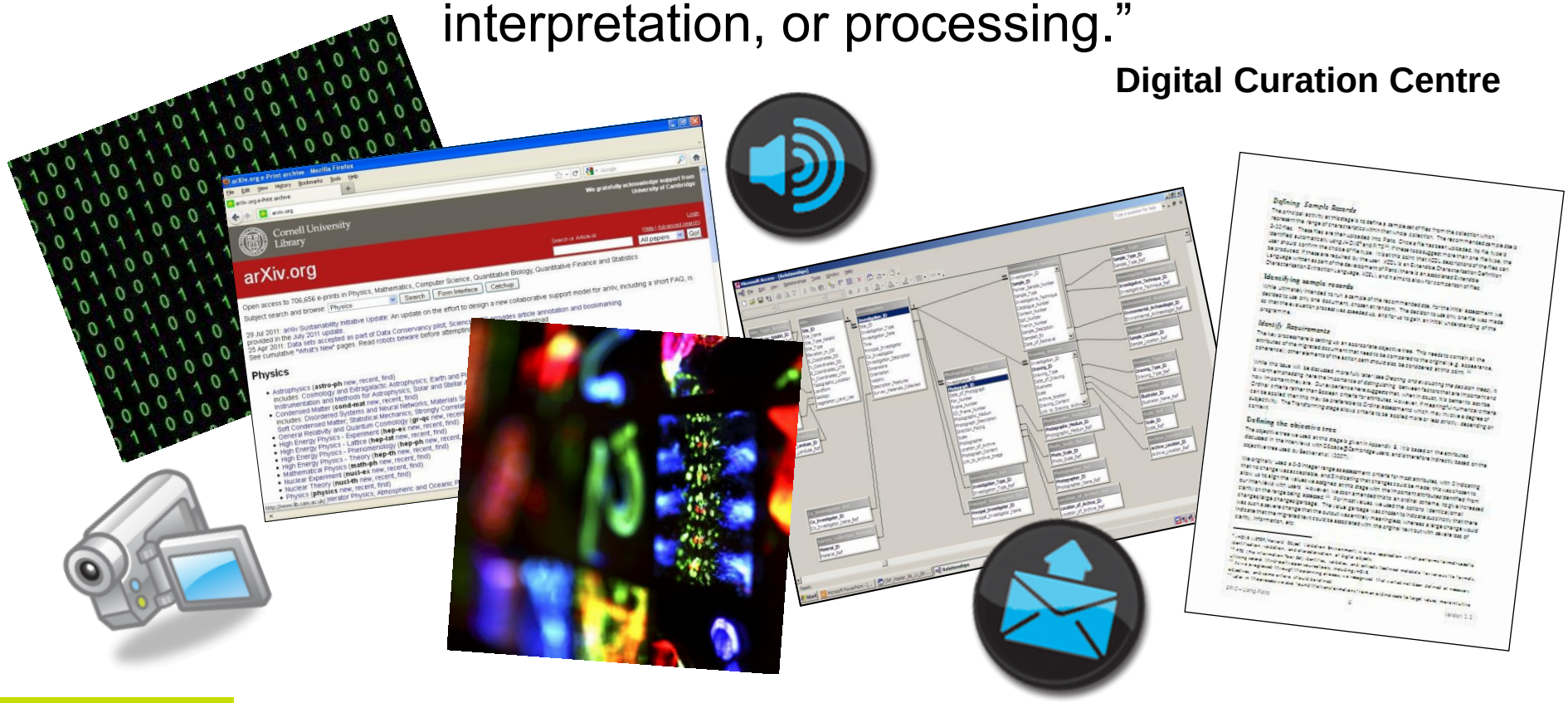
(Knight, 2013)

What is ‘data’?

http://www.lib.cam.ac.uk/dataman/PrePARe/Whatisdata/PrePARe_Whatisdata.pdf

“A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing.”

Digital Curation Centre



What is ‘data’?

http://www.lib.cam.ac.uk/dataman/PrePARe/Whatisdata/PrePARe_Whatisdata.pdf

Any information you use in your research

Vision de l'Australian National Data Service (ANDS)

"*Some of the data might be raw data, the unprocessed observations of particular phenomena. Some might be processed data, the data produced when raw data has been calibrated or corrected. Some might be derived data, which present a summary or specific view of the raw data. Some might be textual data, the publications which result from a research project or the textual data (texts, bibliographies, surveys,etc.) which forms the basis of a research project*".

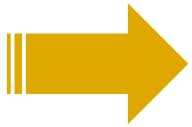
(Research Data Strategy Working Group, 2011)

Guide en ligne
http://ands.org.au/guides/what-is-research-data.html

ANDS enables the transformation of:

Data that are:

- Unmanaged
- Disconnected
- Invisible
- Single use



To Structured Collections that are:

- Managed
- Connected
- Findable
- Reusable

so that Australian researchers can easily publish, discover, access and use/re-use research data.

OCDE
DES POLITIQUES MEILLEURES POUR UNE VIE MEILLEURE

Définition

Enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés **comme sources principales** pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour **valider des résultats de recherche**.

(OCDE, 2007)

Rapport OCDE 2007

Principes et Lignes directrices pour l'accès aux données de la recherche financée sur fonds publics

I. Objectifs

II. Champ d'application et définitions

Données de la recherche

Données de la recherche financée sur fonds publics

Dispositifs d'accès

III. Principes

A. Ouverture

B. Flexibilité

C. Transparence

D. Conformité au droit

E. Protection de la propriété intellectuelle

F. Responsabilité formelle

G. Professionnalisme

H. Interopérabilité

I. Qualité

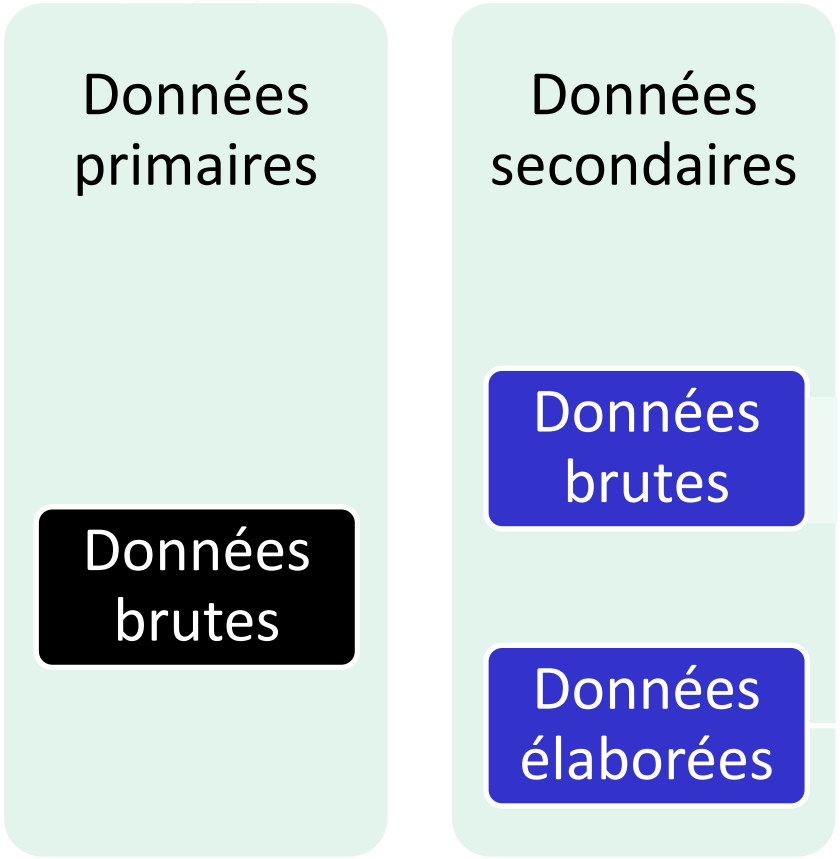
J. Sécurité

K. Efficience

L. Responsabilité de rendre compte

M. Pérennité

Nature et granularité des données

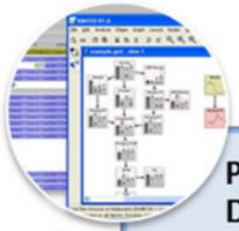


"Secondary data is data collected by someone other than the user... Primary data, by contrast, are collected by the investigator conducting the research". http://en.wikipedia.org/wiki/Secondary_data

Les données dans tous leurs états …

Publication with Data

Données contenues et explicitées dans des articles de recherche

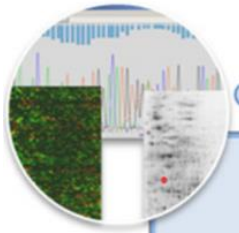


Processed Data & Data Representations

Données dans les fichiers (matériel) supplémentaires liés à des articles

Data Collections and Structured Databases

Données détenues dans des Data Centers et des entrepôts



Raw Data and Primary Data Sets

Données dans « les tiroirs » et sur les disques dans les instituts de recherche

D'après : (Arend et al., 2014)

Nature des données - Exemples

Donnée	Génétique Génomique	Expérimentation Observation	Enquêtes, Analyse textuelle	Données Santé
Brute	• Séquences lues ADN-ARN • Génotypes SNP • Données d'expression (arrays, qPCR)	• Imagerie • Sorties des systèmes de mesure • volt, ampère, ohm, fréquence, kg, °C, etc…	• Données d'enquête produites ou achetées • Corpus textuel (publications par exemple)	• Prescriptions médicales • Données adminis-tratives d'établisse-ments de santé • Enquêtes
Élaborée	• Séquences alignées, assemblées • Données d'expression RNAseq • Annotations des gènes • Données passeport populations/souches	• Combinaison de plusieurs types de données élaborées ou brutes • Carte de température interpolée, flux de chaleur dans un organe, résistance à un stress, etc…	• Données d'enquêtes nettoyées, documentées, anonymisées, … • Métadonnées, référentiels et ontologies en analyse textuelle	• Totaux, évolutions, comparatifs ...

Typologie des données - Exemples

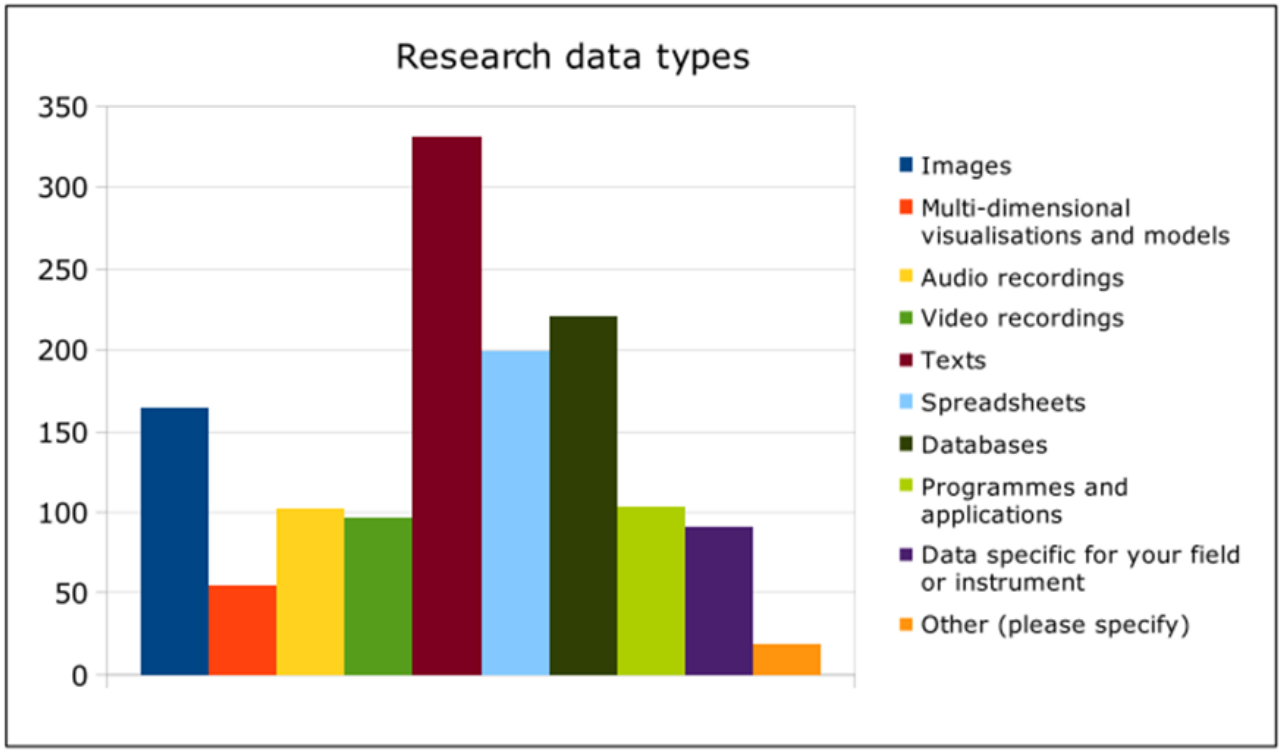


Figure 3: Research data types (Simukovic et al. (2013b))

(Simukovic, Kindling, & Schirmbacher, 2014)

Bibliographie – Qu'est ce que les données ?

- Arend, D., Lange, M., Chen, J., Colmsee, C., Flemming, S., Hecht, D., & Scholz, U. (2014). e!DAL--a framework to store, share and publish research data. *Bmc Bioinformatics, 15*, 214. 10.1186/1471-2105-15-214
- Borgman, C. L. (2014). [Keynote: Data, Data, Everywhere, Nor Any Drop to Drink (slides)]. http://works.bepress.com/borgman/322
- Fayet, S. (2013). « Données » de la recherche, les mal nommées Retrieved from http://urfistinfo.hypotheses.org/2581
- Gaspin, C., & Pontier, D. (2012). Rapport du groupe de travail sur la gestion et le partage des données. Conseil Scientifique de l'Inra (Ed.), (pp. 1-62). http://www.pfl-cepia.inra.fr/uploads/gdp_docs/Rapport-GestionDonnees-web.pdf
- Knight, G. (2013). [Data Management for Librarians : an introduction]. http://fr.slideshare.net/GarethKnight/data-management-for-librarians-an-introduction
- OCDE. (2007). Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics (pp. 1-29). http://www.oecd.org/fr/science/sci-tech/38500823.pdf
- Simukovic, E., Kindling, M., & Schirmbacher, P. (2014). *Unveiling Research Data Stocks: A Case of Humboldt-Universität zu Berlin. iConference 2014, 2014/03/04-07, Berlin - DEU. http://hdl.handle.net/2142/47259*

_03

Gestion des données

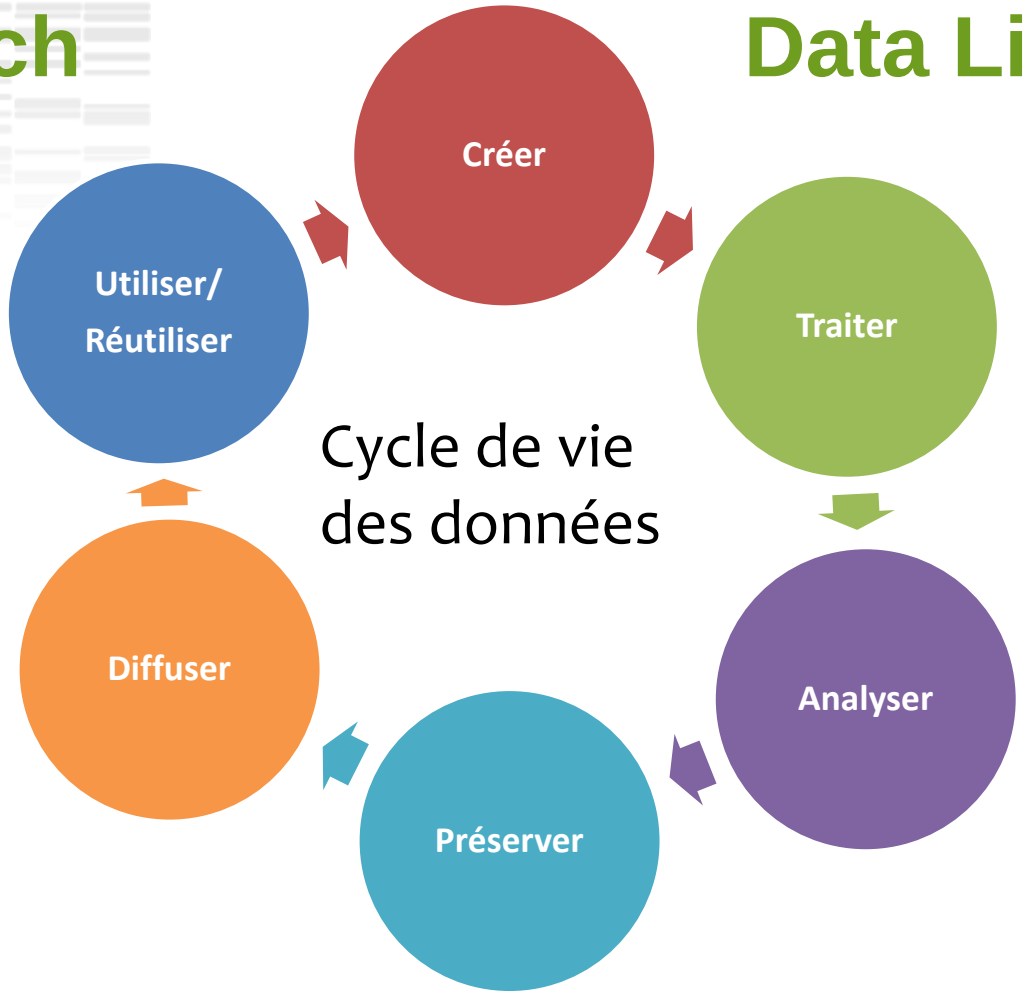
INRA
SCIENCE & IMPACT

3.1- Gestion des données

Cycle de vie des données

INRA SCIENCE & IMPACT

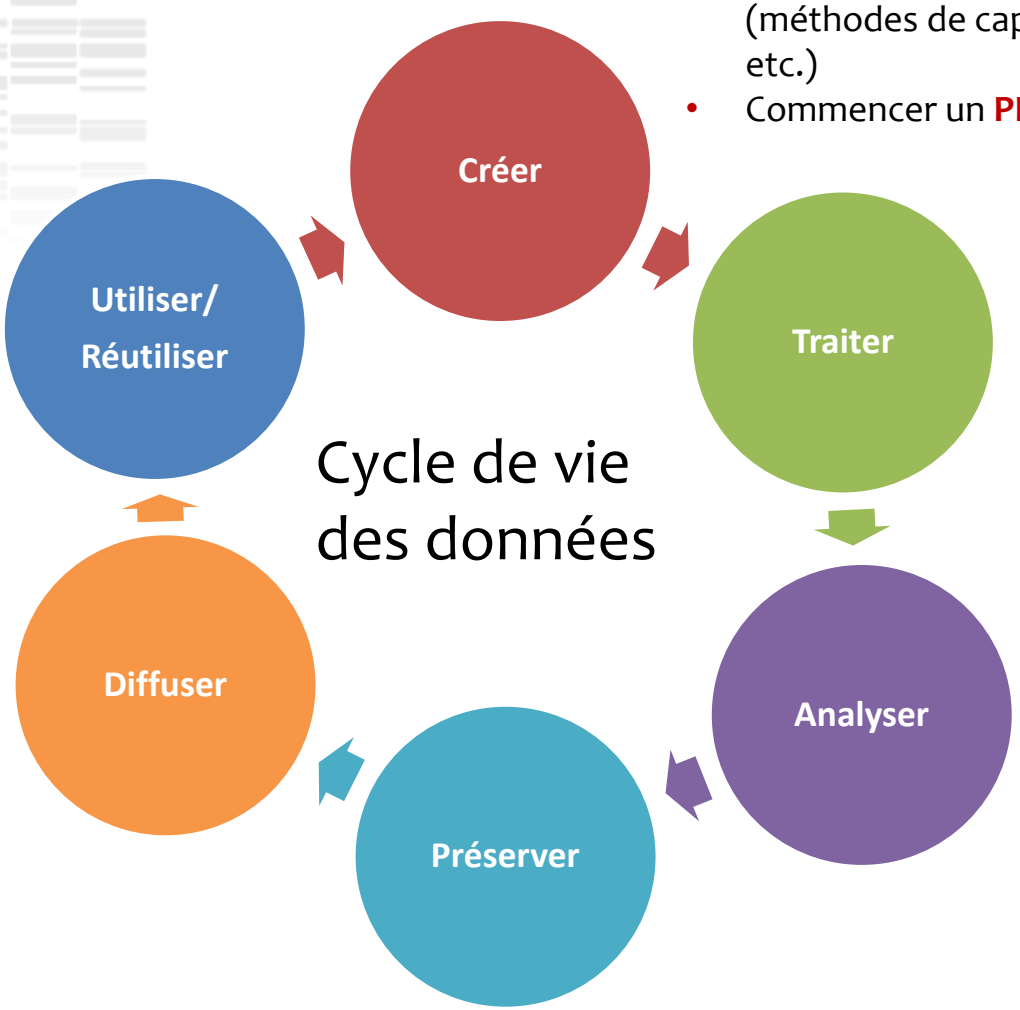
Research Data Lifecycle



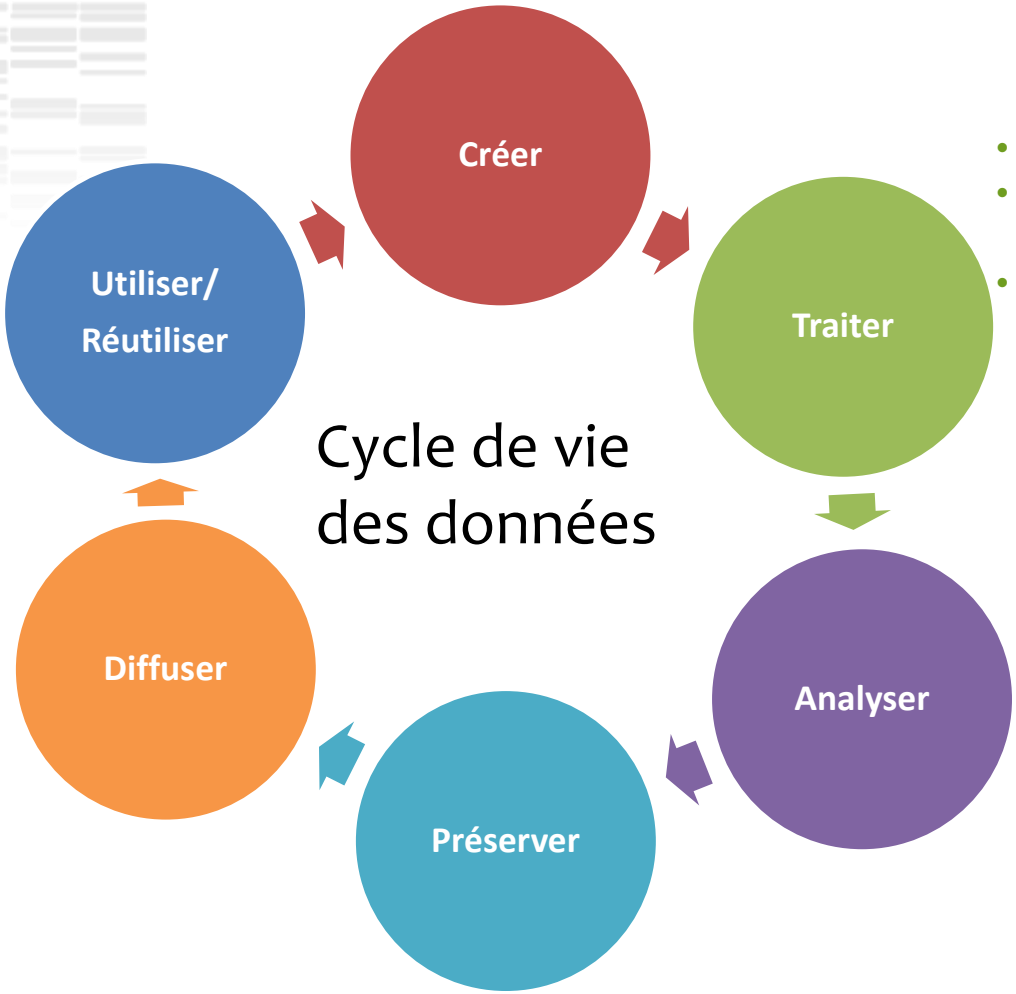
Gestion → Partage → Diffusion - Services

Adapté de http://www.data-archive.ac.uk/create-manage/life-cycle

- Planifier la création et la gestion des données (méthodes de capture, formats, stockage, etc.)
- Commencer un **PLAN de GESTION**



Adapté de ***http://www.data-archive.ac.uk/create-manage/life-cycle***

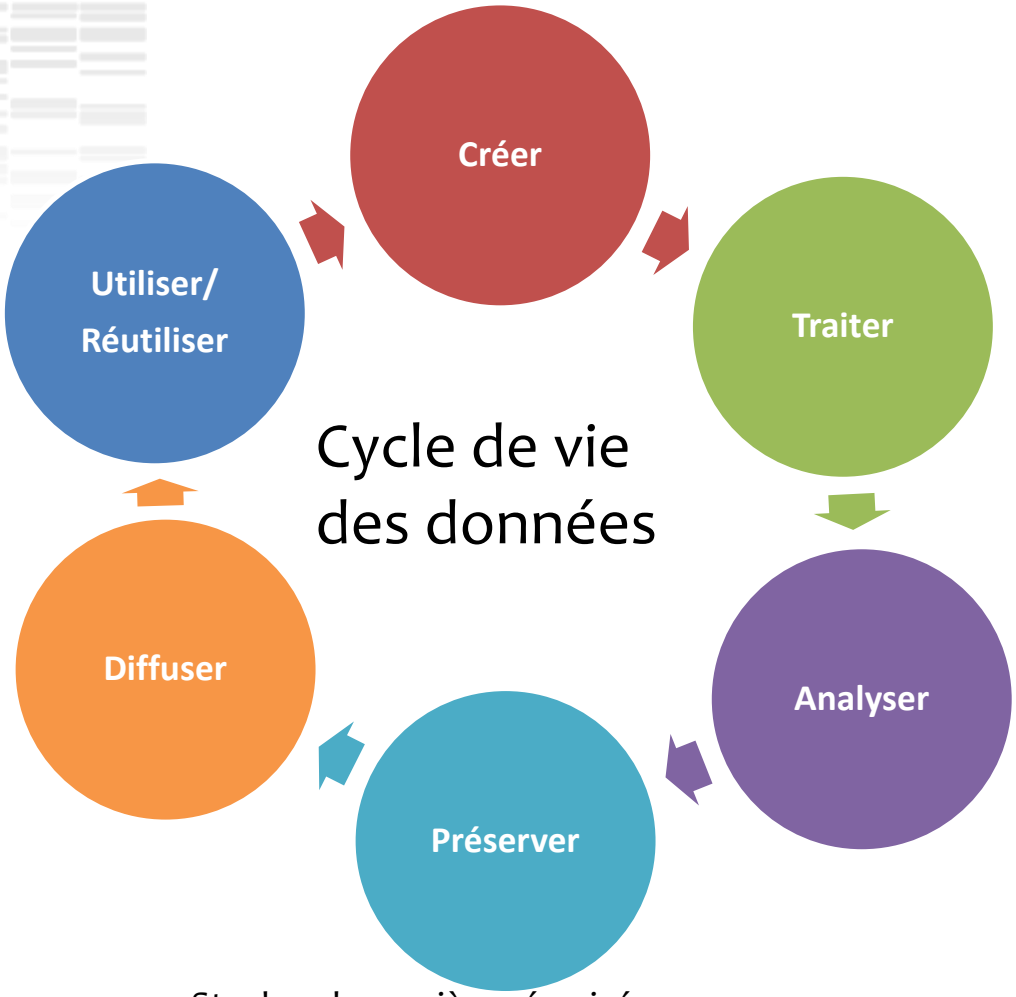


- **Documenter**
- Vérifier, valider, anonymiser, etc.
- Mettre à jour le plan de gestion des données

Adapté de http://www.data-archive.ac.uk/create-manage/life-cycle

- Sélectionner les données qui seront conservées
- Choisir une infrastructure

Adapté de http://www.data-archive.ac.uk/create-manage/life-cycle



- Stocker de manière sécurisée
- Archiver de manière pérenne

Adapté de ***http://www.data-archive.ac.uk/create-manage/life-cycle***

- Définir les droits d'accès et les conditions d'utilisation
- Promouvoir (faire savoir)

Adapté de http://www.data-archive.ac.uk/create-manage/life-cycle

- **Trouver**
- Vérifier les conditions de réutilisation
- Évaluer la qualité
- **Citer**

Adapté de http://www.data-archive.ac.uk/create-manage/life-cycle

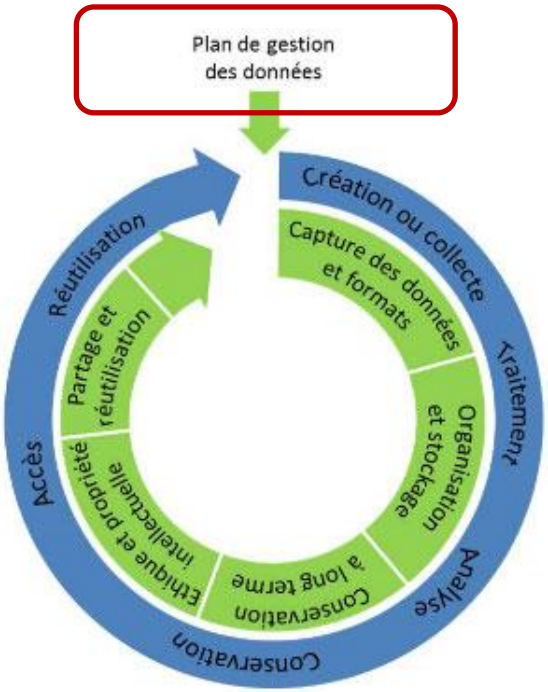
3.2- Gestion des données

Plan de gestion des données

INRA
SCIENCE & IMPACT

Plan de gestion des données
(Data Management Plan = DMP)

- Document, établi au démarrage d'un projet de recherche, qui décrit la façon dont les données seront obtenues, traitées, organisées, stockées, sécurisées, préservées, partagées,… au cours et à l'issue du projet
- Il couvre tout le cycle de vie des données, s'applique à tous les jeux de données du projet
- Il aide à la mise en place de bonnes pratiques de gestion



Inist, d'après UK Data Service

A quelles questions répond un PGD ?

Responsabilités dans le projet
En quoi consiste le projet ? Qui sont les partenaires ? Quelle politique en matière de gestion des données ? Qui est responsable de la gestion des données ?

Collecte des données
Quelles données seront produites/utilisées au cours du projet ? (type, format, volume et accroissement…). Comment seront-elles produites ? Transformées ?

Ressources
Comment la gestion des données est-elle financée, en particulier à long terme ?

Documentation des données
Comment, où, par qui, seront stockées, sauvegardées et sécurisées les données ?

Dans le cadre d'un projet de recherche ou non…

Accès et partage des données
Qui pourra accéder aux données ? Les données seront-elles publiées ? partagées ? Avec qui ? Comment ? Dans quel délai ? Sous quelle licence ?

Sauvegarde des données
Comment les données seront elles identifiées, décrites ? Quels standards de métadonnées utilisera t'on ? Comment seront générées les métadonnées ?

Propriété intellectuelle
Qui sera propriétaire des données produites ? Des données externes seront-elles utilisées ?

Ethique
Des données sensibles seront-elles produites ou utilisées ? Comment sera assurée l'anonymisation des données ?

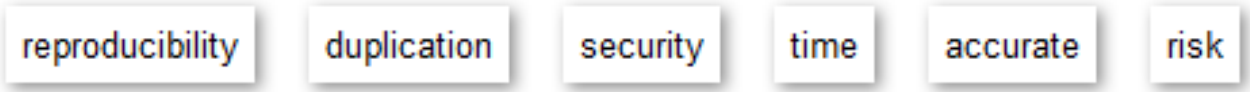
Archivage et préservation des données
Quel plan pour l'archivage et la préservation à long terme ?

Pourquoi rédiger un PGD ? (1)

"Plan ahead to create high-quality shareable research data"

UK Data Service

- **Pour se poser les bonnes questions** dès le début du projet et…
  - **identifier les risques** liés à la gestion des données, assurer la sécurité et la préservation des données sur le long terme,
  - **identifier les responsabilités**, les rôles de chacun dans la gestion des données, **planifier** les ressources et compétences nécessaires à cette gestion,
  - donner accès à des données fiables afin d'assurer la reproductibilité de la recherche et permettre à d'autres de comprendre et d'utiliser les données (**planning for sharing**).



Pourquoi rédiger un PGD ? (2)

- Pour **répondre aux exigences des financeurs**, exemples :

RESEARCH COUNCILS UK	• 6 des 7 principaux Research Councils
National Science Foundation WHERE DISCOVERIES BEGIN NIH National Institutes of Health Turning Discovery Into Health	• National Science Foundation (NSF) • National Institutes of Health (NIH)
	• H2020
Australian Government Australian Research Council	• Australian Research Co • National Health and M Council (NHMRC)
Canada	• Recommandations plus que des exigences

The National Science Foundation (NSF) explains that Data Management Plans are to be "reviewed as an integral part of the proposal, coming under Intellectual Merit or Broader Impacts or both, as appropriate for the scientific community of relevance." (Creamer, 2014)

Exemples de plans rédigés (extrait – Wiki Données INRA)

- DCC (UK) : http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples
- Exemples de plans dans différentes disciplines : http://library.umassmed.edu/necdmc/dmp
- NSF :
  - NSF General: Mauna Loa example
  - NSF General: Rio Grande example
  - NSF General: HDF Map example
  - NSF General: Nutrient Network example
  - NSF BIO: E. afffinis example

- Oxford : projet DAMARO
  - David Shotton's Twenty Questions for Research Data Management - PDF - June 2013 (questions type avec réponses possibles) et Data Management Plan Template - PDF - June 2013.

- 3 études de cas (PowerPoint presentation - June 2013) :
  - Research Data Case Study - Humanities
  - Research Data Case Study - Social Sciences
  - Research Data Case Study - Physical Sciences

- University of Minnesota - Exemples et infos autour des plans de gestion de données https://www.lib.umn.edu/datamanagement/DMP/example

- UC San Diego - Exemples de plans dans différentes disciplines http://idi.ucsd.edu/data-curation/examples.html

- Wageningen :
  - Tableau général
  - exemple de plans pour 2 PhD
    - A Spatially Explicit Modelling Approach to Estimate Waterborne Pathogen Concentrations in the Surface Waters of the World
    - Data Management Plan for the PhD project: Development and Application of a Monitoring System to Assess the Impacts of Climate and Land Cover Changes on Eco-Hydrological Processes in an Eastern Andes Catchment Area

Exemple : Horizon 2020

Lignes directrices pour la gestion des jeux de données produits par un projet

- Identifiant et nom du jeu de données
- Description du jeu de données :
  - origine, nature, échelle, utilité, publication scientifique en relation, existence de données similaires, réutilisations possibles…
- Normes et métadonnées utilisées
- Partage des données
  - modalités, procédures d'accès, embargo, dispositifs techniques et logiciels nécessaires à la réutilisation des données, licence. Entrepôt dans lequel les données seront stockées. Raisons d'une éventuelle impossibilité de partage ;
- Stockage et sauvegarde, archivage et conservation (volumes, coûts)
- Prescriptions supplémentaires pour que les données de recherche soient :
  1. découvrables,
  2. accessibles,
  3. évaluables et compréhensibles,
  4. utilisables au-delà du but premier de leur collecte,
  5. interopérables selon des normes qualitatives spécifiques

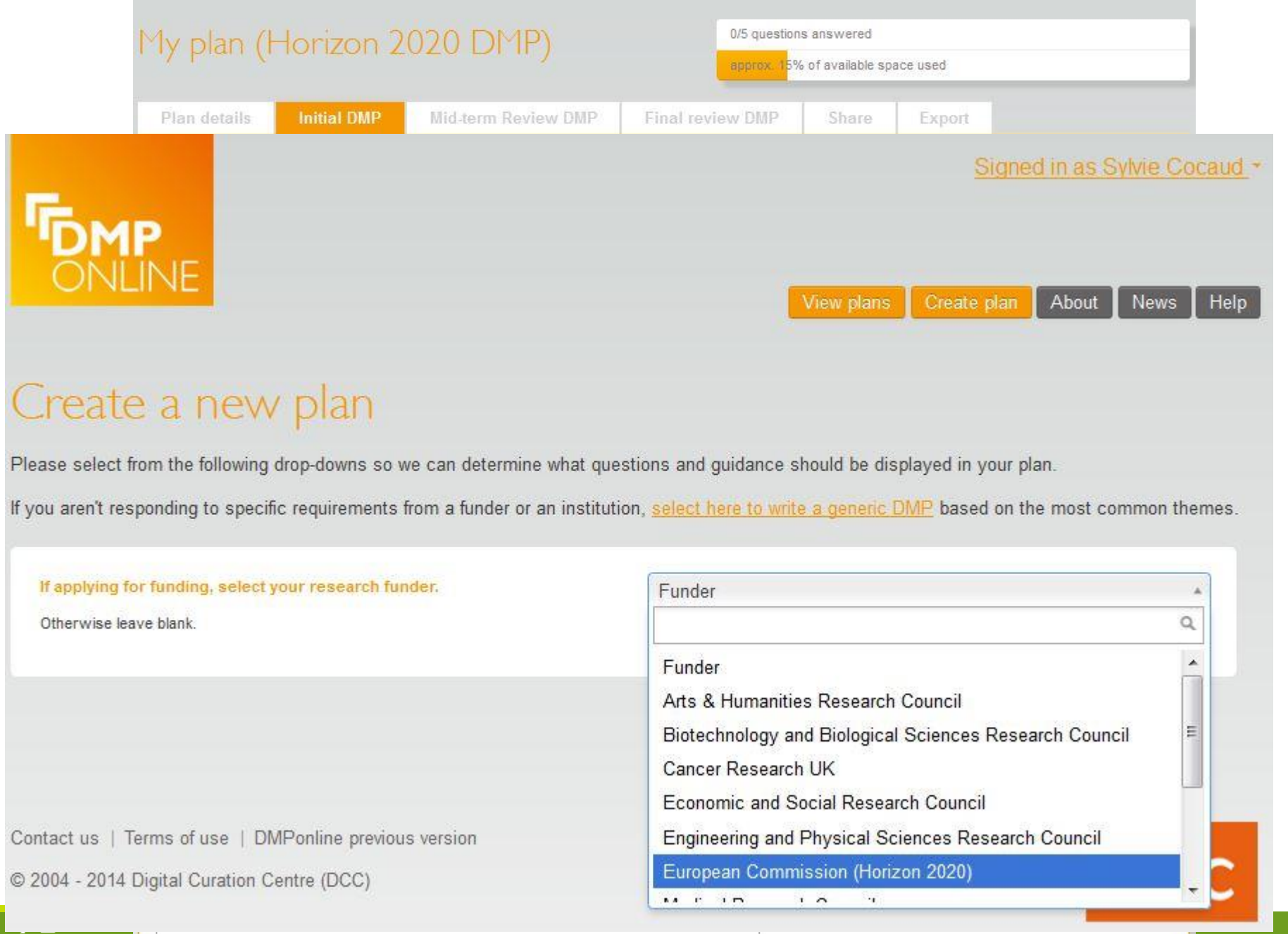
http://openaccess.inist.fr/IMG/pdf/lignes_directrices_pgd_horizon_2020_tr_fr.pdf

Comment rédiger un PGD ?

Quelques outils

- Outils en ligne
  - DMPOnline (DCC, UK) - Intègre les directives H2020
  - DMPTOOL (University of California Curation Center, US)

- Checklists, templates, guides, exemples
  - Checklist for a Data Management Plan.
    DCC (Grille par défaut utilisée par DMPOnline).
  - DMP Requirements Guides et exemples (US)
  - Data management checklist
    UK Data Archive
  - Research data management planning checklist
    Univ. of Sidney (AU)
  - Data Management Training Modules
    University of Guelph (CA)
  - Lignes directrices pour la gestion des données dans Horizon 2020.
    Annexe 1 : modèle de plan de gestion de données (PGD) (UE)
    Programme-cadre européen pour la recherche et l'innovation. Horizon 2020.
  - Template for an individual data management plan
    Wageningen University & Research centre (NL) pour les doctorants par exemple


My plan (Horizon 2020 DMP)
0/5 questions answered
approx. 15% of available space used
Plan details
Initial DMP
Mid-term Review DMP
Final review DMP
Share
Export
Signed in as Sylvie Cocaud
DMP ONLINE
View plans
Create plan
About
News
Help
Create a new plan
Please select from the following drop-downs so we can determine what questions and guidance should be displayed in your plan.
If you aren't responding to specific requirements from a funder or an institution, select here to write a generic DMP based on the most common themes.
If applying for funding, select your research funder.
Otherwise leave blank.
Funder
Funder
Arts & Humanities Research Council
Biotechnology and Biological Sciences Research Council
Cancer Research UK
Economic and Social Research Council
Engineering and Physical Sciences Research Council
European Commission (Horizon 2020)
Contact us | Terms of use | DMPonline previous version
© 2004 - 2014 Digital Curation Centre (DCC)

3.3- Gestion des données

Documentation

INRA SCIENCE & IMPACT

Enjeux de la documentation des données

- Favoriser la compréhension des données
- Favoriser la reproductibilité des données
- Faciliter la recherche et la localisation des données
- Faciliter la réutilisation

Documentation et métadonnées

"I guess it makes sense for a robot to read an e-book [401]" by brianjmatis on flickr

En pratique…

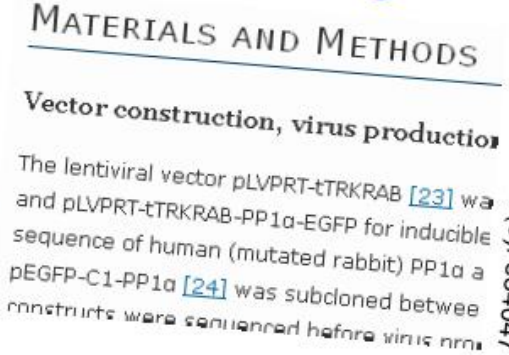
- **Who** created it, **when** and **why**

www.texample.net

- Include:
  - **Description** of the item
  - **Method or theoretical approach**
  - **What** is being **measured**
  - **References** to related data

description ***n.***
A set of characteristics by which something can be recognised

By mdxdt on flickr: www.flickr.com/photos/dxdt22/177749386/



MATERIALS AND METHODS

Vector construction, virus productio

The lentiviral vector pLVPRT-tTRKRAB [23] wa
and pLVPRT-tTRKRAB-PP1α-EGFP for inducible
sequence of human (mutated rabbit) PP1α a
pEGFP-C1-PP1α [24] was subcloned betwee
constructs were sequenced before virus pro

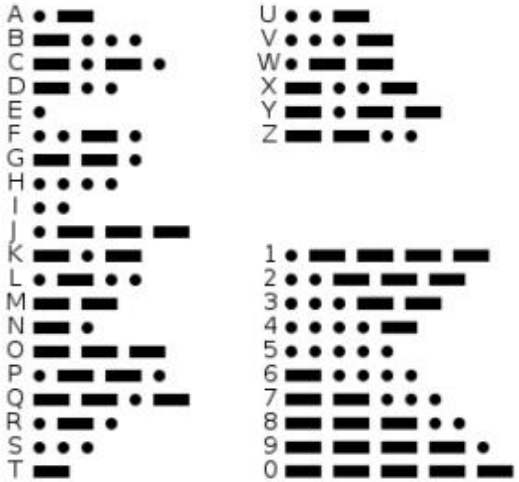
M. Farinelli et al. (2012) PLoS ONE 7(3): e34047

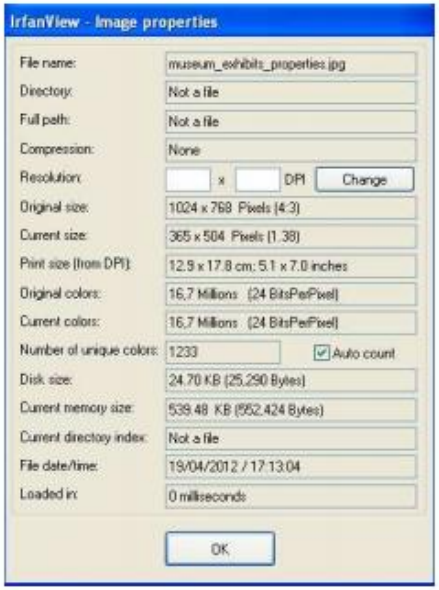
http://www.lib.cam.ac.uk/dataman/PrePARe/Explainit/PrePARe_ExplainIt.pdf

En pratique…

- **Define** jargon, acronyms and code

By Gavin Llewellyn
http://www.flickr.com/photos/gavinjllewellyn/6826303487/





- Provide **technical information** about the file
  (may be generated automatically)

http://www.lib.cam.ac.uk/dataman/PrePARe/Explainit/PrePARe_ExplainIt.pdf

3.4- Gestion des données

Stockage sécurisé et Préservation

INRA SCIENCE & IMPACT

Stockage sécurisé (backup)

- Sécurité physique
  - ✓ Copies multiples : principe LOCKSS ou Lots Of Copies Keeps Stuff Safe
  - ✓ Lieux de stockage multiples
- Sécurité informatique
  - ✓ Protection contre les virus
  - ✓ Accès contrôlé
  - ✓ Cryptage
  - ✓ Etc.

Préservation des données

- Enjeux : être capable d'accéder à, de lire, comprendre, utiliser la donnée sur le long terme
- Plusieurs risques doivent être prévenus
  - ✓ Détérioration des supports physiques → LOCKSS
  - ✓ Absence d'informations décrivant le contenu → Documentation
  - ✓ Obsolescence des formats de fichiers → Formats durables
  - ✓ Disparition/évolution logiciels ou matériels de lecture → Migration
  - ✓ Etc.

Exemples de formats recommandés

Ce tableau dérive de plusieurs recommandations : université d'Edinburgh (http://www.data-archive.ac.uk/create-manage/format/formats-table), TGE ADONIS (http://www.huma-num.fr/sites/default/files/ressourcesdoc/la-lettre-fev-mars2013.pdf), PURR (https://purr.purdue.edu/legal/preservation-strategies)

Types de données	Formats recommandés pour la préservation
Données quantitatives tabulaires	• SPSS format portable (.por) • Fichiers texte .CSV, .tab ou délimités avec un caractère donné (ce caractère doit être absent des données elles mêmes) • Fichiers de commandes ('setup') contenant des métadonnées : SPSS, Stata, SAS • Fichiers texte structurés ou balisés contenant des métadonnées : DDI XML file
Données géographiques Données vectorielle ou matricielles (raster)	• ESRI Shapefile (essential - .shp, .shx, .dbf, optional - .prj, .sbx, .sbn) • geo-referenced TIFF (.tif, .tfw) • Données CAD (.dwg) • Données tabulaires d'attributs GIS
Données qualitatives Données textuelles	• eXtensible Mark-up Language (XML) conforme à un Document Type Definition (DTD) ou un schema (.xml) • Rich Text Format (.rtf) • Données plein texte, ASCII (.txt)
Données images numériques	• TIFF version 6 non compressé (.tif)
Données audio numériques	• Free Lossless Audio Codec (FLAC) (.flac)
Données vidéo numériques	• MPEG-4 (.mp4) • motion JPEG 2000 (.mj2)
Documents et scripts	• Rich Text Format (.rtf) • PDF/A or PDF (.pdf) • HTML (.htm) • OpenDocument Text (.odt)

Préservation des données : faire des choix

- Tout préserver est
  - ✓ Coûteux
  - ✓ Inefficace
- Critères de sélection des données
  - ✓ Critères liés au cadre juridique, à la stratégie de son établissement, etc.
  - ✓ Coûts vs bénéfices
  - ✓ Considérations personnelles

Préservation des données : à qui s'adresser

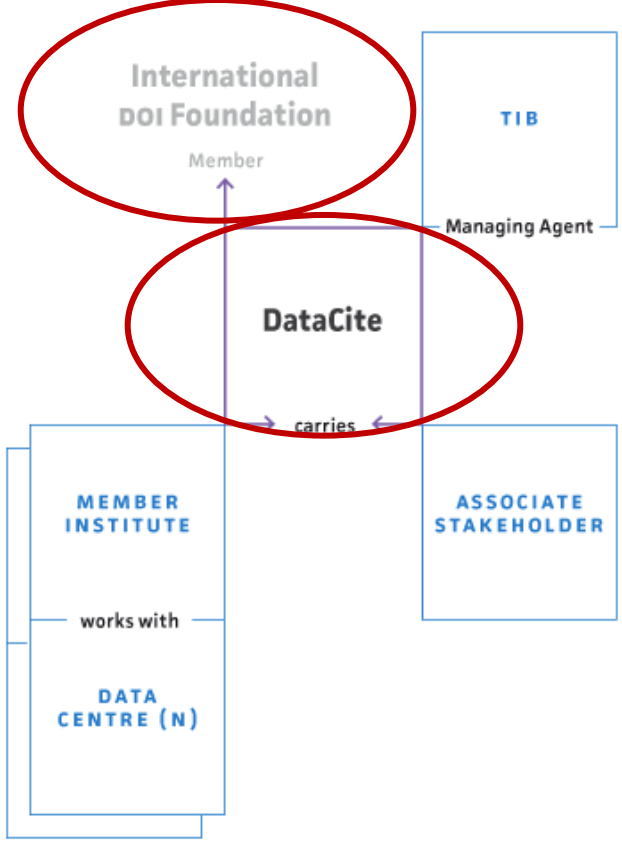
- Entrepôt institutionnel
- Entrepôts thématiques (Pangea, Knowledge Network for Biocomplexity, etc.)
- Entrepôts pluridisciplinaires
  - ✓ Portés par des organisations privées (Dryad, Figshare, etc.)
  - ✓ Portés par des organisations publiques (Eudat B2SHARE, Zenodo, 3TU.Datacentrum, etc.)

3.5- Gestion des données

DOI

Digital Object Identifier (DOI)

Permet de référencer, citer, fournir un lien unique et pérenne vers un objet numérique



DataCite

Helping you to find, access, and reuse data

Inist-CNRS
Member Institute

DOI: 10.5072/xfwugk

RESOLVER

Page descriptive publique

10.5072
Préfixe

xfwugk
Suffixe

INRA
Data center

DOI : les services proposés par DataCite

- DataCite Metadata Store (MDS)
  - Création de DOI et enregistrement de métadonnées associées : https://mds.datacite.org/
- DataCite Metadata Search
  - Recherche des métadonnées associées à l'objet scientifique enregistré dans DataCite : http://search.datacite.org/ui
- DataCite Statistics
  - Statistiques d'enregistrement et de résolution de DOI : http://stats.datacite.org/
- DOI Citation Formatter
  - En collaboration avec CrossRef, création de différents formats de citation pour les DOI DataCite et CrossRef : http://crosscite.org/citeproc/
- Content Negotiation
  - Possibilité d'obtenir des métadonnées dans des formats divers et/ou d'accéder automatiquement et directement à un objet plutôt que par la « landing page » : http://data.datacite.org/static/index.html
- DataCite Test Environment
  - Environnement de test DataCite : http://test.datacite.org/
- DataCite OAI Provider
  - Exposition des métadonnées en OAI-PMH : http://oai.datacite.org

TP - DOI

Interface DataCite

INRA SCIENCE & IMPACT

3.6- Gestion des données

Enjeux

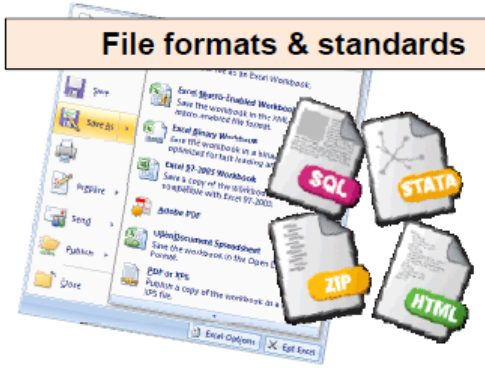
Les enjeux d'une bonne gestion des données

- Les données numériques sont « fragiles »
  - ✓ Détérioration des supports
  - ✓ Sinistres
  - ✓ Etc.
- Les données numériques peuvent devenir illisibles ou inutilisables
  - ✓ Obsolescence du format
  - ✓ Perte des logiciels ou matériels de lecture
  - ✓ Manque de documentation ou métadonnées
- Enjeu patrimonial des données

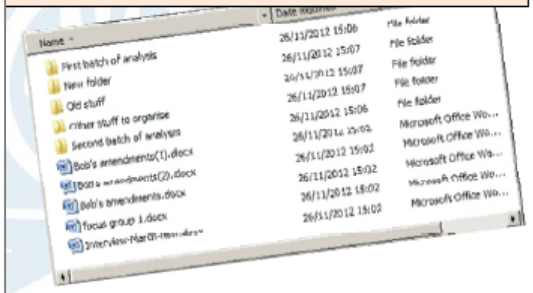
Short-term decisions with long-term implications

Software products

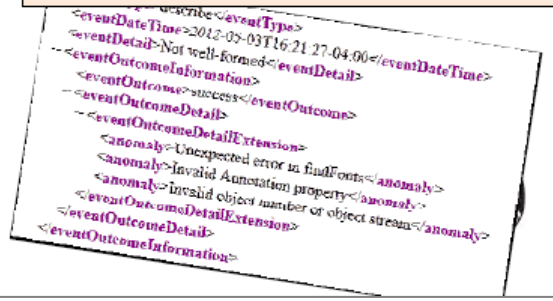
File formats & standards



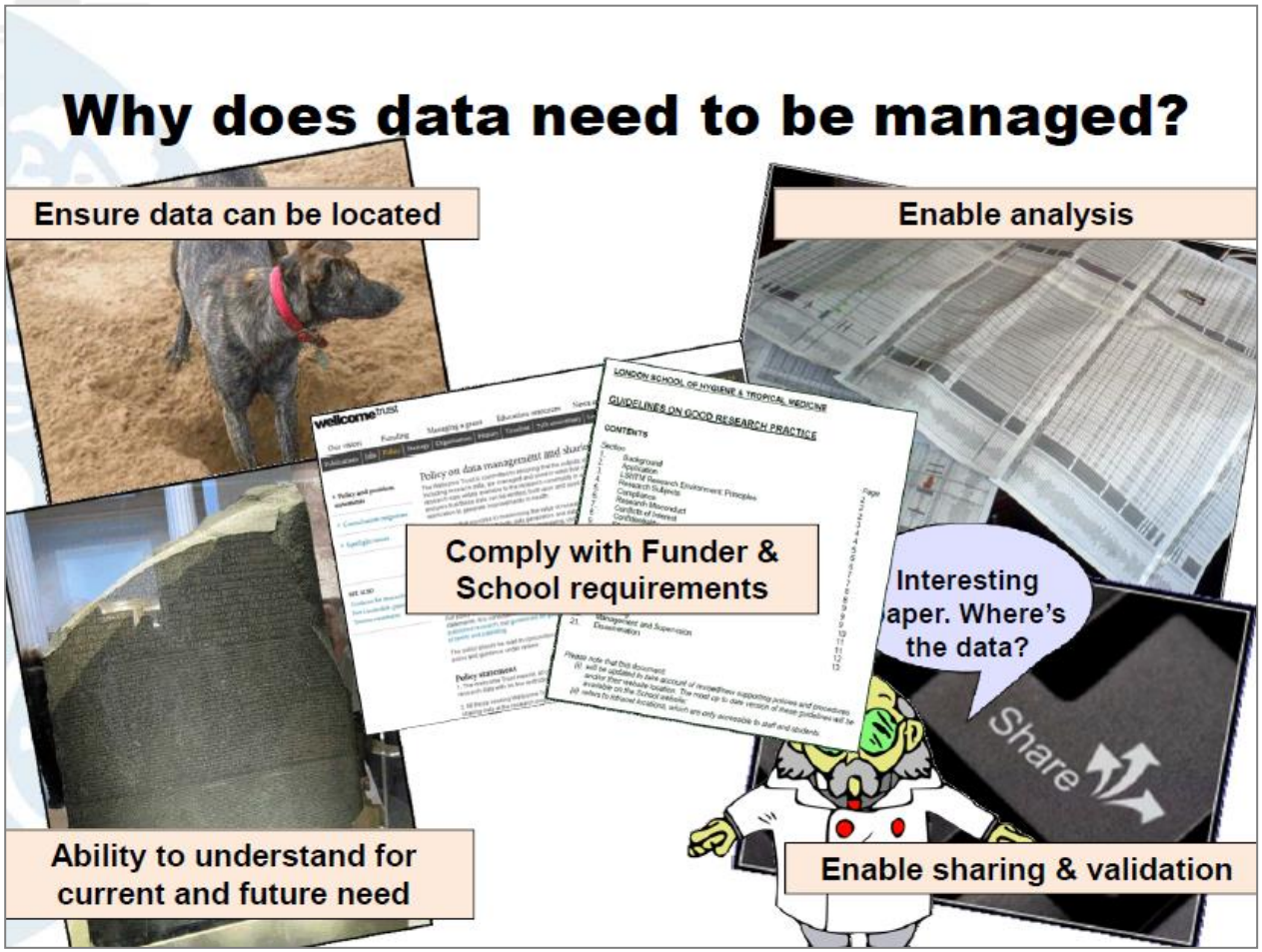
Data organisation & labelling



Quality Controls

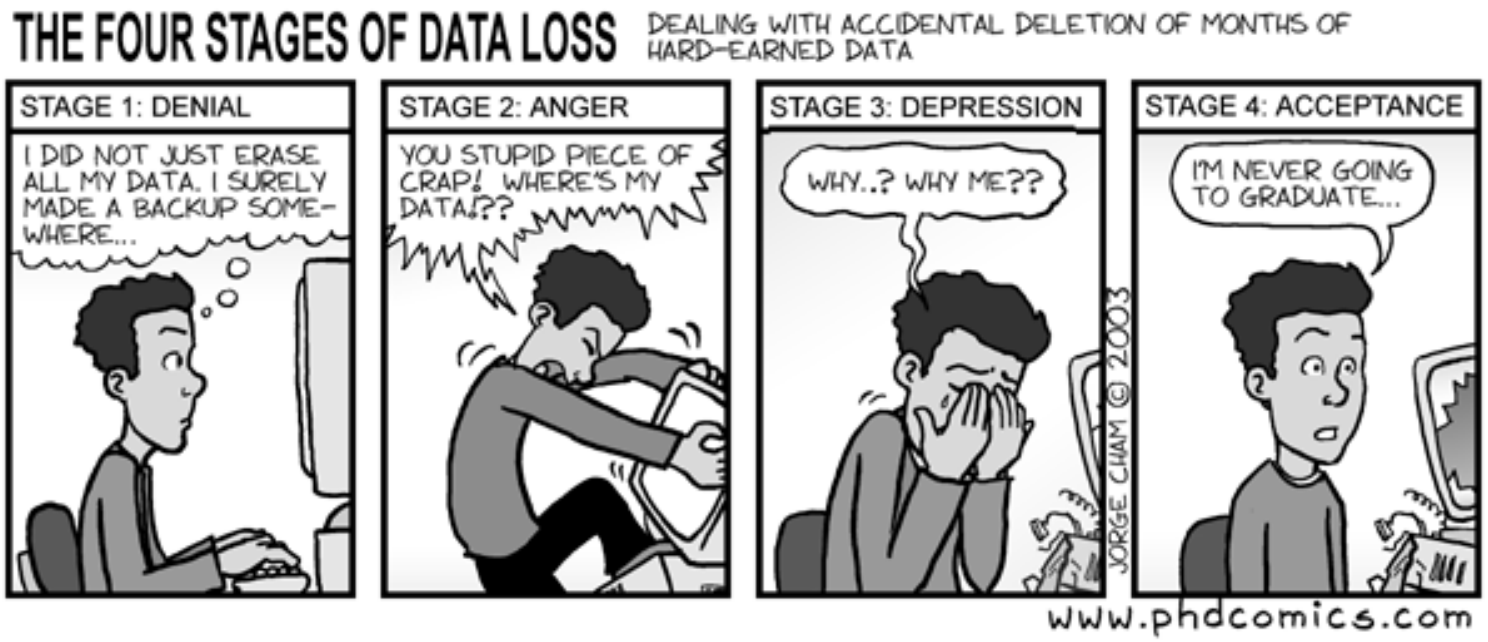


http://fr.slideshare.net/GarethKnight/data-management-for-librarians-an-introduction

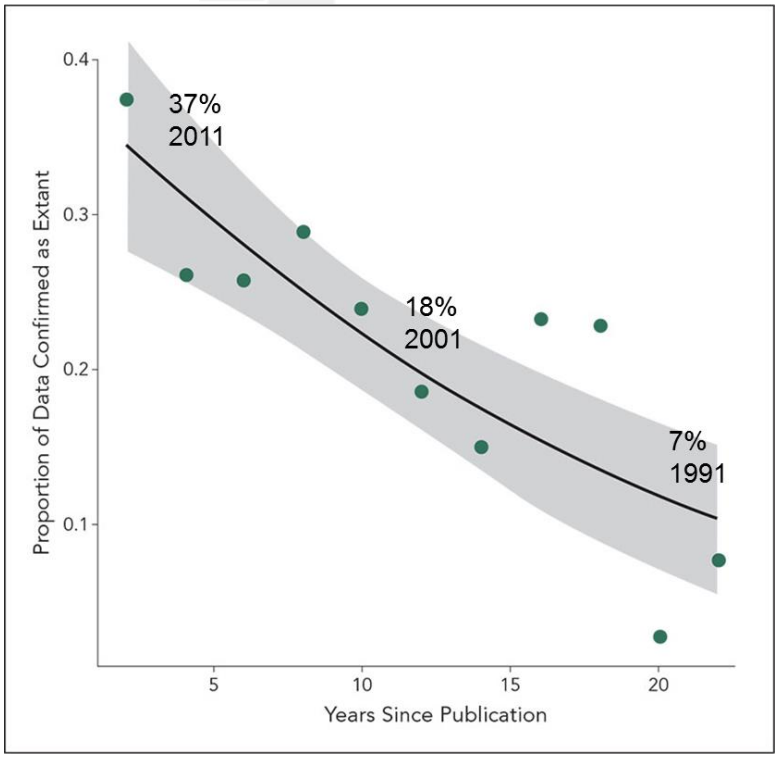


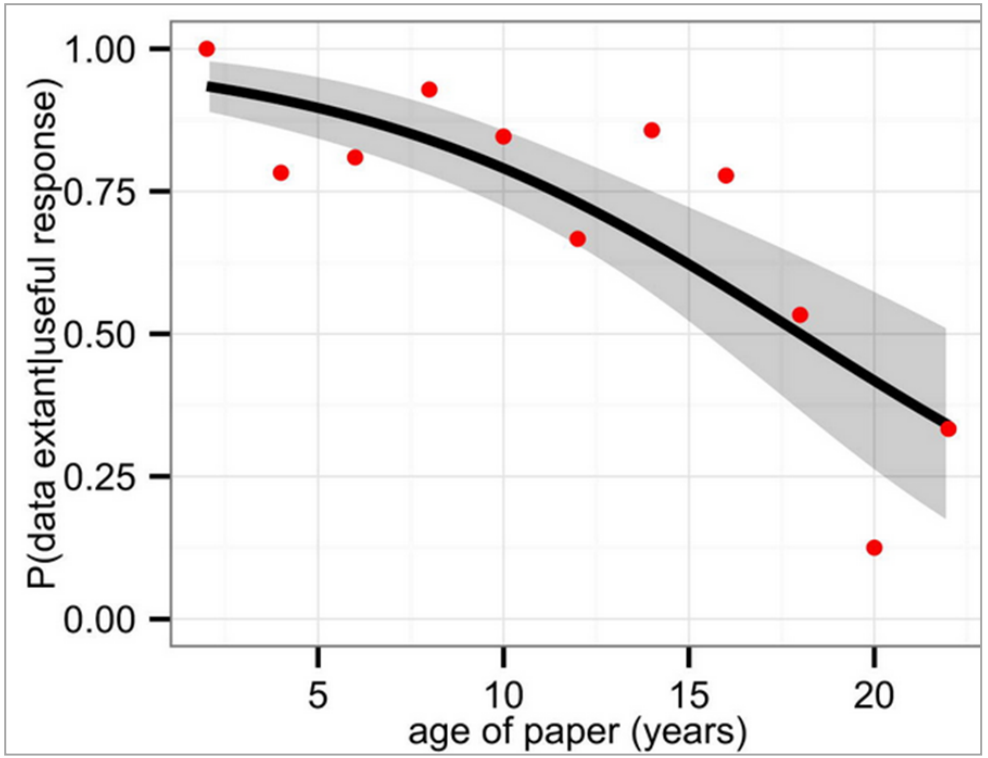
http://fr.slideshare.net/GarethKnight/data-management-for-librarians-an-introduction

Risques de perte des données !



20 ans après publication : 80% des données perdues





Etude portant sur 516 articles publiés entre 1991 et 2011. Domaine : “Morphological data from plants, animals, or other organisms”. La disponibilité des données chute fortement au cours du temps !

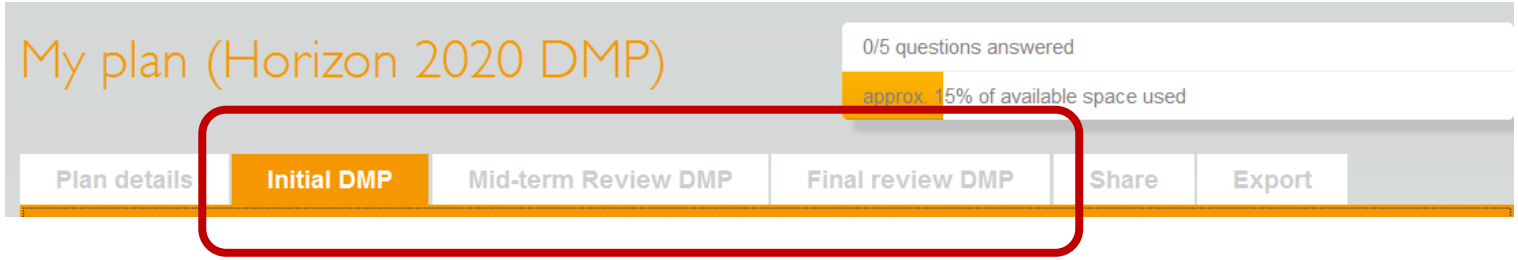
(Vines et al., 2014)

3.7- Gestion des données

Difficultés et freins

Difficultés liées au cycle de vie…

- « La vie d'un projet de recherche n'est pas un long fleuve tranquille »
  - Le contexte et le workflow peuvent changer en cours de projet : le PGD doit être entretenu et mis à jour selon ces évolutions. Exemple : Versioning prévu dans le modèle DMPOnline



  - Quelles données, quelles métadonnées de contexte seront utiles plus tard à un tiers (données partagées) voire même au producteur de données ?

Freins…
au niveau des chercheurs

- Les chercheurs ne sont pas formés à la gestion des données,
  - Difficultés liées à la méconnaissance des formats de données et de métadonnées
- Perception d'une perte de temps, manque de volonté
  - s'ils ne perçoivent pas un impact direct sur leurs recherches et leurs productions
- Beaucoup restent sceptiques quand à l'utilité de leurs données sur le long terme

Freins… organisationnels et techniques

- Des pratiques de gestion des données trop diverses (formats, métadonnées)
- Pas de stratégie de préservation
- Manque d'infrastructures adéquates
  - Espaces de stockage sécurisés adéquats
  - Espaces collaboratifs avec une gestion fine des droits d'accès aux données.
- Manque d'incitation
  - Absence de politiques institutionnelles
  - Manque de reconnaissance dans le processus d'évaluation du chercheur

TP-TD
Découverte de sites

- Liste de supports de formation
- Outils en ligne
- Sites web de référence (ANDS, DCC ..)

INRA SCIENCE & IMPACT

Supports de formation

- Europe
  - ✓ https://joinup.ec.europa.eu/community/ods/description
- France
  - ✓ http://www.inist.fr/donnees/co/Donnees_recherche_web.html
- UK
  - ✓ http://www.data-archive.ac.uk/create-manage/training-resources
  - ✓ http://datalib.edina.ac.uk/mantra
  - ✓ http://www.lib.cam.ac.uk/dataman/training.html
  - ✓ http://eprints.soton.ac.uk/338816
  - ✓ http://library.leeds.ac.uk/roadmap-project-outputs en particulier cette page : http://library.leeds.ac.uk/info/377/roadmap/123/roadmap_events/3
  - ✓ http://www.uel.ac.uk/trad/outputs/resources/
  - ✓ http://rdmrose.group.shef.ac.uk/?page_id=10
  - ✓ http://datalib.edina.ac.uk/mantra/libtraining.html
  - ✓ http://www.bath.ac.uk/research/data
- Canada
  - ✓ Clairoux, N. (2014). Introduction à la gestion des données de la recherche. http://fr.slideshare.net/nclairoux/gestion-donnes-recherche-plan-de-gestion-des-donnes-archivage-prservation

Plan de gestion – Sites utiles

- **Pays Bas**
  - ✓ L'**université de Wageningen** propose différents services autour des données de la recherche et en particulier une aide à la constitution de plans de gestion de données : Data management plans.
- **Royaume Uni**
  - ✓ Site du DCC (UK) : FAQ, checklist, exemples
  - ✓ http://ukdataservice.ac.uk/manage-data/handbook/
  - ✓ Welcome Trust
  - ✓ University of Bristol
  - ✓ University of Edimburgh (guides, services, cours en ligne : MANTRA)
- **USA**
  - ✓ MIT: http://libraries.mit.edu/guides/subjects/data-management/plans.html
  - ✓ US Geological Survey : http://www.usgs.gov/datamanagement/index.php
  - ✓ University of Wisconsin-Madison:
    - http://researchdata.wisc.edu/make-a-plan/data-plans/ (in general)
    - http://researchdata.wisc.edu/make-a-plan/nsf-data-plans4/ (NSF-specific)
  - ✓ ICPSR (Inter-university Consortium for Political and Social Research) : http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/
  - ✓ California Digital Library: http://www.cdlib.org/services/uc3/datamanagement/
  - ✓ Thea P. Atwood. "NSF Data Management Plans" University of Massachusetts. Dec. 2014. http://works.bepress.com/tpatwood/5
- **Australie**
  - ✓ http://ands.org.au/guides/data-management-planning-awareness.html

Plan de gestion - Outils & formulaires

Outil	Proposé par	Pays	Commentaires
DMPOnline	DCC (Digital Curation Center)	UK	
DMP20	Université d'Oxford	UK	Formulaire simplifié pour les doctorants et les jeunes chercheurs. Utilise également DMP Online. Voir aussi le projet Oxford DMPOnline
IEDA DMP Tool	IEDA (Integrated Earth Data Applications) (US-NSF)	US	Pour les projets soumis à la NSF. Bien que spécialisé dans les données d'observation en Geoscience (marines et terrestres), le DMP est conçu comme un plan générique applicable à d'autres domaines. Voir aussi : Data Management Plan (TEMPLATE Web site) Data Management & Sharing Frequently Asked Questions (FAQs)
DMPTOOL	University of California Curation Center of the California Digital Library	US	Le template de saisie dépend de l'organisme de financement choisi par le chercheur. DMP templates and requirements by funders, with examples Vidéo courte (1.46 mn) pour découvrir cet outil

Bibliographie – Gestion des données

- Atwood, T. P. (2014). [NSF Data Management Plans. Tips & Guidance on applying to your SBE DDRIG]. http://works.bepress.com/tpatwood/5
- Ayris, P., & UCL Library Services (2014). [Science 2.0 : Research Data Management]. http://discovery.ucl.ac.uk/1419857/
- Cocaud, S. (2014). *Les plans de gestion des données de la recherche*. Séminaire IST INRA, 2014/11/27, Paris – France
- Creamer, A. (2014). Broader impacts and Data Management Plans. Extrait de: http://escience community.umassmed.edu/2014/11/13/broader-impacts-and-data-management-plans/?utm_source=rss&utm_medium=rss&utm_campaign=broader-impacts-and-data-management-plans
- Van den Eynden, V., Corti, L., Woollard, M., Bishop, L., & Horton, L. (2011). *Managing and sharing data. Best practice for researchers*. University of Essex (UK): UK Data Archive. http://ukdataservice.ac.uk/manage-data/handbook/
- Vines, T. H., Albert, A. Y., Andrew, R. L., Debarre, F., Bock, D. G., Franklin, M. T., . . . Rennison, D. J. (2013). *How Does the Availability of Research Data Change With Time Since Publication?* Meeting Abstract for Seventh International Congress on Peer Review and Biomedical Publication, 2013/09/08-10, Chicago, IL (USA). http://www.peerreviewcongress.org/abstracts_2013.html

_04

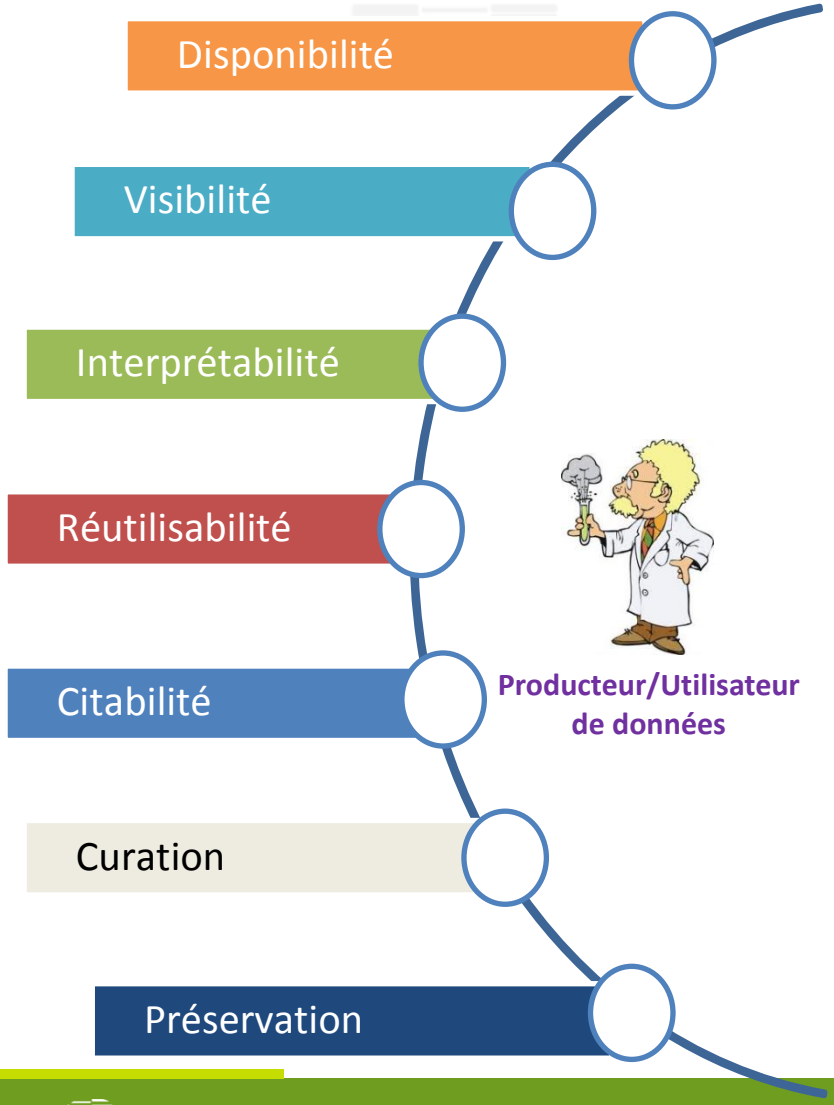
Diffusion des données

Bien gérer les données : préalable à ouverture et partage

"*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*".

Odile Hologne - cité dans : Gaillard, R., 2014. De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ?". Université de Lyon, ENSSIB, Lyon. 1-104 p.
http://www.enssib.fr/bibliotheque-numerique/notices/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche

Rendre visibles, partager ses données…



7 critères ~ qualités à atteindre pour faciliter le partage et la réutilisation des données (*)

Où et comment publier ses données?

- Quelles voies utiliser ?
- Publier dans un Data Paper, quels avantages ?
- Quels coûts ?

Comment décrire et citer ses données ?

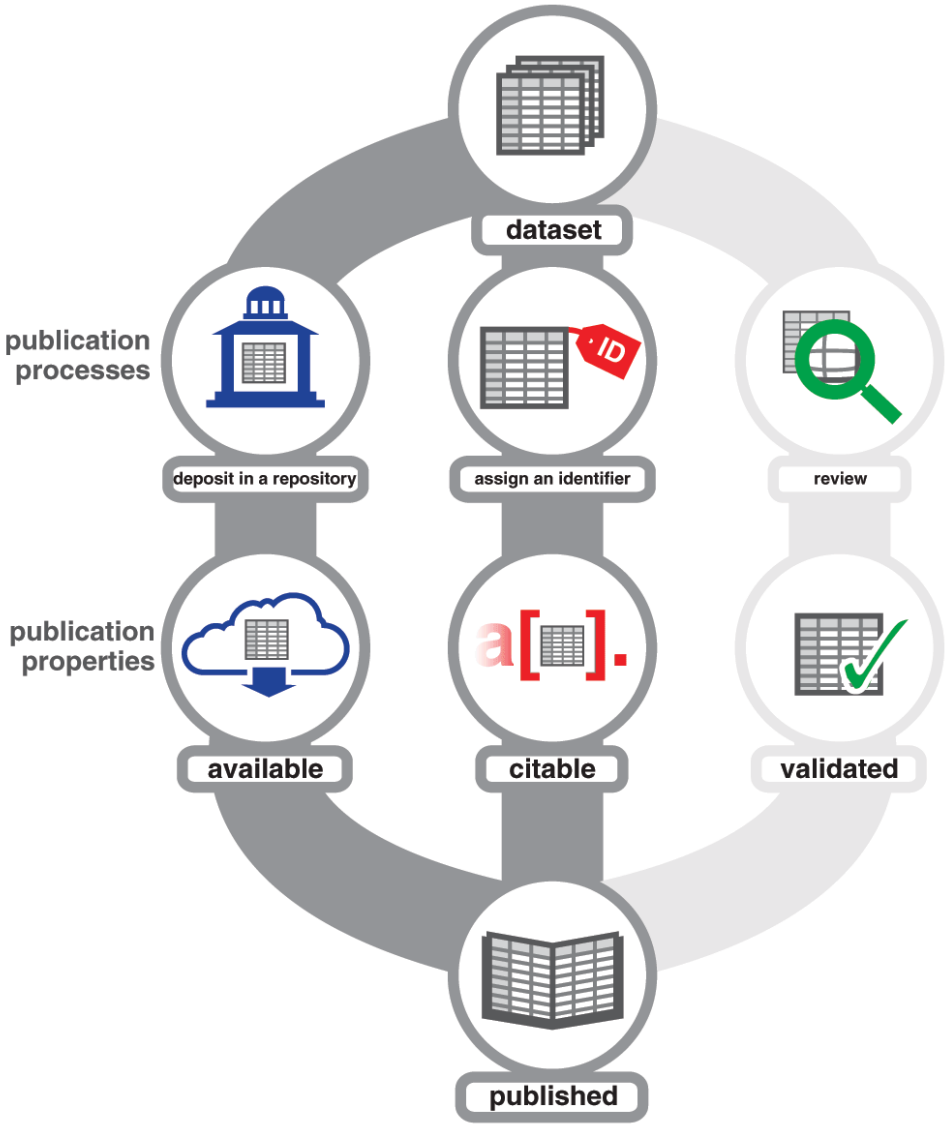
- Quelle documentation? Quelles métadonnées?
- Quel identifiant utiliser ?, Comment l'obtenir ?, Qui peut m'aider ?
- Bonnes pratiques, contraintes : comment les connaitre ?

(*) Rapport ODE = Opportunities for Data Exchange (Reilly, Schallier, Schrimpf, Smit, & Wilkinson, 2011)

4.1- Diffusion des données

Modes de diffusion

Données publiées ?



"published" traduit le fait que les données sont **disponibles** au public (car déposées dans un entrepôt) et **citables** (grâce à un identifiant), mais pas obligatoirement validées par l'existence d'un processus d'évaluation (Peer Review)

F1000Research

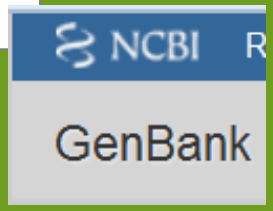
(Kratz & Strasser, 2014)

Donner accès à ses données : plusieurs stratégies …

- Publier (=déposer) ses données dans un entrepôt
- Fournir ses données sous la forme de matériel supplémentaire à la publication
- Publier ses données dans un Data Paper, publication scientifique spécifique décrivant les données, et publié :
  - soit dans une revue «classique»
  - soit dans un **Data Journal** qui peut fournir ou préconiser des entrepôts de confiance.

- Publier dans le web des données (linked data)

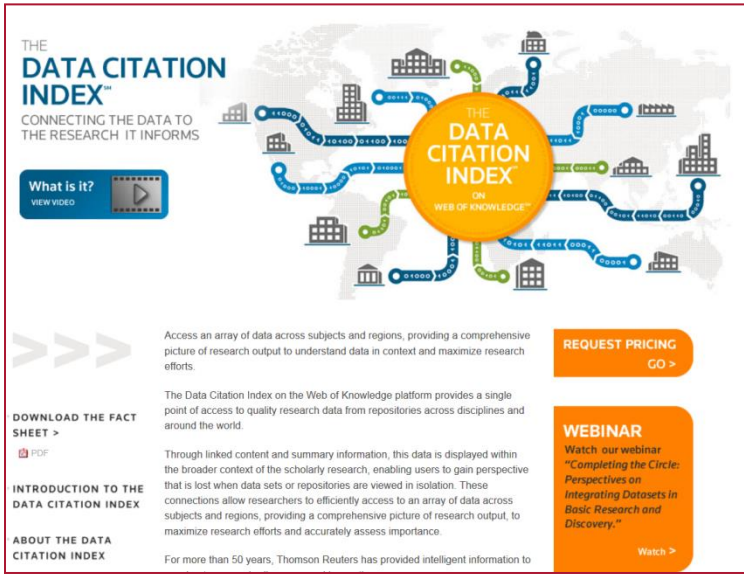
4.1.1- Déposer dans un entrepôt



Comment trouver un entrepôt ?

- Multiplication et hétérogénéité des entrepôts
- Conseil : utiliser des annuaires ou répertoires

Accès gratuit



Accès payant

Comment choisir un entrepôt ?

- Selon les recommandations d'un financeur, d'un éditeur, de son organisme de rattachement
- Selon les types d'entrepôts et leurs caractéristiques
  - Discipline, modèle économique, type d'identification, licence, partenariat éditeurs, certification

Disciplinaire / Propriétaire de l'entrepôt	Institution publique	Organisation à but non lucratif	Organisation à but lucratif
Thématique	PANGAEA GenBank Knowledge Network for Biocomplexity (KNB)	Gene Expression Omnibus (GEO)	
Pluridisciplinaire	Zenodo 3TU.Datacentrum	Dryad Datahub	Figshare

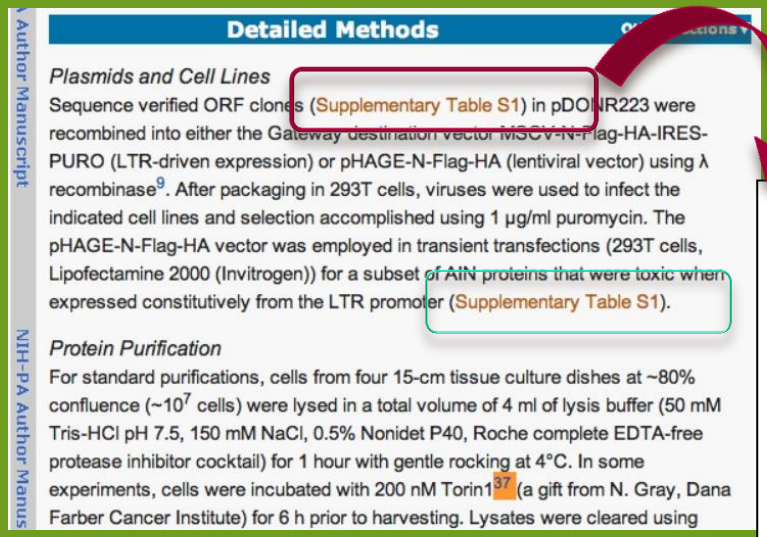
Exemple d'entrepôt

Propriétaire	Dryad
Thématique	Pluridisciplinaire (Biology and Biochemistry; Ecology and Environment; Health and Medicine)
Modèle économique	gratuit pour les chercheurs si le dataset <10GB, coût si >10GB. Sponsorship proposé aux institutions. Voir http://datadryad.org/pages/pricing
Formats de données supportés	Tout format (texte, tableurs, vidéos, photographies, code, y compris des archives compressées de fichiers multiples)
Plateforme utilisée	DSpace
Identifiant attribué au jeu de données	DOI
Licence des données publiées	CC0
Sécurité, persistance, préservation	Partenariat avec CLOCKSS qui garanti l'accès aux données indéfiniment
Accessibilité, réutilisabilité des données	Oui
Compatibilité OAI-PMH pour l'ouverture des données	Oui
Partenariat avec des éditeurs	Liste des éditeurs membres de Dryad : http://datadryad.org/pages/membershipOverview#members Integrated journals table : http://datadryad.org/pages/integratedJournals
Liens externes	Données liées à la fois vers et depuis la publication correspondante. Lorsque c'est approprié, liées également vers et depuis des entrepôts spécialisés (e.g. GenBank).
Contenu de l'entrepôt	Uniquement des données associées à des publications scientifiques
Gestion des versions des fichiers	oui
Gestion/Curation des données	Curation : vérification de l'intégrité des fichiers, vérification de la complétude et de la qualité des métadonnées, conversion des fichiers dans des formats adaptés à la preservation. S'intègre au workflow de soumission du manuscript aux journaux partenaires. Accès réservé aux relecteurs durant la revue par les pairs. Possibilité d'embargo post publication.

Dryad Membership: Members

- American Association for the Advancement of Science *
- American Society of Nauralists *
- The American Genetic Association *
- British Ecological Society *
- BMJ Publishing Group, Ltd. *
- The Biological Journal of the Linnean Society (Linnean Socie
- BioMed Central *
- Ecology Letters *
- Ecological Society of America *
- Elementa: Science of the Anthropocene
- European Society for Evolutionary Biology *
- Evolutionary Applications *
- The Genetics Society *
- German National Libary of Medicine
- HighWire - starting January 1, 2014
- Molecular Ecology *
- Molecular Ecology Resources *
- Molecular Phylogenetics and Evolution *
- Oikos *
- Oxford University Press *
- The Paleontological Society *
- Pensoft Publishers *
- PLOS *
- The Royal Society
- Society for Molecular Biology and Evolution *
- Society for the Study of Evolution *
- Society of Systematic Biologists *
- United States Fish and Wildlife Service *
- Wiley-Blackwell *

4.1.2- Publier des données comme matériel supplémentaire d'un article



Detailed Methods

Plasmids and Cell Lines

Sequence verified ORF clones (Supplementary Table S1) in pDONR223 were recombined into either the Gateway destination vector MSCV-N-Flag-HA-IRES-PURO (LTR-driven expression) or pHAGE-N-Flag-HA (lentiviral vector) using λ recombinase[9]. After packaging in 293T cells, viruses were used to infect the indicated cell lines and selection accomplished using 1 μg/ml puromycin. The pHAGE-N-Flag-HA vector was employed in transient transfections (293T cells, Lipofectamine 2000 (Invitrogen)) for a subset of AIN proteins that were toxic when expressed constitutively from the LTR promoter (Supplementary Table S1).

Protein Purification

For standard purifications, cells from four 15-cm tissue culture dishes at ~80% confluence (~10^7 cells) were lysed in a total volume of 4 ml of lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.5% Nonidet P40, Roche complete EDTA-free protease inhibitor cocktail) for 1 hour with gentle rocking at 4°C. In some experiments, cells were incubated with 200 nM Torin1[37] (a gift from N. Gray, Dana Farber Cancer Institute) for 6 h prior to harvesting. Lysates were cleared using

Part 1. Supplementary Methods

Data Processing and Analysis. Mass spectral data was processed using *CompPASS*, as previously described[1] with modifications discussed below. Briefly, Sequest summary files were processed into a high threshold dataset based on a 2% protein false-positive rate by keeping the XCorr thresholds for each charge state constant while varying the ΔCn (thresholds: XCorr 2+ ≥ 2.5; XCorr 3+ ≥ 3.2; XCorr 4+ ≥ 3.5; +1 charge states were not collected). These processed data sets were merged for each duplicate run and used to populate a "stats table" consisting of each dataset for the AIN as well as 102 unrelated proteins (Dubs and their selected HCIPs[1]; https://harper.hms.harvard.edu/*CompPASS*_Dubs.html). The D^N-score and Z-score are calculated from total spectral counts (TSCs) for each protein found in association with each bait.

Because *CompPASS* was originally designed for analysis of mostly non-reciprocal datasets, we devised a new weighted D^N-score (WD^N-score) (Supplementary Fig. S2), which aids in the identification of HCIPs that are associated with multiple baits in a network. WD^N-scores were calculated as:

$$WD_{i,j} = \sqrt{(\lambda\omega_j)^p (x_{i,j})} \qquad \text{(Eq. 1)}$$

$$\lambda = \left(\frac{k}{\sum_{i=1}^{n1} f_{i,j}}\right), \quad f_{i,j} = \begin{cases} 1 ; x_{i,j} > 0 \\ x_{i,j} \end{cases} \qquad \text{(Eq. 2)}$$

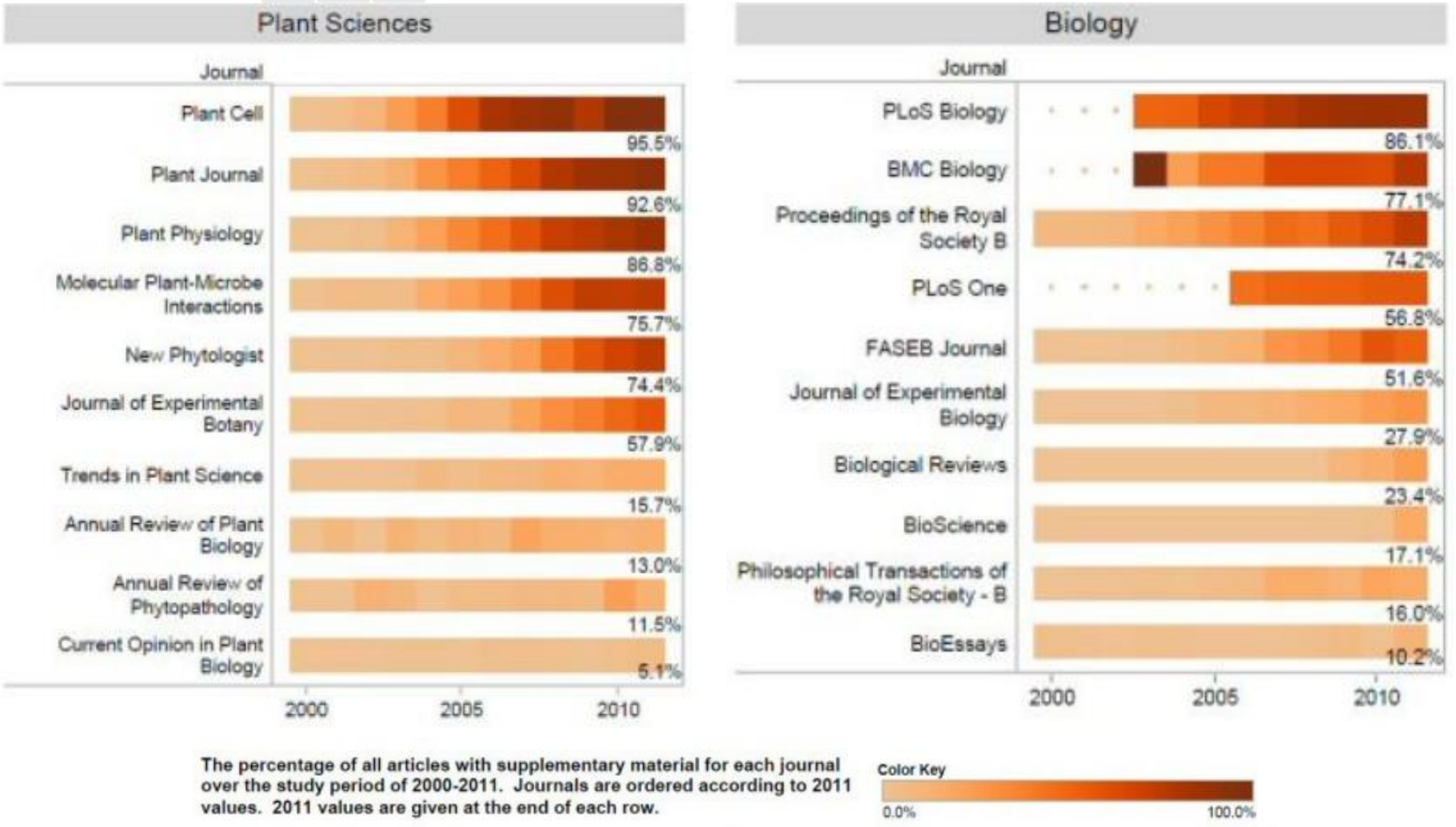
$$\omega_j = \left(\frac{\sigma_j}{\bar{x}_j}\right), \quad \bar{x}_j = \frac{\sum_{i=1,\ldots,n} x_{i,j}}{k} \quad ; n = 1,2,\ldots,m \qquad \text{(Eq. 3)}$$

$x_{i,j}$ = total peptides for interactor j from bait i

Tendances…

- De plus en plus d'articles sont publiés avec du « matériel supplémentaire »
  - Libellés variés : Supplemental material, Supplemental data, Auxiliary information, Supporting information, Supplementary content, Additional content …
  - Contenus variés : fichiers audio, vidéo, images à haute résolution, analyses statistiques, explications méthodologiques approfondies …
  - Mis à disposition par les auteurs ou à la demande des reviewers
- Augmentation significative du nombre de documents supplémentaires dans les revues scientifiques → **impact sur les politiques éditoriales**

The percentage of articles with supplementary materials by journal.



(Kenyon & Sprague, 2014)

L’édition scientifique s’adapte…

- Les données sont (devraient être) de plus en plus considérées comme des produits de la recherche, au même titre que les publications scientifiques,
- Les politiques éditoriales intègrent de plus en plus le dépôt et le partage de données mais dans un contexte d’augmentation des volumes des données, les éditeurs :
  - imposent parfois des limites de taille aux données fournies comme matériel supplémentaire des articles,
  - externalisent la gestion la curation et le stockage des données en recommandant souvent des entrepôts spécifiques (ex : Groupe Nature)
  - certains font de la mise à disposition des données une condition préalable à l’acceptation de l’article (Nature ou PLoS)

EDITORIAL

Making Data Maximally Available

Brooks Hanson is Deputy Editor for physical sciences at *Science*.

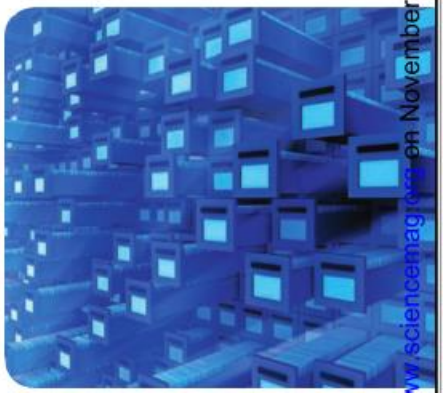
Andrew Sugden is Deputy Editor for biological sciences and International Managing Editor at *Science*.

Bruce Alberts is Editor-in-Chief of *Science*.

SCIENCE IS DRIVEN BY DATA. NEW TECHNOLOGIES HAVE VASTLY INCREASED THE EASE OF DATA collection and consequently the amount of data collected, while also enabling data to be independently mined and reanalyzed by others. And society now relies on scientific data of diverse kinds; for example, in responding to disease outbreaks, managing resources, responding to climate change, and improving transportation. It is obvious that making data widely available is an essential element of scientific research. The scientific community strives to meet its basic responsibilities toward transparency, standardization, and data archiving. Yet, as pointed out in a special section of this issue (pp. 692–729), scientists are struggling with the huge amount, complexity, and variety of the data that are now being produced.

Recognizing the long shelf-life of data and their varied applications, and the close relation of data to the integrity of reported results, publishers, including *Science*, have increasingly assumed more responsibility for ensuring that data are archived and available after publication. Thus, *Science* and other journals have strengthened their policies regarding data, and as publishing moved online, added supporting online material (SOM) to expand data presentation and availability. But it is a growing challenge to ensure that data produced during the course of reported research are appropriately described, standardized, archived, and available to all.

Science's policy for some time has been that "all data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*" (see www.sciencemag.org/site/feature/contribinfo/). Besides prohibiting references to data in unpublished papers (including those described as "in press"), we have encouraged authors to comply in one of two ways: either by depositing data in public databases that are reliably supported and



Hanson, B.; Sugden, A.; Alberts, B., 2011. Making data maximally available. *Science, 331 (6018): 649.* 10.1126/science.1203354

plos.org create account sign in

PLOS ONE | Subject Areas | For Authors | About Us | Search | advanced search

OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

2,649	2	8	100
VIEWS	CITATIONS	SAVES	SHARES

Sine Systemate Chaos? A Versatile Tool for Earthworm Taxonomy: Non-Destructive Imaging of Freshly Fixed and Museum Specimens Using Micro-Computed Tomography

Rosa Fernández, Sebastian Kvist, Jennifer Lenihan, Gonzalo Giribet, Alexander Ziegler

Published: May 16, 2014 • DOI: 10.1371/journal.pone.0096617

Citation: Fernández R, Kvist S, Lenihan J, Giribet G, Ziegler A (2014) *Sine Systemate Chaos?* A Versatile Tool for Earthworm Taxonomy: Non-Destructive Imaging of Freshly Fixed and Museum Specimens Using Micro-Computed Tomography. PLoS ONE 9(5): e96617. doi:10.1371/journal.pone.0096617

Editor: Jacob Guy Bundy, Imperial College London, United Kingdom

Received: December 11, 2013; **Accepted:** April 9, 2014; **Published:** May 16, 2014

Funding: Funding for RF was provided by the Ramón Areces Foundation, for SK by the Wenner-Gren Foundations and Helge Ax:son Johnson's Foundation, and for AZ by the Deutsche Forschungsgemeinschaft through grants no. ZI-1274/1-1 and ZI-1274/1-2. Research was also supported by internal funds from the Museum of Comparative Zoology and from the Department of Invertebrate Zoology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Data Availability: The authors confirm that all data underlying the findings are freely available without restriction. Specimen data may be accessed through the Museum of Comparative Zoology at Harvard University's MCZbase (http://mczbase.mcz.harvard.edu/) under the sample IDs http://mczbase.mcz.harvard.edu/guid/MCZ:IZ:24804, http://mczbase.mcz.harvard.edu/guid/MCZ:IZ:24805, http://mczbase.mcz.harvard.edu/guid/MCZ:IZ:95557, http://mczbase.mcz.harvard.edu/guid/MCZ:IZ:95901. MicroCT scans are available through GigaScience's GigaDB (http://gigadb.org/site/index) at http://dx.doi.org/10.5524/100092.

Exemple : Groupe Nature (1)

authors & referees > Policies > Availability of data & materials

Site content

- Homepage
- Policies
  - Publication ethics
  - Bioethics
  - Availability of data & materials
  - Peer-review policy
  - Embargo
  - Corrections
  - License to publish
  - Feedback

Availability of data and materials

The policy outlined on this page applies to *Nature* journals (those with the word "*Nature*" in their title). NPG publishes many other journals, each of which has separate publication policies described on its website. A current list of these journals, with links to each journal's homepage is available.

Reporting requirements for life sciences research

As of May, 2013, Nature journals require authors of life sciences research papers that are sent for external review to include in their manuscripts relevant details about several elements of experimental and analytical design. This initiative aims to improve the transparency of reporting and the reproducibility of published results. It focuses on elements of methodological information that are frequently poorly reported. During peer review, authors will be asked to confirm that these elements are included in the manuscript by filling out a checklist that will be made available to the editors and reviewers.

Exemple : Checklist / Groupe Nature

Corresponding Author Name: _____

Manuscript Number: _____

Reporting Checklist For Life Sciences Articles

This checklist is used to ensure good reporting standards and to improve the reprodu
please read Reporting Life Sciences Research.

Data deposition

Reported in

17. Provide accession codes for deposited data.

Data deposition in a public repository is mandatory for:

a. Protein, DNA and RNA sequences
b. Macromolecular structures
c. Crystallographic data for small molecules
d. Microarray data

Deposition is strongly recommended for many other datasets for which structured public repositories exist; more details on our data policy are available here. We encourage the provision of other source data in supplementary information or in unstructured repositories such as Figshare and Dryad. We encourage publication of Data Descriptors (see Scientific Data) to maximize data reuse

18. If computer code was used to generate results that are central to the paper's conclusions, include a statement in the Methods section under "**Code availability**" to indicate whether and how the code can be accessed. Include version information as necessary and any restrictions on availability.

November 2014

Mandatory deposition	Suitable repositories
Protein sequences	Uniprot
DNA and RNA sequences	Genbank
	DNA DataBank of Japan (DDBJ)
	EMBL Nucleotide Sequence Database (ENA)
DNA and RNA sequencing data	NCBI Trace Archive
	NCBI Sequence Read Archive (SRA)
Genetic polymorphisms	dbSNP
	dbVar
	European Variation Archive (EVA)
Linked genotype and phenotype data	dbGAP
	The European Genome-phenome Archive (EGA)
Macromolecular structure	Worldwide Protein Data Bank (wwPDB)
	Biological Magnetic Resonance Data Bank (BMRB)
	Electron Microscopy Data Bank (EMDB)
Microarray data (must be MIAME compliant)	Gene Expression Omnibus (GEO)
	ArrayExpress
Crystallographic data for small molecules	Cambridge Structural Database

Nature (Availability of data and materials)
http://www.nature.com/authors/policies/availability.html
A condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to others without undue qualifications. Data sets must be made freely available to readers from the date of publication, and must be provided to editors and peer-reviewers at submission, for the purposes of evaluating the manuscript. For the following types of data set, submission to a community-endorsed, public repository is mandatory. Accession numbers must be provided in the paper. Approximately 40 community-endorsed repositories are listed and include notable ones such as GenBank (DNA Sequences), Protein Data Bank (PDB) (Protein Structures), PANGAEA (Earth and Environmental Sciences) and Cambridge Crystallographic Data Centre (CCDC) (Chemical Structures)

Science (Data and materials availability)
http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail
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). The list of approved repositories is similar to those approved by Nature but this policy states that large data sets with no relevant approved repository be deposited as supplementary material at the Science web site.

Plos One (**Sharing of Data, Materials, and Software**)
http://www.plosone.org/static/policies.action#sharing
Publication is conditional upon the agreement of the authors to make freely available any materials and information described in their publication that may be reasonably requested by others.
Data Availability
PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception[1]. When submitting a manuscript online, authors must provide a *Data Availability Statement* describing compliance with PLOS's policy. If the article is accepted for publication, the data availability statement will be published as part of the final article. Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection. PLOS journal editors encourage researchers to contact them if they encounter difficulties in obtaining data from articles published in PLOS journals. If restrictions on access to data come to light after publication, we reserve the right to post a correction, to contact the authors' institutions and funders, or in extreme cases to retract the publication.

Elsevier (Supplementary data)
http://www.elsevier.com/journals/learning-and-instruction/0959-4752/guide-for-authors#87000
"Elsevier accepts electronic supplementary material to support and enhance scientific research. Supplementary files offer the author additional possibilities to publish supporting applications, high-resolution images, background data sets, sound clips and more. Supplementary files supplied will be published online alongside the electronic version of your article in Elsevier Web products, including ScienceDirect".
Elsevier has, and continues to go one step further than its peers by offering two-way or bi-directional linking between repositories such as PANGAEA, CCDC and Dryad (A Scientific and Medical Repository). Data sets housed in PANGAEA are hyperlinked via Digital Object Identifiers (DOIs) to the related article on the ScienceDirect platform. In reverse, readers are hyperlinked from a banner on the article page in ScienceDirect to the PANGAEA data set record. Similar interoperability exists between Elsevier publications and data sets housed in the CCDC and Dryad. In the summer of 2013 Elsevier added 28 new journals with a reciprocal linking option between its Science Direct platform and Dryad. As ofMarch 2014, approximately 230 scientific and medical journals were linked with Dryad.

Guide de bonnes pratiques !


NISO
How the information world CONNECTS

NFAIS
SERVING THE GLOBAL INFORMATION COMMUNITY

NISO RP-15-2013

Recommended Practices for Online Supplemental Journal Article Materials

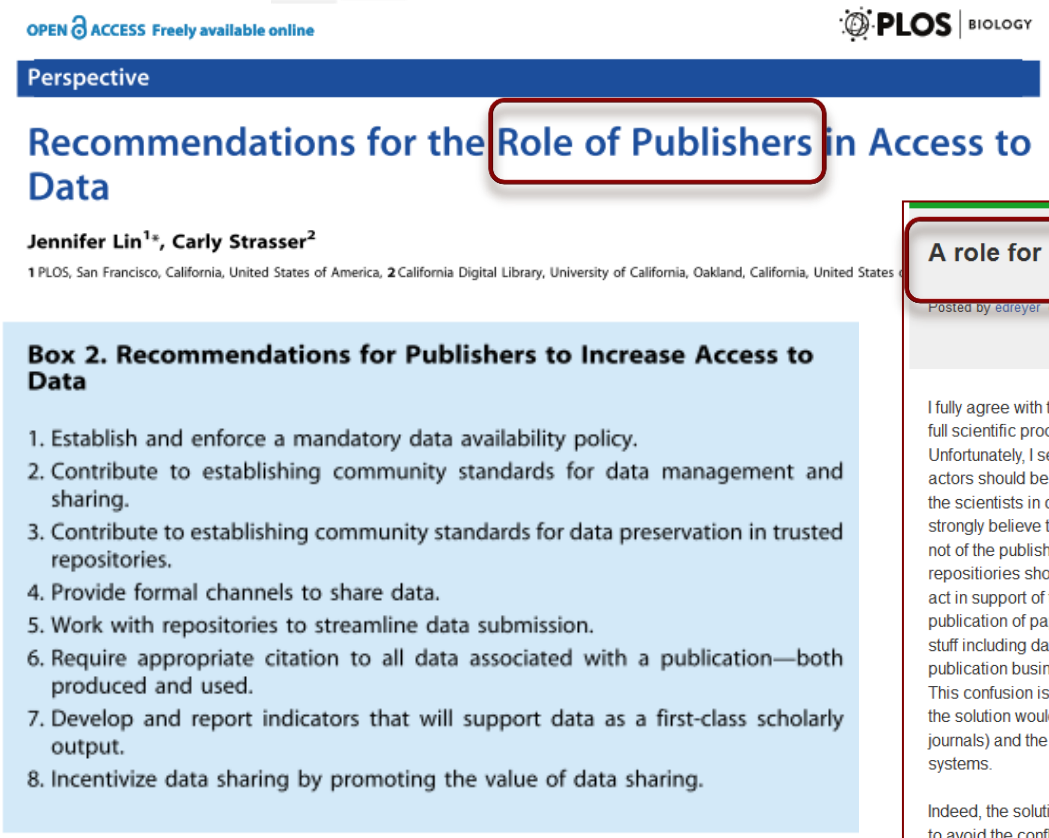
January 2013

A Recommended Practice of the National Information Standards Organization and the National Federation of Advanced Information Services

Recommandations pour les éditeurs, les auteurs, les relecteurs et les **documentalistes** afin de :

- Garantir « discoverability and findability »
- Fournir de bonnes métadonnées,
- Attribuer des identifiants persistants,
- Assurer une qualité d'archivage, de référencement, et de migration des matériaux supplémentaires pour maintenir leur accessibilité à long terme.

Quels rôles pour les différents acteurs ?



OPEN ACCESS Freely available online

PLOS | BIOLOGY

Perspective

Recommendations for the Role of Publishers in Access to Data

Jennifer Lin[1]*, Carly Strasser[2]

1 PLOS, San Francisco, California, United States of America, 2 California Digital Library, University of California, Oakland, California, United States

Box 2. Recommendations for Publishers to Increase Access to Data

1. Establish and enforce a mandatory data availability policy.
2. Contribute to establishing community standards for data management and sharing.
3. Contribute to establishing community standards for data preservation in trusted repositories.
4. Provide formal channels to share data.
5. Work with repositories to streamline data submission.
6. Require appropriate citation to all data associated with a publication—both produced and used.
7. Develop and report indicators that will support data as a first-class scholarly output.
8. Incentivize data sharing by promoting the value of data sharing.

Lin J, Strasser C (2014) Recommendations for the Role of Publishers in Access to Data. PLoS Biol 12(10): e1001975. 10.1371/journal.pbio.1001975

A role for publishers.... or for journal editors?

Posted by edreyer on 30 Oct 2014 at 12:28 GMT

I fully agree with the content of this paper. Institutions and journal editors are getting fumlly aware of the importance of open data as a full scientific production that needs be made available for re use, and for quality control as companions to published papers. Unfortunately, I see an important ambiguity in the whole paper that requires some clarification. It is that we never knwo whether the actors should be the "publishers" (i.e., Elsevier, Springer, francis and Taylor, Nature publishing group, etc) of the journal editors (i.e., the scientists in charge of the editorial policy of the journal and the warrants of the scientific quality of the published material). I strongly believe that this is issue, as well as the issue of the ethics in scientific publication, is a central duty of the journal editors and not of the publishers. The procedues for submission and evaluation, the guidelines and rules for access and for the data repositiories should be a major concerns for the scietists in charge of the editoriazl policy of the journals, and the publishers should act in support of this policy. Otherwise, we would rapidly run into the classical conflict of interest which is already visible todate for the publication of papers: publishers need to publish to make their economic model run. Scientists need to publish relevant and solid stuff including data sets. I will not list again the concerns raised by many scientists about the policy of publishers towards the publication business.
This confusion is a real problem for me, and I would urge the authors of the present "recommendations" to clarify this point. Actually, the solution would rely on a tight cooperation between editors (who should take the lead and develop editorial policies for their journals) and the publishers (who may bring an added value due to their expertise in publication techniques and data management systems.

Indeed, the solution is a clear and transparent cooperation between editors and publishers, each with his aims and responsibilities to avoid the conflict of interest that will arise if we do not clarify this matter.

Erwin Dreyer, Editor of Annals of Forest Science, Inra Nancy (France).

http://ist.blogs.inra.fr/afs/2014/11/01/open-data-a-role-for-publishers-or-for-journal-editors/

4.1.3- Publier un Data Paper

Qu'est ce qu'un Data Paper ?

« Data as the subject of a paper »

- Principal objectif
  - Décrire les données et leur méthode d'obtention, et non les analyser et en tirer des conclusions comme dans un article classique.
- **Valeur ajoutée** :
  - Informer la communauté de l'existence du jeu de données
  - Valoriser les données :
    - en leur apportant une bonne visibilité (objectif d'ouverture)
    - en explicitant leur potentiel de réutilisation
  - Faciliter la réutilisation des données : métadonnées rigoureuses
  - Publication citable → crédit aux auteurs
    (et non considérée comme une publication antérieure par beaucoup de journaux)

Data Paper : caractéristiques

- il **décrit** les données,
  - intègre les métadonnées associées et toutes les informations techniques (méthodes, formules, applications logicielles...) utiles à la compréhension de l'obtention des données et de leur réutilisation par d'autres scientifiques,
- il est associé à un **identifiant** et relié aux données par des **hyperliens pérennes**, éventuellement issus d'entrepôts différents,
- il fait l'objet d'un **examen (reviewing) par des pairs** du domaine disciplinaire (pas toujours)
- il doit pouvoir être corrigé (avec un gain de qualité au fur et à mesure des utilisations, annotations, corrections...). Les corrections doivent alors être expliquées et l'article **versionné.**

Data Paper : publication et structure

- Publiés soit dans des revues classiques soit dans des « Data journals »
- Une structure spécifique parfois minimale

Liens pérennes

General structure

- Titre
- Authors, affiliations
- Abstract
- Keywords
- Context (spatial coverage, temporal coverage)
- Methods
  - Steps, sampling strategy, quality control, contraints, ethical considerations
- Dataset description
  - Object names, data type, format names & versions, creators, creation dates, language, license, location (DOI), publication date
- Reuse potential
- Acknowledgements
- References

Data Paper

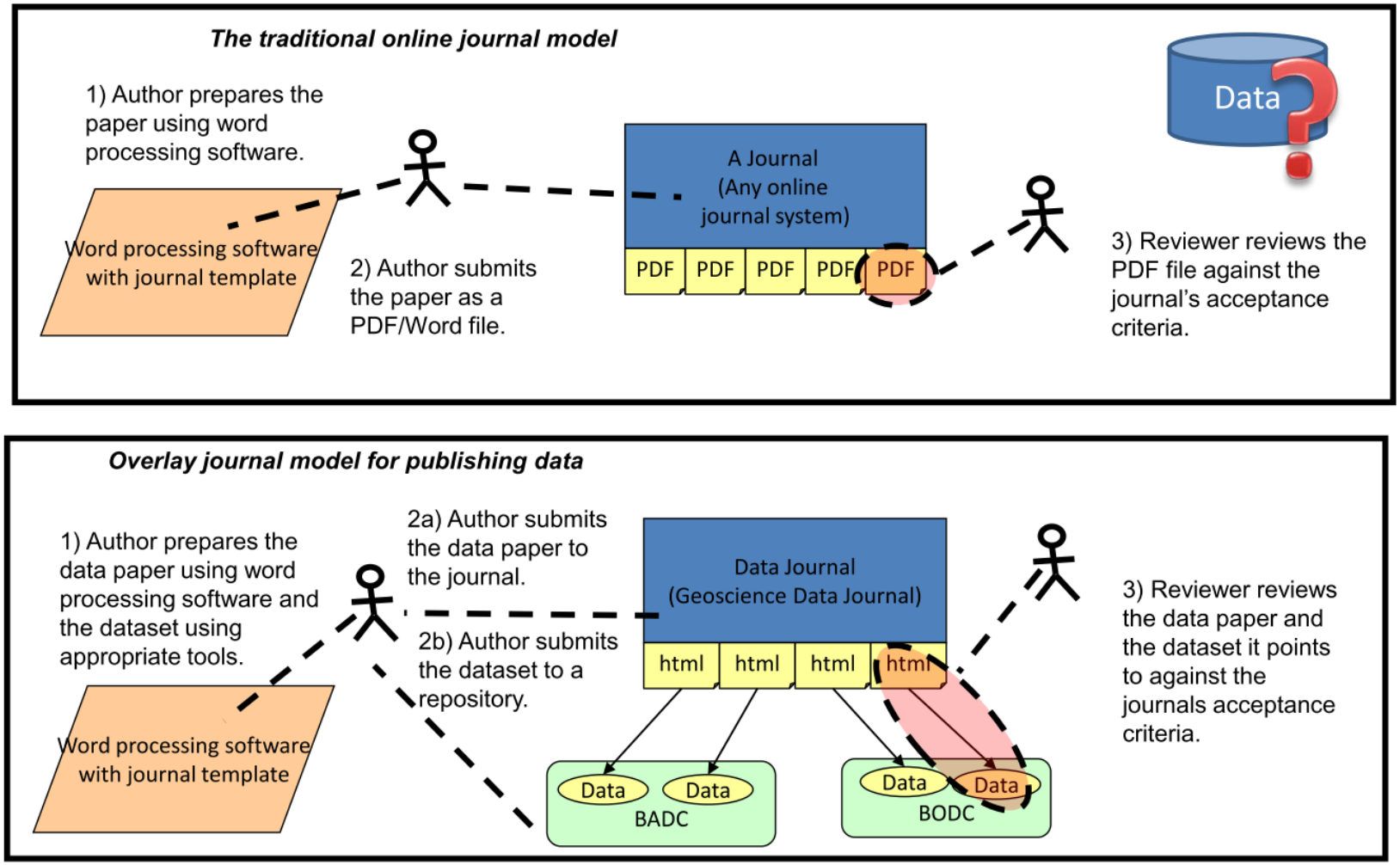
Citation vers les données

Citation vers l'article

- Dépôt au sein du journal ou dans un entrepôt externe (recommandé ou au choix de l'auteur)
- En libre accès ou avec une restriction d'accès temporaire

Dataset

Peer review possible … mais pas généralisé



(Whyte & Callaghan, 2013)

Data Paper dans « Ecology »

Kerry D. Woods 2009. Multi-decade, spatially explicit population studies of canopy dynamics in Michigan old-growth forests. Ecology 90:3587–3587. **http://dx.doi.org/10.1890/09-0565.1**

ECOLOGICAL SOCIETY OF AMERICA

esa ECOLOGY

ESA Home | Ecosphere | Ecology | Ecological Monographs | Ecological Applications | Frontiers | Bulletin | Ecological Archives | Quick Search | Go

All Publications > Ecology > December 2009 > Multi-decade, spatially explicit population studies of canopy dynamics... Advanced Search

Volume 90, Issue 12 (December)

< Previous Next >

Current Issue
Available Issues
Preprints

< Previous Article | Volume 90, Issue 12 (December 2009) | Next Article >

Add to Favorites | Email | Download to Citation Manager | Track Citations | Permissions

PDF

Kerry D. Woods 2009. Multi-decade, spatially explicit population studies of canopy dynamics in Michigan old-growth forests. Ecology 90:3587–3587. http://dx.doi.org/10.1890/09-0565.1

Data Papers

Multi-decade, spatially explicit popul… studies o… canopy dynamics in Michigan old-growth forests
***Ecological Archives* E090-251**

Kerry D. Woods[1]

Natural Sciences, Bennington College, Bennington, Vermont 05201

Established in 1935, a regular grid of 256 permanent plots includes
Natural Area in northern Michigan, USA. Woody stems have been re-
population and community dynamics over periods of up to 72 years.
about half of the study plots, have been mapped and individually trac
dominated by *Fraxinus nigra* and *Thuja occidentalis*. Detailed, long-t
general; this data set is both of exceptional duration and unusual in
regular array over the stand, they can support analyses of spatiotem
2002 provides a unique opportunity to compare disturbance respons
data set have already provided new insights into late-successional p
should permit a range of further comparative and integrative analyse
to the archived data set.

The complete data sets corresponding to abstracts published in t
electronically in *Ecological Archives* at {http://esapubs.org/archive
directly beneath the title.)

Keywords: *Acer saccharum*; *Betula alleghaniensis*; *Fagus grar*
hardwood forest, old-growth forest, permanent plots, succession

Received: March 30, 2009; Revised: July 20, 2009; Accepted: July 22

[1] E-mail: kwoods@bennington.edu

Corresponding Editor: W. K. Michener.

Share this Article
Share |

Journal Information
ISSN: 0012-9658
Frequency: Monthly

Mission and Scope
Types of contributions
Editorial Board
Staff
Instructions for Authors
Reviewer Guidelines
Permissions

Kerry D. Woods. 2009. Multi-decade, spatially explicit population studies of canopy dynamics in Michigan old-growth forests. *Ecology* 90:3587.

Data Paper

***Ecological Archives* E090-251-D1.**

Copyright

Authors
Data Files
Abstract
Metadata

Jeux de données déposés dans « Ecological Archives »

http://esapubs.org/archive/ecol/E090/251/default.htm

Author(s)

Kerry D. Woods
Natural Sciences
Bennington College
Bennington, VT 05201 USA
E-mail: kwoods@bennington.edu

Data Files

Files are ASCII text, tab-delimited. No compression schemes were used.

all_plots_1935_1948.txt -- data for all stems measured in 1935 and 1948.

all_plots_1974-1980.txt -- data for all stems measured in 1974 through 1980.

upland_plots_89-07.txt -- data for upland plots mapped and measured two or more times, 1989 through 2007.

swamp_all_modern.txt -- data for wetland plots censused from 1992 through 2007.

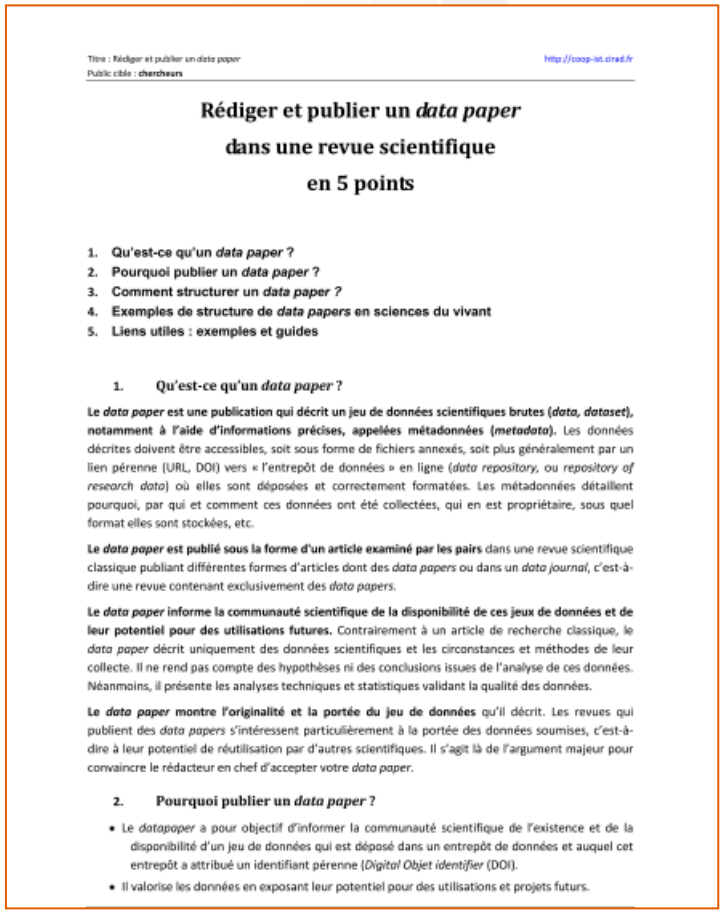
species_codes.txt -- four-letter codes and full names for all species.

sampling_history.txt -- table summarizing sampling history for all plots.

INRA SCIENCE & IMPACT

Données de la re

Exemples (Data Papers, Data Journals)



Titre : Rédiger et publier un data paper — http://coop-ist.cirad.fr
Public cible : chercheurs

Rédiger et publier un *data paper* dans une revue scientifique en 5 points

1. Qu'est-ce qu'un *data paper* ?
2. Pourquoi publier un *data paper* ?
3. Comment structurer un *data paper* ?
4. Exemples de structure de *data papers* en sciences du vivant
5. Liens utiles : exemples et guides

1. Qu'est-ce qu'un *data paper* ?

Le *data paper* est une publication qui décrit un jeu de données scientifiques brutes (*data, dataset*), notamment à l'aide d'informations précises, appelées métadonnées (*metadata*). Les données décrites doivent être accessibles, soit sous forme de fichiers annexés, soit plus généralement par un lien pérenne (URL, DOI) vers « l'entrepôt de données » en ligne (*data repository*, ou *repository of research data*) où elles sont déposées et correctement formatées. Les métadonnées détaillent pourquoi, par qui et comment ces données ont été collectées, qui en est propriétaire, sous quel format elles sont stockées, etc.

Le *data paper* est publié sous la forme d'un article examiné par les pairs dans une revue scientifique classique publiant différentes formes d'articles dont des *data papers* ou dans un *data journal*, c'est-à-dire une revue contenant exclusivement des *data papers*.

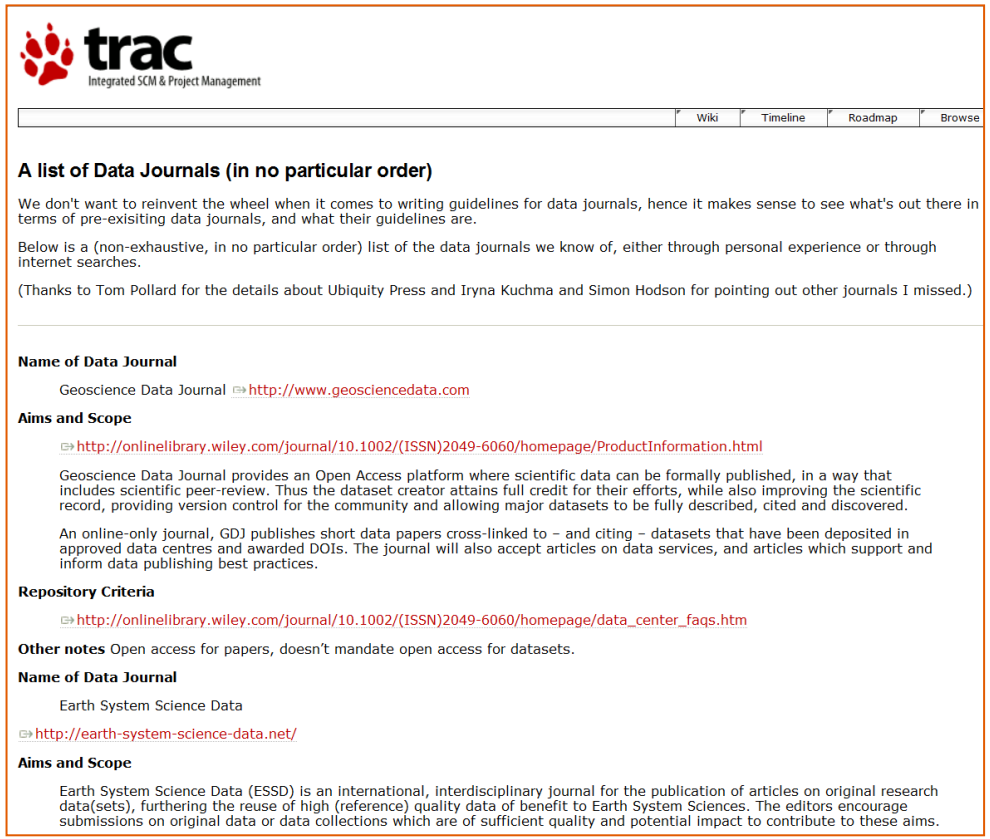
Le *data paper* informe la communauté scientifique de la disponibilité de ces jeux de données et de leur potentiel pour des utilisations futures. Contrairement à un article de recherche classique, le *data paper* décrit uniquement des données scientifiques et les circonstances et méthodes de leur collecte. Il ne rend pas compte des hypothèses ni des conclusions issues de l'analyse de ces données. Néanmoins, il présente les analyses techniques et statistiques validant la qualité des données.

Le *data paper* montre l'originalité et la portée du jeu de données qu'il décrit. Les revues qui publient des *data papers* s'intéressent particulièrement à la portée des données soumises, c'est-à-dire à leur potentiel de réutilisation par d'autres scientifiques. Il s'agit là de l'argument majeur pour convaincre le rédacteur en chef d'accepter votre *data paper*.

2. Pourquoi publier un *data paper* ?

- Le *datapaper* a pour objectif d'informer la communauté scientifique de l'existence et de la disponibilité d'un jeu de données qui est déposé dans un entrepôt de données et auquel cet entrepôt a attribué un identifiant pérenne (*Digital Objet Identifier* (DOI).
- Il valorise les données en exposant leur potentiel pour des utilisations et projets futurs.

Dedieu, L. (2014). *Rédiger et publier un Data Paper en 5 points.* http://url.cirad.fr/ist/data-paper



trac
Integrated SCM & Project Management

Wiki | Timeline | Roadmap | Browse

A list of Data Journals (in no particular order)

We don't want to reinvent the wheel when it comes to writing guidelines for data journals, hence it makes sense to see what's out there in terms of pre-exisiting data journals, and what their guidelines are.

Below is a (non-exhaustive, in no particular order) list of the data journals we know of, either through personal experience or through internet searches.

(Thanks to Tom Pollard for the details about Ubiquity Press and Iryna Kuchma and Simon Hodson for pointing out other journals I missed.)

Name of Data Journal

Geoscience Data Journal http://www.geosciencedata.com

Aims and Scope

http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060/homepage/ProductInformation.html

Geoscience Data Journal provides an Open Access platform where scientific data can be formally published, in a way that includes scientific peer-review. Thus the dataset creator attains full credit for their efforts, while also improving the scientific record, providing version control for the community and allowing major datasets to be fully described, cited and discovered.

An online-only journal, GDJ publishes short data papers cross-linked to – and citing – datasets that have been deposited in approved data centres and awarded DOIs. The journal will also accept articles on data services, and articles which support and inform data publishing best practices.

Repository Criteria

http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060/homepage/data_center_faqs.htm

Other notes Open access for papers, doesn't mandate open access for datasets.

Name of Data Journal

Earth System Science Data

http://earth-system-science-data.net/

Aims and Scope

Earth System Science Data (ESSD) is an international, interdisciplinary journal for the publication of articles on original research data(sets), furthering the reuse of high (reference) quality data of benefit to Earth System Sciences. The editors encourage submissions on original data or data collections which are of sufficient quality and potential impact to contribute to these aims.

http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList

Voies de partage des données	Avantages	Limites
Données intégrées **Articles soumis au peer-review**	• Intégration maximale des données et de l'article : citable, recherchable • Paternité des données / crédits immédiat aux auteurs	• Données difficiles à trouver indépendamment de l'article (réservées à l'abonné selon le modèle éditorial) et dans une forme peu ou pas réutilisable
Données intégrées **Supplementary files** associés à un article	• Bonne intégration des données et de l'article • Format des données libéré des contraintes de rédaction de l'article • Paternité des données / crédits aux auteurs	• Taille souvent limitée • Peu de standardisation sur le signalement des fichiers «supplémentaires» • Identification des données indépendamment de l'article possible (via DOI) mais rare
Données déposées dans des entrepôts reconnus Liens réciproques entre l'article et les données (dépôt soit dans un entrepôt interne à la revue, soit dans un système externe)	• Entrepôts reconnus par une communauté disciplinaire • Données normalisées, standardisées, conservées de façon pérenne • Pas de restriction en volume • Liens réciproques sécurisés	• entrepôts disciplinaires (biologie, sciences de la vie, sciences du sol, chimie), ou génériques (Zenodo …) • Dépend du maintien de financement par les gouvernements (soumis aux aléas budgétaires)

Données publiées dans des
Data Papers
	• Paternité des données / crédits aux auteurs
	• Citation aisée
	• Réutilisation des données facilitée
		• Interrogation sur la qualité :
		  • du Peer-Review ?
		  • des liens bidirectionnels entre données et Data Paper

Valeur ajoutée de la liaison données & publication

- Permet de retrouver plus facilement les données (pointées par des liens),
- « Améliore la valeur contextuelle » des données et de l'article associé
  - Un lecteur peut lire le rôle ou la fonction d'une protéine ou d'un gène particulier dans l'article et accéder facilement via un lien vers les données supplémentaires enregistrées dans un entrepôt approuvé (par exemple Protein Data Bank (PDB) , http://www.rcsb.org/pdb/home/home ou GenBank http://www.ncbi.nlm.nih.gov/genbank/) pour une analyse plus approfondie
- Facilite l'interprétation des données et donc leur réutilisation en fournissant une explication sur les traitements et les résultats obtenus,
- Permet à l'auteur d'avoir une reconnaissance immédiate et supplémentaire de son travail.

9-30% increase depending on e.g. discipline (Piwowar H. et al, 2007, 2013)

Data Citation Advantage ?

Astrophysik

- Henneken, E. A., & Accomazzi, A. (2011). Linking to Data - Effect on Citation Rates in Astronomy. Digital Libraries; Instrumenthttp://www.forschungsdaten.org/index.php/Data_citationation and Methods for Astrophysics. Retrieved from http://arxiv.org/abs/1111.3618v1
- Dorch, B. (2012). On the Citation Advantage of linking to data: Astrophysics. Retrieved from http://hprints.org/hprints-00714715/

Geowissenschaften

- Belter, C.W. (2014). Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. PLOS ONE 9(3): e92590. doi: 10.1371/journal.pone.0092590
- Sears, J. R. (2012). Data Sharing Effect on Article Citation Rate in Paleoceanography. IN53B-1628. AGU Fall Meeting 2011. Retrieved from http://static.coreapps.net/agu2011/html/IN53B-1628.html

Biomedizin

- Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE, 2(3), e308. doi:10.1371/journal.pone.0000308
- Botstein, D. (2010). It's the data! Molecular Biology of the Cell, 21(1), 4–6. doi:10.1091/mbc.E09-07-0575
- Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. PeerJ, 1, e175. doi:10.7717/peerj.175

http://www.forschungsdaten.org/index.php/Data_citation

Deposition Summary — Hide

Authors: Gao, X., Song, J., Galan, J.

Deposition: 2013-04-16
Release: 2013-07-17
Last Mo 4-10-29

RCSB PDB PROTEIN DATA BANK

Primary Citation

Structure and function of the Salmonella Typhi chimaeric A(2)B(5) typhoid toxin.

Song, J., Gao, X., Galan, J.E.

Journal: (2013) Nature **499:** 350-354

PubMed: 23842500
DOI: 10.1038/nature12377
Search Related Articles in PubMed

PubMed Abstract:

Salmonella enterica serovar Typhi (S. Typhi) differs from most other salmonellae in that it causes a life-threatening systemic infection known as t

Molecular Description

Classification: Toxin
Structure Weight: 116819.55
Molecule: Putative pertussis-like toxin subunit
Polymer: 1
Chains: A, B, C, D, E
Fragment: unp residues 24-1
Organism: Salmonella enteric
Gene Names: STY1891 t1107
UniProtKB: Protein Feature

Q8Z6A3
PDB Sites
Secstruc
6L.A

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed — Advanced

Display Settings: Abstract — Send to:

Nature. 2013 Jul 18;499(7458):350-4. doi: 10.1038/nature12377. Epub 2013 Jul 10.

Structure and function of the Salmonella Typhi chimaeric A(2)B(5) typhoid toxin.

Song J[1], Gao X, Galán JE.

Author information

Abstract

Salmonella enterica serovar Typhi (S. Typhi) differs from most other salmonellae in that it causes a life-threatening systemic infection known as typhoid fever. The molecular bases for its unique clinical presentation are unknown. Here we find that the systemic administration of typhoid toxin, a unique virulence factor of S. Typhi, reproduces many of the acute symptoms of typhoid fever in an animal model. We identify specific carbohydrate moieties on specific surface glycoproteins that serve as receptors for typhoid toxin, which explains its broad cell target specificity. We present the atomic structure of typhoid toxin, which shows an unprecedented A2B5 organization with two covalently linked A subunits non-covalently associated pentameric B subunit. The structure provides insight into the toxin's receptor-binding specificity and delivery mechanisms and reveals how the ities of two powerful toxins have been co-opted into a single, unique toxin that can induce many of the symptoms characteristic of typhoid fever. se findings may lead to the development of potentially life-saving therapeutics against typhoid fever.

mment in
teriology: toxins in tandem. [Nature. 2013]

23842500 [PubMed - indexed for MEDLINE] PMCID: PMC4144355 Free PMC Article

nature International weekly journal of science

Structure and function of the *Salmonella* Typhi chimaeric A_2B_5 typhoid toxin

Jeongmin Song, Xiang Gao & Jorge E. Galán

Affiliations | Contributions | Corresponding author

Nature **499**, 350–354 (18 July 2013) | doi:10.1038/nature12377
Received 18 February 2013 | Accepted 12 June 2013 | Published online 10 July 2013

Referenced accessions

Protein Data Bank

PDB 4K6L 3D 4K6L

Supplementary Material

1
Click here to view.(1.8M, pdf)

Supplementary Materials

Conferring Virulence: Structure and Function of the chimeric A_2B_5 Typhoid Toxin

Jeongmin Song, Xiang Gao, and Jorge E. Galán
Department of Microbial Pathogenesis, Yale University School of Medicine, New Haven, CT 06536

This file contains:

*15 Supplementary figures
*4 Supplementary Tables

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4144355/bin/NIHMS492770-supplement-1.pdf

4.1.4- Publier dans le web des données

Que veut dire publier dans le Web de données?

Données liées
Web de données

Open Data

Données sur le Web, dans un format structuré et non propriétaire, (données) identifiées par des URIs et reliées à d'autres

Données sur le Web, dans un format structuré et non propriétaire, (données) identifiées par des URIs

Données sur le Web, dans un format structuré et non propriétaire

Données sur le Web, dans un format structuré

Données sur le Web, quel que soit le format

D'après http://5stardata.info/

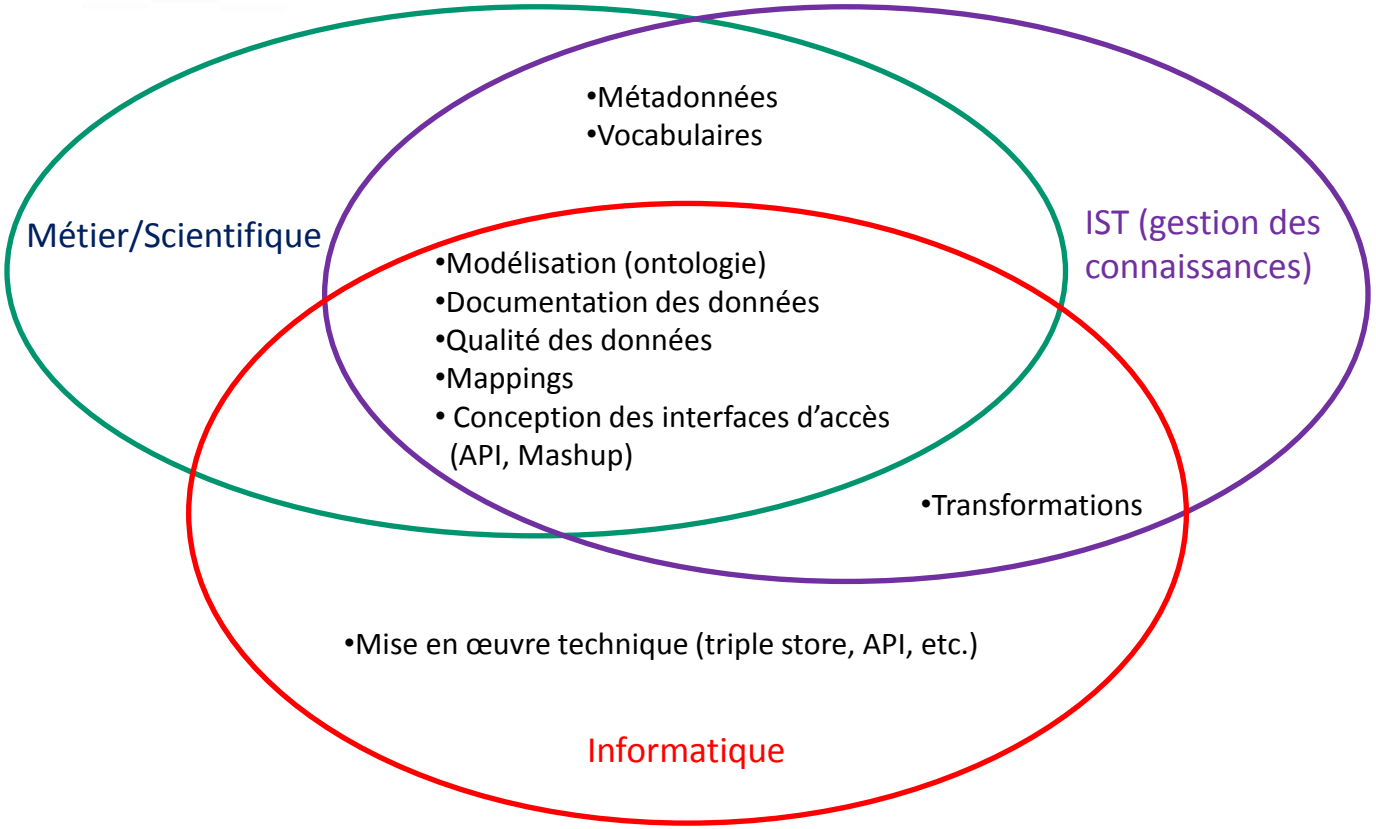
Enjeux

- Favoriser la réutilisation des données y compris par les machines,
- Favoriser l'interopérabilité des données,
- Enrichir, mettre en contexte les données grâce aux liens.

Principales technologies

- **RDF** (Resource Description Framework) pour la représentation des données
- **RDFS** (RDF Schema), **SKOS** (Simple Knowledge Organisation System) pour la description des vocabulaires
- **OWL** (Web Ontology Language) pour la modélisation du domaine;
- **SPARQL** pour l'interrogation des bases de données RDF.

Compétences



EUCLID

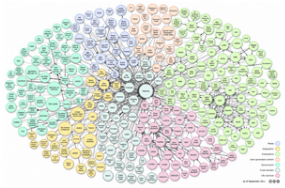
EdUcational Curriculum for the usage of LInked Data

http://www.euclid-project.eu/

Study our modules

MODULE 1 | MODULE 2 | MODULE 3 | MODULE 4 | MODULE 5 | MODULE 6

Module 1: Introduction and Application Scenarios



This module introduces the main principles of Linked Data, the underlying technologies and background standards. It provides basic knowledge for how data can be published over the Web, how it can be queried, and what are the possible use cases and benefits. As an example, we use the development of a music portal (based on the MusicBrainz dataset), which facilitates access to a wide range of information and multimedia resources relating to music. The module also includes some multiple choice questions in the form of a quiz, screencasts of popular tools and embedded videos.

eBook: HTML | iBook | ePUB | Kindle

Course (includes screencasts and exercises): HTML | iTunes U | Slides | Webinar I | Webinar II

Follow our learning pathways

The following table describes which Euclid training modules are most relevant for which data professional training. Naturally, the introductory and intermediate modules are suitable for all data expert types, even if certain topics might be more suitable for a given audience than others.

	Data Architect	Data Manager	Data Analyst	Data Application Developer
Introductory Level	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios	Module 1: Introduction and Application Scenarios
Intermediate Level		Module 2: Querying Linked Data	Module 2: Querying Linked Data	Module 2: Querying Linked Data
	Module 3: Providing Linked Data	Module 3: Providing Linked Data		Module 3: Providing Linked Data
			Module 4: Interaction with Linked Data	Module 4: Interaction with Linked Data
Advanced Level				Module 5: Creating Linked Data Applications
	Module 6: Scaling up	Module 6: Scaling up		

4.2- Citation des données…

Using citations to link research outputs

- We already have a working method for linking between publications which is
  - commonly used
  - understood by the research community
  - used to create metrics to show how much of an impact something has (citation counts)
  - applied to digital objects (digital versions of journal articles)

- We can extend citation to other things like
  - data
  - code
  - multimedia

http://www.flickr.com/photos/anton41/6588935181/

And the best bit is, we don't need to teach researchers a new method of linking – they cite like they normally would!

(Callaghan et al., 2013)

La citation : un élément clé

- Un système robuste de citation :
  - Assure une meilleure visibilité,
  - Facilite la localisation, la découverte et l'accès aux données,
  - Traduit une bonne pérennité des jeux de données (une bonne infrastructure de citation suppose un accès permanent aux données citées dans des entrepôts fiables)
  - Facilite la vérification /validation des résultats de recherche
    - possibilité d'accéder aux données brutes pour les ré-analyser, les contextualiser, tester/évaluer les méthodes utilisées …

Comment citer ses données ?

- Nécessité pour le chercheur/producteur de données de :
  - Connaître et appliquer les standards de citation des données,
  - Utiliser des standards pour décrire les jeux de données et leur provenance afin de permettre aux utilisateurs de juger de la valeur des données,
  - Bien gérer les différentes versions des jeux de données

Comment citer une donnée dans un article ?

La revue recommande un format de citation

- OUI → Suivre la recommandation
- NON → L'entrepôt de stockage de la donnée recommande un format de citation
  - OUI → Suivre la recommandation
  - NON → La donnée est identifiée par un DOI

OUI: utiliser le service DOI formatter de DataCite pour générer un format de citation dans le style qui convient

NON : construire la citation en s'appuyant sur le format standard de Datacite

Creator (PublicationYear): Title, Publisher, Identifier

Ou

Creator (PublicationYear): Title, Version, Publisher, ResourceType, Identifier.

Table 1: DataCite Mandatory Properties

ID	Property
1	Identifier (with type sub-property)
2	Creator (with name identifier and affiliation sub-properties)
3	Title (with optional type sub-properties)
4	Publisher
5	Publication Year

Table 2: DataCite Recommended and Optional Properties

ID	Property
6	Subject (with scheme sub-property)
7	Contributor (with type, name identifier, and affiliation sub-properties)
8	Date (with type sub-property)
9	Language
10	ResourceType (with general type description sub-property)
11	AlternateIdentifier (with type sub-property)
12	RelatedIdentifier (with type and relation type sub-property)
13	Siez
14	Format
15	Version
16	Rights
17	Description (with type sub-property)
18	Geolocation (with point and box sub-properties)

DataCite
DataCite - International Data Citation

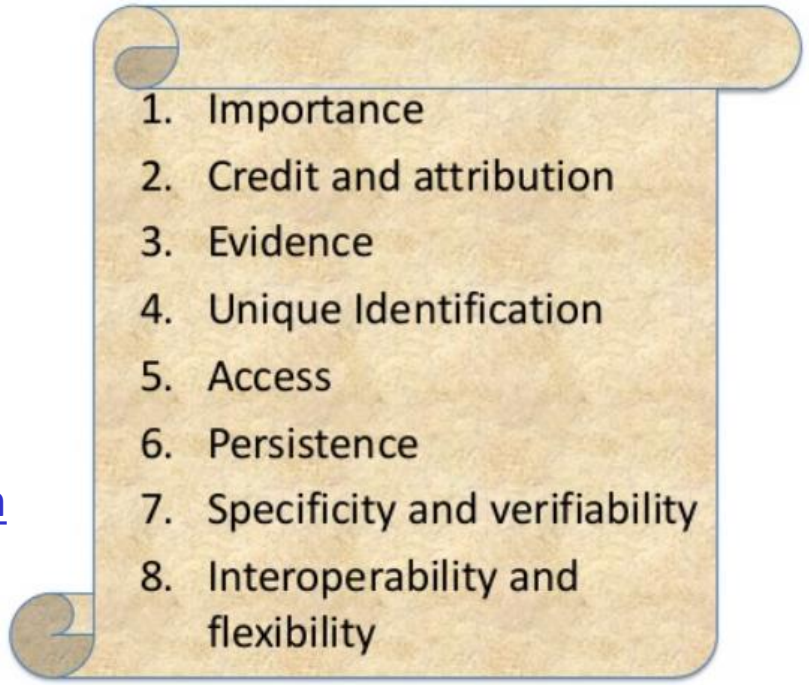
DataCite Metadata Schema for the Publication and Citation of Research Data

http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf

DC1

Data Citation Principles

"Joint Declaration of Data Citation Principles"
Force 11 - 2013
https://www.force11.org/datacitation



"data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.
In support of this assertion, and to encourage good practice, we offer a set of guiding principles for data within scholarly literature, another dataset, or any other research object."

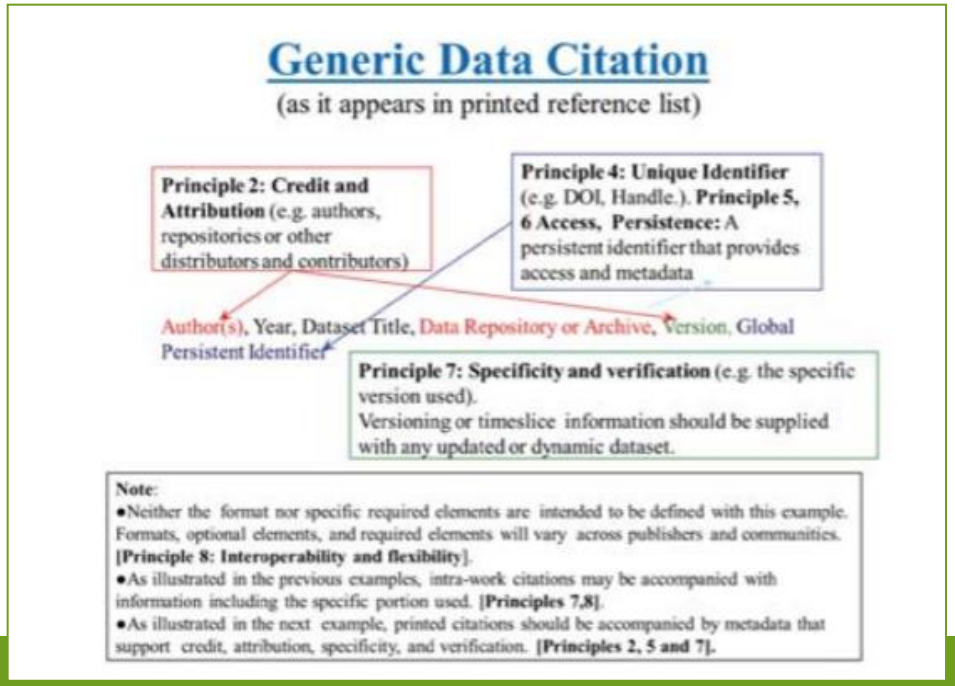
Exemple

Citation / Article

Callaghan, S.A., Waight, J., Agnew, J.L., Walden, C.J., Wrench, C.L. and Ventouras, S. (2013), The GBS dataset: measurements of satellite site diversity at 20.7 GHz in the UK. Geoscience Data Journal. doi: 10.1002/gdj3.

Citation / Dataset

Science and Technology Facilities Council (STFC), Chilbolton Facility for Atmospheric and Radio Research, [Callaghan SA, Waight J,Walden CJ, Agnew J, Ventouras S]. 2009a. GBS 20.7 GHz slant path radio propagation measurements, Sparsholt site. NERC British Atmospheric Data Centre. doi: 10.5285/E8F43A51-0198-4323-A926-FE69225D57DD



Bibliographie - Diffusion des données

- Callaghan, S., Murphy, F., Tedds, J., Allan, R., Kunze, J., Lawrence, R., Mayernik, M. S., & Whyte, A. (2013). Connecting data repositories and publishers for data publication. http://cedadocs.badc.rl.ac.uk/951/1/OAplus_InteroperabilityWorkshop_PREPARDE.pdf
- Callaghan, S., 2014. Preserving the integrity of the scientific record: data citation and linking. *Learned publishing, 27(5), 15-24.* *10.1087/20140504*
- Callaghan, S., 2014. Citing Bytes - Adventures in Data Citation: The Joint Declaration of Data Citation Principles from http://citingbytes.blogspot.fr/2014/10/the-joint-declaration-of-data-citation.html (Blog)
- Dedieu, L., 2014. *Rédiger et publier un Data Paper en 5 points.* *http://url.cirad.fr/ist/data-paper*
- Gruttemeier, H., 2013. *DataCite - identifiants pérennes pour le partage des données. Journées FRéDoc, 3013/10/09, Aussois - France. 62 p.* http://renatis.cnrs.fr/IMG/pdf/DataCite_FreDoc.pdf
- Hrynaszkiewicz, I.; Shintani, Y., 2014. Scientific Data : An open access and open data publication to facilitate reproducible research. *Journal of Information Processing and Management, 57 (9): 629-640.* http://dx.doi.org/10.1241/johokanri.57.629
- Kenyon, J., & Sprague, N. R., 2014. Trends in the Use of Supplementary Materials in Environmental Science Journals. *Issues in Science and Technology Librarianship.* *10.5062/F40Z717Z*
- Kratz, J.; Strasser, C., 2014. Data publication consensus and controversies. *F1000Res, 3: 94.* 10.12688/f1000research.3979.2
- Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). *Report on integration of data and publications* 1-87. http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf
- Van den Eynden, V.; Corti, L.; Woollard, M.; Bishop, L.; Horton, L., 2011. *Managing and sharing data. Best practice for researchers. University of Essex (UK): UK Data Archive. pp. 40.* http://ukdataservice.ac.uk/manage-data/handbook/
- Whyte, A., & Callaghan, S. (2013). *Perspectives on the Role of Trustworthy Repository Standards in Data Journal Publication*. Paper presented at the IASSIST, Cologne - Allemagne. http://fr.slideshare.net/angusawhyte/iassist-preparde-whyte

_05

Partage des données

INRA SCIENCE & IMPACT

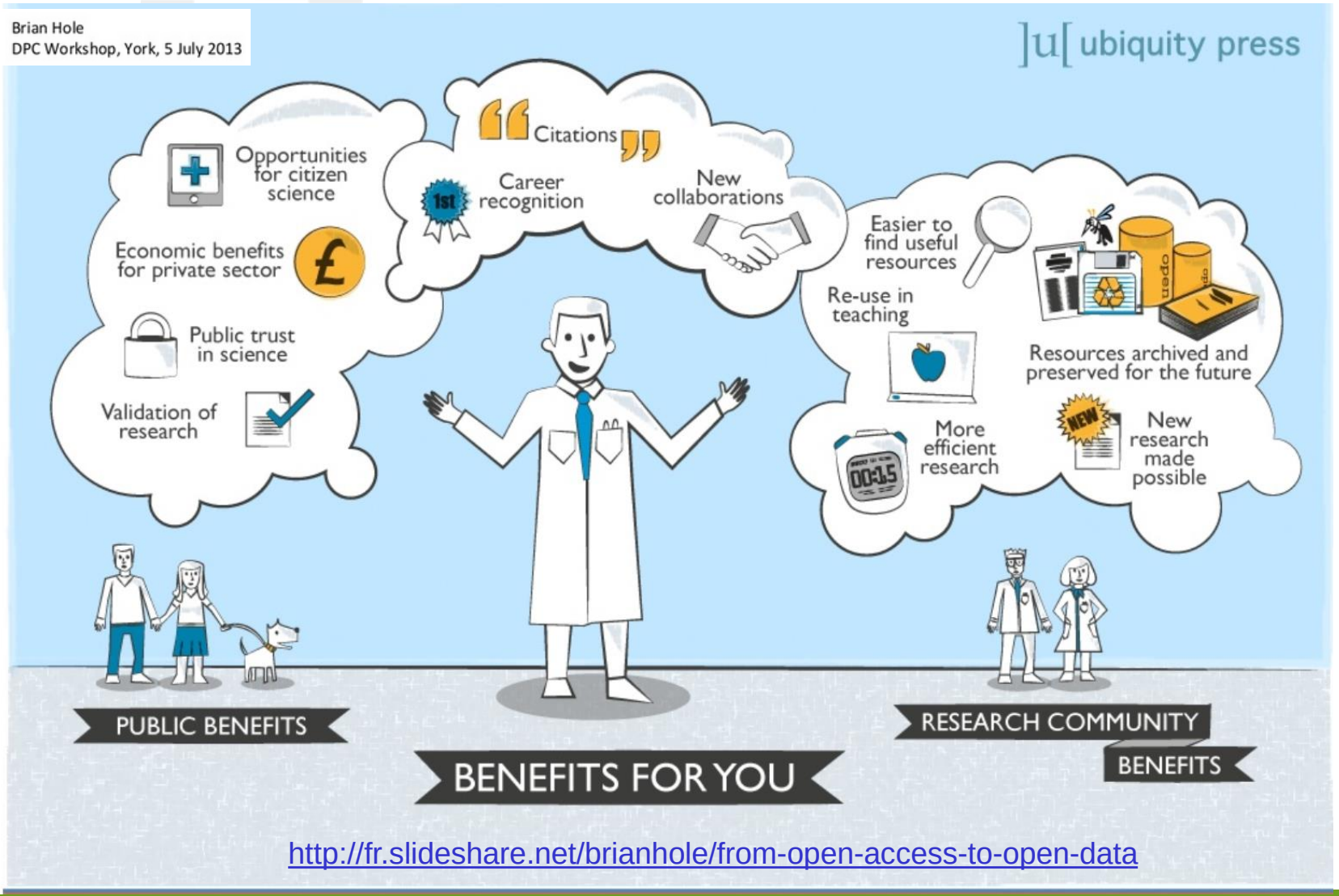
5.1- Partage des données

Enjeux, pourquoi partager ?

Pourquoi ne pas partager ?

INRA SCIENCE & IMPACT

Pourquoi partager les données (enjeux) ?



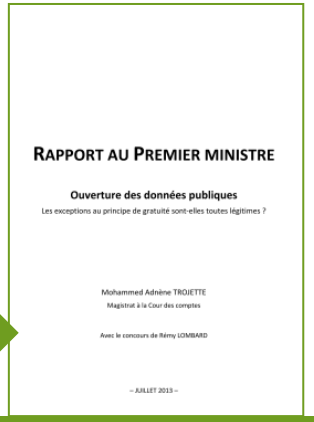
http://fr.slideshare.net/brianhole/from-open-access-to-open-data

Pourquoi partager les données (enjeux) ?

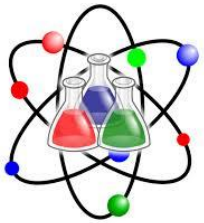
- Enjeux patrimoniaux
  - ✓ Preuve,
  - ✓ Mémoire
- Enjeux économiques

  - ✓ Valeur économique de la donnée
  - ✓ Ré-utilisation gratuite ou payante des données, exploitation des résultats de recherches antérieures
  - ✓ Open data : accélérer l'innovation et le retour sur Investissement dans la R&D

- « *la réutilisation de l'ensemble des données publiques représente aujourd'hui un impact économique et social potentiel évalué à quelques 140 milliards d'euros par an en Europe* »
  http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF
- Consulter le "Rapport Trojette"



Pourquoi partager les données (enjeux) ?



- Enjeux scientifiques
  - ✓ De hypothesis-driven à « data driven innovation » ou « data driven science ».
    Exemples :
    « en sciences dures, par exemple en génomique, avec la fouille d'immenses quantités d'articles scientifiques pour en extraire les données de séquençage qu'ils recèlent. En sciences sociales et humaines : la possibilité de fouiller de grandes masses de données, parfois issues de plusieurs études ou enquêtes statistiques (voire des archives du Web), en sociologie, en économie, permettent de réaliser des avancées considérables, dans une économie de moyens inenvisageable jusqu'alors» (Colcanap & Perales, 2014)
  - ✓ Big data et métiers associés
  - ✓ Changements d'échelle (gène->écosystème), interdisciplinarité

- Enjeux sociétaux
  - ✓ Participation des citoyens et de la société civile (transparence accrue du processus scientifique) – Citizen Science
  - ✓ Confiance dans la recherche

Quand ne pas partager ?

- Quand la loi ne le permet pas…
  En France, cadre juridique de l'ouverture des données publiques et des informations concernées :
  - Loi dite CADA (78-753 – 17/07/78)
  - Loi « Informatique et Libertés » (06/01/78)
  - Loi 51-711 sur les données individuelles couvertes par le secret statistique

- La loi CADA reconnaît à toute personne le droit d'obtenir communication des documents administratifs détenus par une administration ou une personne de droit privé dans le cadre d'une mission de service public. Sont exclus du champ des données communicables (article 6)
  - les documents dont la communication porterait atteinte **« à la protection de la vie privée, au secret médical et au secret en matière commerciale et industrielle »** : les documents comportant des données à caractère personnel doivent être rendus **anonymes** avant communication ;
  - les documents dont la communication porterait préjudice au **bon fonctionnement de l'Etat** (secret de la défense nationale, secret des délibérations du Gouvernement, déroulement des procédures judiciaires engagées, etc. tels que listés dans l'article 6).

Cas des données de santé

(données sensibles)


Commission open data en santé

RAPPORT

Remis à Madame Marisol TOURAINE,
Ministre des Affaires sociales et de la Santé,
Le 9 juillet 2014

Rapport « Commission Open data en santé » (07/14)
Cadre juridique de la protection et de l'utilisation des données personnelles de santé

http://www.drees.sante.gouv.fr/IMG/pdf/rapport_final_commission_open_data-2.pdf

- Défend une ouverture large des données de santé, permettant leur meilleure utilisation.
- Liste les impacts positifs attendus : démocratie sanitaire, autonomie des patients, développement recherche et innovation, amélioration des pratiques professionnelles…
- Pose des limites nécessaires pour garantir la protection de la vie privée des patients
  - Accès libre et soutenu aux données strictement anonymes,
  - Données avec risque faible de ré-identification : procédure simplifiée CNIL après avis d'un comité technique
  - Données détaillées avec risque plus fort de ré-identification : canal unique d'autorisation par la CNIL après avis de différents comités …

5.2- Partage des données

Comment partager (cadre juridique)

Licences

Ouverture et partage des données : complexité juridique

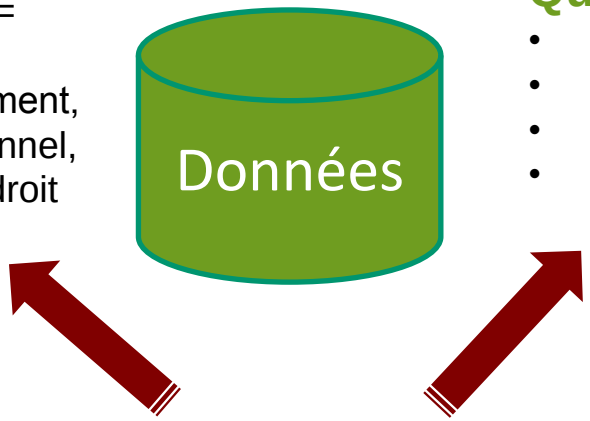
- Droit de propriété des données
  - Les données brutes ne sont a priori pas protégées par le droit d'auteur
  - dans certaines conditions, le droit *sui generis* des bases de données qui protège les collections des données peut/pourrait s'appliquer (mais différemment selon les pays)
- A qui appartiennent les jeux de données, et a-t-on le droit de les rendre publiquement accessibles ?
  - *Rapport pour le conseil scientifique de l'Inra* (Gaspin & Pontier, 2012) « *développement de la recherche au sein de consortiums internationaux regroupant des organismes n'ayant pas les mêmes environnements juridiques ni les mêmes positions stratégiques. Comment respecter les mesures légales et les politiques des différents organismes ou pays ?* »
  - Clause de confidentialité intégrées à certains contrats de recherche
- Droit de réutilisation des données publiques

Problématique

Quel cadre juridique pour les données issues de la recherche ?

Quelles données ?

- Données primaires, brutes = données non réglementées
- Données issues d'un traitement,
- Données à caractère personnel,
- Données protégées par le droit d'auteur



Que et qui protège t-on ?

- Les données ?
- Les personnes ?
- Le travail du chercheur ?
- Des acteurs privés ?

légales

- Protection des personnes (régime spécifique « données sensibles)
- Secret professionnel
- Différences de législations nationales
- Différence d'appréciation ou d'application du doit *sui generis* des bases de données aux données…

Contexte spécifique

- Clause de confidentialité
- Partenaires multiples

répondant à l'intérêt des producteurs

- Privilège d'exploitation
- Droit de citation

D'après : (Moriceau, 2013)

Sous quelles conditions partager ?

- Partenariat
  - Dans le cadre d'un projet de recherche
    - Défini dans le cadre des accords de "consortium" et/ou dans le PGD
  - Hors projet de recherche, dans un cadre contractuel
- Modalités d'accès
  - ouvert ou non (restriction à des groupes spécifiques…)
  - gratuit ou non
- Comment libérer les données et définir des conditions de réutilisation ?
  - «*Dans ce contexte en mouvement et, juridiquement, incertain, le meilleur moyen de définir les obligations juridiques attachées à une donnée est d'avoir recours à une licence ou une waiver*» (Gaillard, 2014)
  - « *Même dans le cas où les données ne sont pas protégées par un droit (droit d'auteur, droit sui generis du producteur de base de données), il est recommandé de les diffuser accompagnées d'un contrat de licence. »* (Battisti, 2010)

Licences (1)

http://fr.wikipedia.org/wiki/Donn%C3%A9es_ouvertes#Les_licences

Open Data Commons - http://opendatacommons.org/guide/
Projet UK (Open Knowledge Foundation)
Licences applicables aux bases de données et aux données prises isolément.
3 types de licences de données libres (droit anglosaxon)

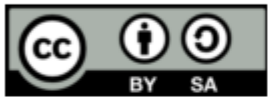
- ***Public Domain Dedication and License*** (PDDL) : Donne la possibilité d'utiliser, copier, modifier, redistribuer une base de données sans aucune restriction. Licence libre de tout droit, de type domaine public. L'auteur abandonne son droit moral.
- ***Open Database Commons*** (ODC-by) : autorise l'utilisation, la copie, la redistribution, la modification, la réalisation de travaux dérivés de la base de données, sous réserve d'indiquer le nom de l'auteur de la base de données originale. On retrouve ces principes dans la licence Creative Commons By.
- ***Open Database License*** : La licence ODC-ODbl[64] est fondée sur le droit d'auteur et le droit sui generis des bases de données. Elle donne la possibilité aux utilisateurs de copier, distribuer, utiliser, modifier et produire une œuvre dérivée à partir d'une base de données sous réserve de la redistribuer sous les mêmes conditions imposées par la licence originale. Elle implique aussi d'indiquer le nom de l'auteur de la base de données d'origine.

Licences (2)

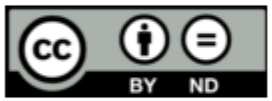
Licences de type Creative Commons : médiatisées et largement utilisées

Domaine public/Pas de droit réservé – Des tiers peuvent librement se baser, améliorer et réutiliser les œuvres pour toute fin sans aucune restriction de copyright ou de loi sur les bases de données.

Attribution – Des tiers peuvent de distribuer, remixer, arranger, et adapter votre œuvre, même à des fins commerciales, tant qu'on vous accorde le mérite de la création originale.



Attribution-Partage dans les Mêmes Conditions – Des tiers peuvent remixer, arranger, et adapter votre œuvre, même à des fins commerciales, tant qu'on vous en accorde le mérite et qu'on diffuse les nouvelles créations selon des conditions identiques.



Attribution-Pas de Modification – Autorise la redistribution, à des fins commerciales ou non, tant que l'œuvre est diffusée sans modification et dans son intégralité, en vous créditant.

Attribution-Pas d'Utilisation Commerciale – Des tiers peuvent remixer, arranger, et adapter votre œuvre à des fins non-commerciales et, bien que les nouvelles œuvres doivent vous créditer et ne pas constituer une utilisation commerciale, elles n'ont pas à être diffusées selon les mêmes conditions.

Attribution-Pas d'Utilisation Commerciale-Partage dans les Mêmes Conditions – Des tiers peuvent remixer, arranger, et adapter votre œuvre à des fins non-commerciales tant qu'on vous crédite et que les nouvelles œuvres sont diffusées selon les mêmes conditions.

Attribution-Pas d'Utilisation Commerciale-Pas de Modification – Des tiers peuvent télécharger vos œuvres et à les partager tant qu'on vous crédite, mais on ne peut les modifier de quelque façon que ce soit ni les utiliser à des fins commerciales.

proche de la PDDL et la plus plébiscitée

Extrait de Dekkers (2013)

Licences (3)
(infos Wikipedia)

- En France

Licence Ouverte / Open Licence https://wiki.data.gouv.fr/wiki/Licence_Ouverte_/_Open_Licence 	Créée par la mission Etalab pour encadrer l'ouverture des données publiques. Compatible avec les licences Open Governement Licence (UK), Open Data Commons Attribution, et Creative Commons Attribution 2.0 (CC) https://www.etalab.gouv.fr/en/qui-sommes-nous
Open Database Licence (ODbL)	Relative aux bases de données. Issue du projet opendatacommons.org de l'Open Knowledge Foundation et traduite en français (Mairie de Paris)

Critères de choix des licences

- Elles peuvent être recommandées, ou imposées :
  - Par l'entrepôt de dépôt des données (Exemple CC0 : Dryad, Figshare …)
  - Par l'éditeur dans le cas d'un Data Paper publié dans un Data Journal (PloS, PeerJ, Biomed Central, certaines branches de Nature et AAAs…)
- Sinon : choisir la licence la plus adaptée à la nature des données (et restrictions éventuelles associées), la volonté d'ouverture et les modalités de réutilisation possibles.

Voir le document :
(Dekkers, Loutas, De Keyzer, & Goedertier, 2013)

5.3- Partage des données

Freins et leviers :
un consensus favorable mais des pratiques en décalage

INRA
SCIENCE & IMPACT

Vidéo

[Open Access Interview] Salma Mesmoudi : Les budgets ont des limites. Pas les idées et la science !

Ouvrir les données de la recherche, pour les budgets, les idées et la visibilité

- Vidéo réalisée dans le cadre de l'Open Access Week 2014 : Salma Mesmoudi explique les avantages que le partage de données peut apporter à tous les niveaux, du chercheur à la science dans son ensemble.

https://www.mysciencework.com/news/11770/open-access-interview-salma-mesmoudi-les-budgets-ont-des-limites-pas-les-idees-et-la-science

https://www.youtube.com/watch?feature=player_embedded&v=ZZpymYuI5OY

Benefits of data sharing	Barriers to data sharing
• Reduction of error and fraud • Increased return on investment in research • Compliance with funder and journal mandates • Reduce duplication and bias • Reproduction/validation of research • Testing additional hypotheses • Use for teaching • Integration with other data sets • Increased citations	• Concerns over inappropriate reuse • Limited time/resources • Costs associated with data sharing • Human privacy concerns • Unclear ownership of data/authority to release data • Lack of academic incentives/recognition • Lack of repositories or lack of awareness of repositories • Protecting commercially sensitive information

(Hrynaszkiewicz & Shintani, 2014)

Motivations du chercheur

« *Faire avancer la science et être crédité pour son travail* »
(Tabor, 2014)

- Par intégrité scientifique
  - Les pairs peuvent vérifier, valider, répliquer, corriger, compléter, etc. les résultats
- Pour accroître son impact
  - Réutilisation des données dans d'autres domaines, pays, secteurs
  - Citation par ceux qui utilisent les données exposées = plus de notoriété → « Open Data Advantage »
- Pour préserver ses données pour ses propres usages futurs
- Pour contribuer à l'enseignement et la formation

Autres arguments du partage

- la faible accessibilité des données de recherche est un frein pour répondre à certaines questions scientifiques,
- différentes interprétations ou approches à partir de données brutes peuvent contribuer au progrès scientifique,
  - en particulier dans les recherches interdisciplinaires comme par exemple les questions liées au changement climatique et au réchauffement planétaire,
- la réutilisation de données brutes permet de valoriser les ressources existantes et d'alléger de nouveaux processus de collecte,
- la disponibilité des données est une garantie contre les problèmes de mauvaise conduite et de falsification,
- la réutilisation des données facilite leur vérification,
- l'accès aux données est justifié en particulier pour les recherches financées par le public.

Freins au partage

- Enquête « Parse Insight » - 2009

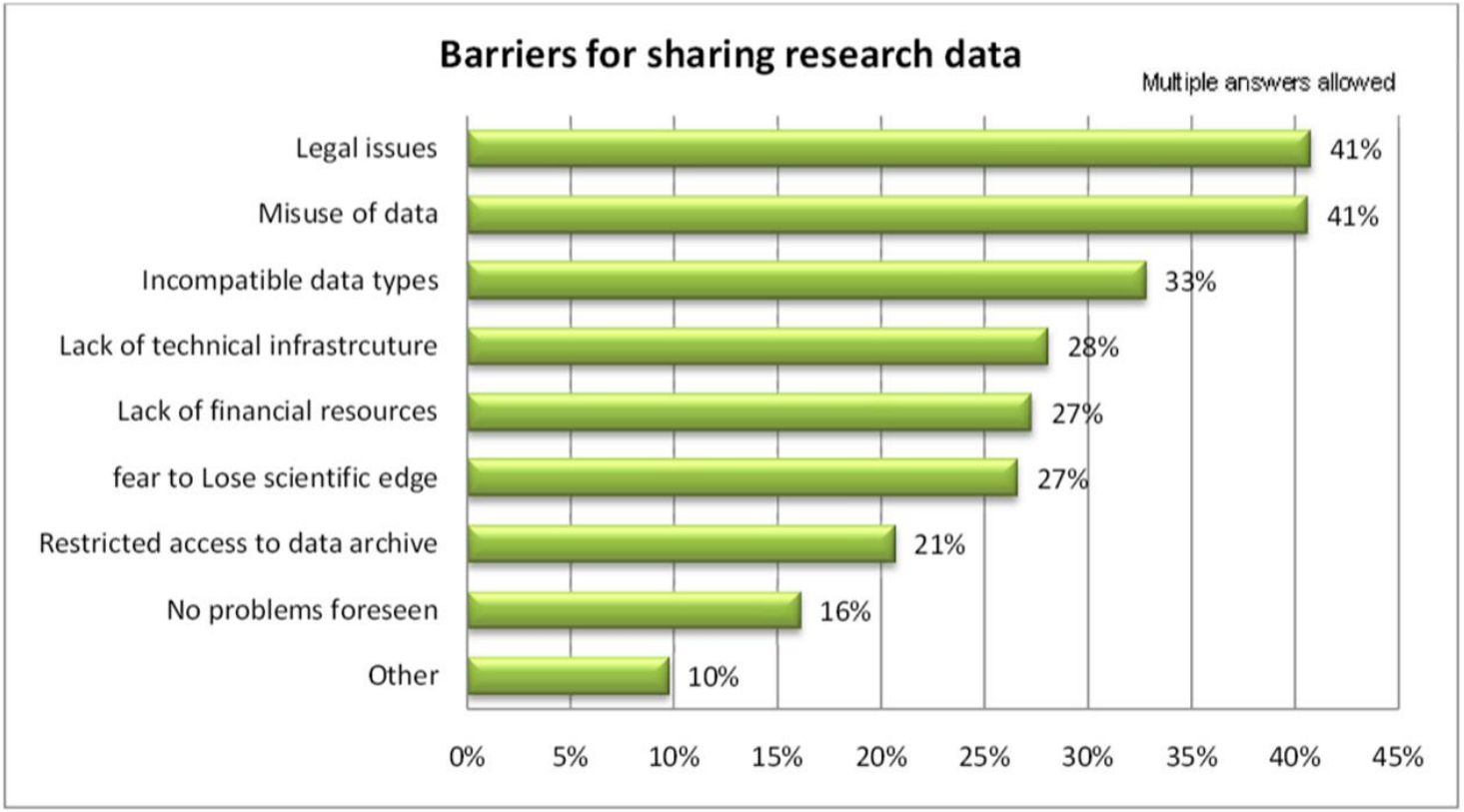


Figure 20: barriers for sharing research data, n = 1270

(Kuipers & Van der Hoeven, 2009)

Table 1 Evidence for barriers to sharing of routinely collected public health data

Category	Barrier	Peer-reviewed		Non peer-reviewed
		Empirical data	Non-empirical*	
Technical	1. Data not collected	[6,21,24,31]	[2,4,7,18,22,14,26-28,30]	[3,23,25]
	2. Data not preserved		[33]	[3,32,34,35]
	3. Data not found		[45]	[3,34]
	4. Language barrier			[36]
	5. Restrictive data format		[40]	[3,34,36-39,41]
	6. Technical solutions not available		[42]	[37]
	7. Lack of metadata and standards	[21,24,43]	[40,44,45]	[1,35-37,39,41,46]
Motivational	8. No incentives		[27,45,49]	[35]
	9. Opportunity cost	[51,52]	[13,33,50,53]	[35]
	10. Possible criticism		[33]	[32]
	11. Disagreement on data use	[21]	[49]	
Economic	12. Possible economic damage		[7,26,27,30]	[55]
	13. Lack of resources	[56,21]	[13,27,28,30,42,53,57]	[3,23,34-36,39,37]
Political	14. Lack of trust	[19,59,60]	[33,61]	[34-37]
	15. Restrictive policies		[30]	
	16. Lack of guidelines		[45,62,65]	[37,41,63,64]
Legal	17. Ownership and copyright		[62,65,66,69]	[37,63,64,67]
	18. Protection of privacy	[12,19,59,73,75]	[44,57,62,66,72,74]	[36,37,64,67,68,70,71]
Ethical	19. Lack of proportionality			[76]
	20. Lack of reciprocity	[51,52]	[50,77,78]	
Number of unique documents (% of total)		14 (21.5%)	30 (46.2%)	21 (32.3%)

** No or little original data presented.*

(Van Panhuis et al., 2014)

Défis pour lever ces freins

- Reconnaissance/rétribution :
  - prise en compte des données dans l'évaluation et la carrière des scientifiques.
- Pérennité : mise en œuvre d'infrastructures pérennes, durables (i.e. des infrastructures dont le fonctionnement et la qualité de service ne se dégradent pas avec les années)
  - Ce problème se pose avec d'autant plus d'acuité que le volume des données augmente (cas en génomique avec Genbank)
  - Le volume des données peut être un facteur limitant le stockage et l'échange (encombrement des BD internationales, avec un temps d'attente important au dépôt)
- Confiance, fiabilité : Les infrastructures doivent être fiables.
  - La provenance, la validation et la fiabilité des données doivent être garanties.
  - La confiance dans les entrepôts ou dans les Data Centers est cependant souvent plus une question de culture et de proximité entre les entrepôts et les communautés scientifiques.
- Références : les données doivent être référençables

5.4- Partage des données

Etat des lieux

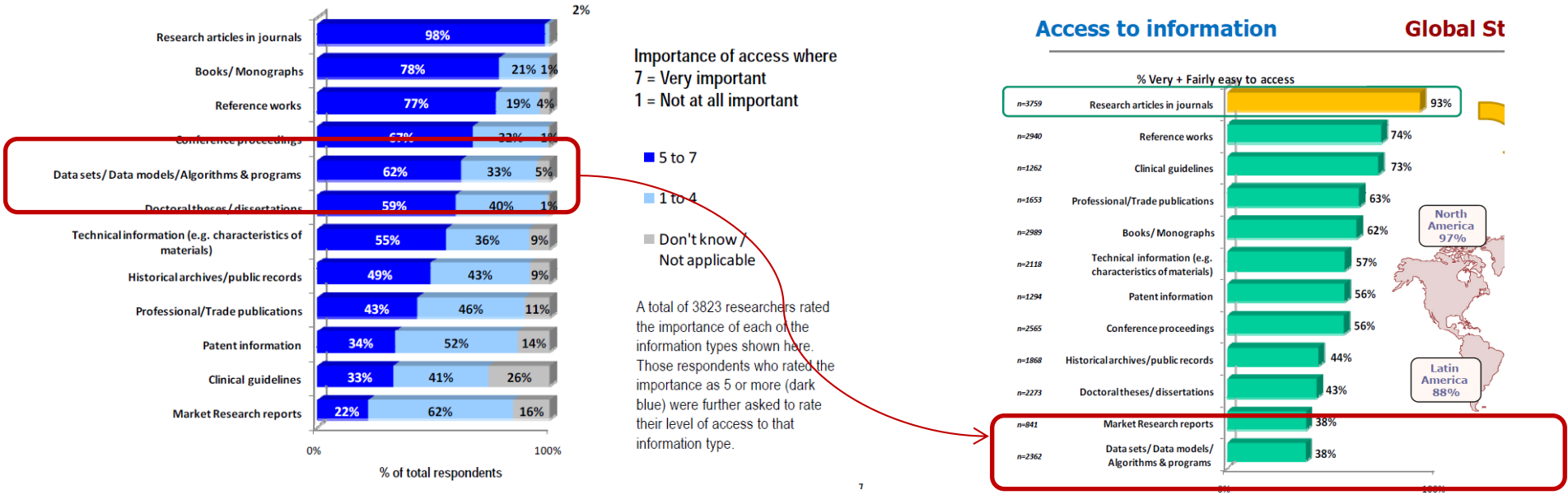
Attentes de la communauté scientifique

84% des chercheurs aimeraient utiliser des jeux de données d'autres chercheurs mais seulement 36% déclarent leurs données accessibles à des tiers

Etude de 2010 auprès de 1300 scientifiques (Monastersky, 2013)

Attentes de la communauté scientifique

- Dans une enquête du Publishing Research Consortium (UK) de 2010 auprès de 3823 chercheurs, l'accès aux données de recherches est estimé comme important ou très important par les répondants. Mais c'est aussi le type d'information jugé le moins accessible.



(Publishing Research Consortium, 2010)

Des pratiques en décalage …

- Enquête « Parse Insight » - 2009
  - ✓ seulement 25 % des répondants partagent leurs données à tout public. 58% les réservent à leur seul groupe de recherche, 11 % aux chercheurs de leur discipline, plus de 20% ne les partagent pas et 6% ne souhaitent pas le faire à l'avenir.

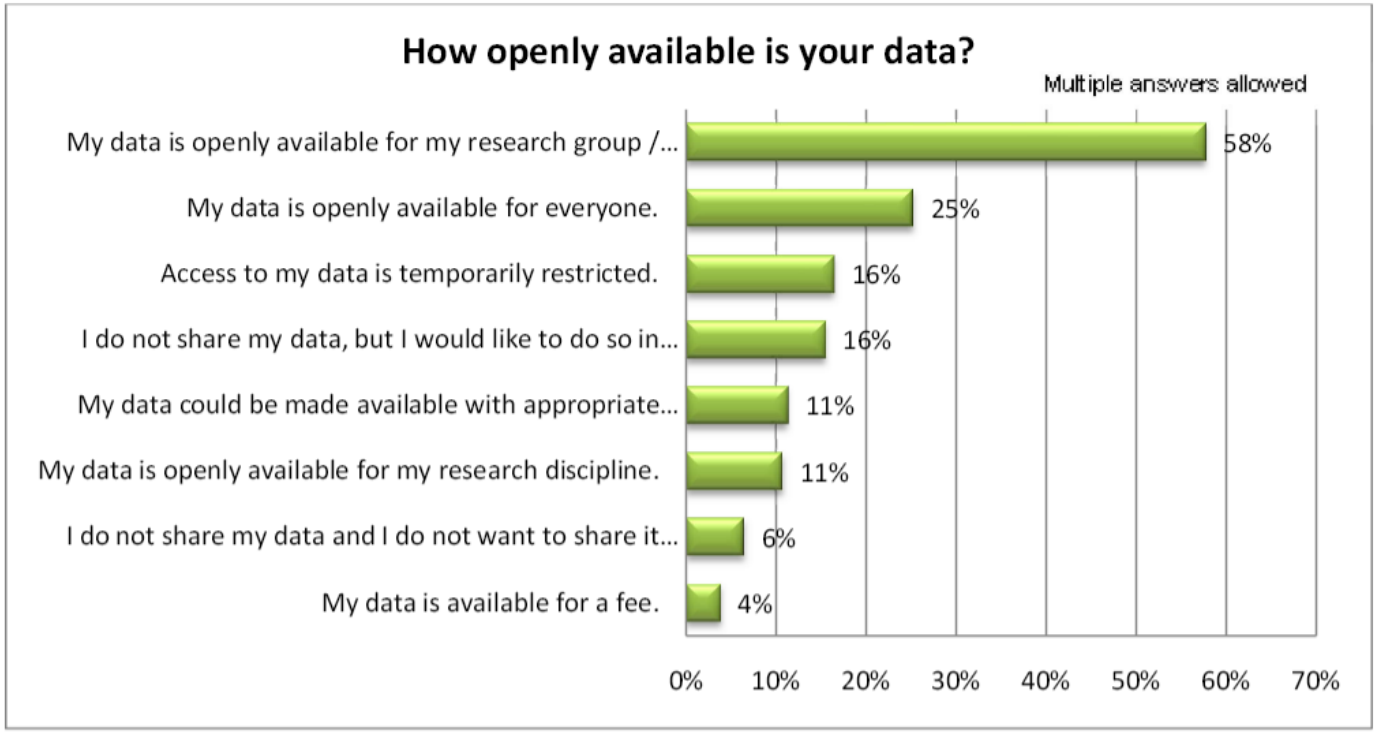


Figure 19: how openly available is your data? n = 1270

(Kuipers & Van der Hoeven, 2009)

RESEARCHER DATA SHARING INSIGHTS

WILEY

Etude 2014

- Wiley's Researcher Data Insights Survey was launched earlier this year to understand how and why researchers make their research data publicly available. The study's results, highlighted below, are intended to advance the global conversation about data sharing and help Wiley better meet the needs of our researchers, authors, and partners in the rapidly evolving landscape of scientific research and communications.
- The survey was deployed in March 2014 and received more than 2,250 responses from researchers around the world.

GLOBAL DATA SHARING TRENDS

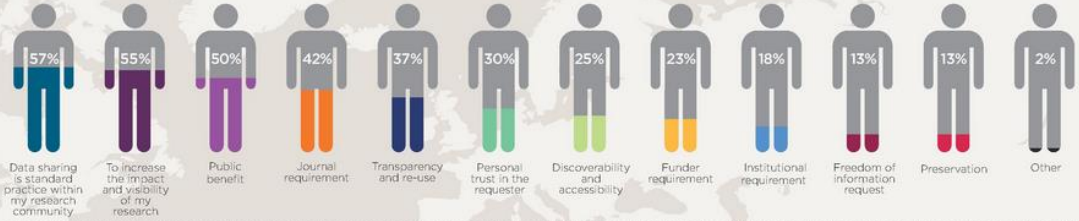
Data sharing practices vary widely across research fields and geographic areas. Just over half of researchers report making their data publicly available, though archiving results in repositories is not yet the norm.

WAYS DATA IS SHARED

- **67%** As supplementary material in a journal
- **37%** Personal, institutional or project webpage
- **26%** Institutional data repository *(i.e. university or institute-sponsored)*
- **19%** Discipline-specific data repository
- **6%** General-purpose data repository *(e.g. Dryad, figshare)*
- ✓ **5%** Other

Globally, researchers also report sharing their data in limited and non-permanent ways: 57% are sharing data at a conference while 42% of researchers share their data upon informal request (e.g. email, direct contact, etc.).

RESEARCHER MOTIVATIONS FOR SHARING DATA



DATA SHARING TRENDS BY COUNTRY

UNITED STATES — 46% SHARING / 54% NOT SHARING

Among researchers in the US sharing their data publicly, two out of three do so because it is standard practice in the communities and because they believe it benefits the public. Similar to their counterparts in the UK, the majority of US-based researchers also share data to increase the impact or visibility of their research.

UNITED KINGDOM — 43% SHARING / 57% NOT SHARING

While more than 40% of UK researchers are sharing data, only about 14% are using discipline-specific or other public repositories like Dryad and figshare. The two key drivers that motivate UK researchers to share their data are the prospect of gaining increased impact or visibility for their work and to satisfy funder requirements.

JAPAN — 44% SHARING / 56% NOT SHARING

Compared with their counterparts around the world, researchers in Japan cite concerns about being scooped as a reason for not sharing data more frequently. Nearly five out of ten Japanese researchers point to this as a reason for not sharing their data, roughly double the global average.

CHINA — 36% SHARING / 64% NOT SHARING

Nearly five in ten Chinese researchers say they are not sharing data because they are not required to do so by their funders or institutions. They are more likely than their global counterparts to say that they do not see data sharing as a personal responsibility and plan to take direction from funders to guide their data sharing decisions in the future.

BRAZIL — 52% SHARING / 48% NOT SHARING

Two out of three researchers in Brazil say that a guarantee of proper credit or attribution would compel them to share more of their data publicly in the future.

AUSTRALIA — 41% SHARING / 59% NOT SHARING

Researchers in Australia say they would be most incentivized to make their data accessible in the future to ensure preservation as well as transparency and re-use. The majority of researchers also ranked funder requirements among top reasons to share in the future.

GERMANY — 55% SHARING / 45% NOT SHARING

Among German researchers sharing their data publicly, three out of four are driven to share data because they believe it will increase the visibility of their research and want to ensure public transparency and re-use. About 20% of German researchers are making use of general purpose repositories (like figshare and Dryad), significantly more than their counterparts around the world, including those in the US and UK.

REASONS WHY RESEARCHERS ARE HESITANT TO SHARE THEIR DATA

- **42%** Intellectual property or confidentiality issues
- **36%** My funder/institution does not require data sharing
- **26%** I am concerned that my research will be scooped
- **26%** I am concerned about misinterpretation or misuse
- **23%** Ethical concerns
- **22%** I am concerned about being given proper citation credit or attribution
- **21%** I did not know where to share my data
- **20%** Insufficient time and/or resources
- **16%** I did not know how to share my data
- **12%** I don't think it is my responsibility
- **12%** I did not consider the data to be relevant
- **11%** Lack of funding
- **7%** Other

DATA SHARING BY DISCIPLINE

Data sharing, specifically by way of data repositories, is most prevalent amongst life scientists, particularly those in the earth and environmental and agriculture and food sciences.

Where Health Scientists share their work:

- **68%** As supplementary material in a journal
- **29%** Personal/institutional/lab webpages
- **29%** Institutional data repositories *(i.e. university or institute-sponsored)*
- **21%** Discipline-specific data repositories
- **5%** General-purpose data repositories *(e.g. Dryad, figshare)*

A typical *Health Science researcher* says she would be motivated to share her data in the future in order to benefit the public, so long as privacy and ethical concerns are assuaged.

Where Life Scientists share their work:

- **76%** As supplementary material in a journal
- **42%** Discipline-specific data repositories
- **29%** Personal/institutional/lab webpages
- **23%** Institutional data repositories *(i.e. university or institute-sponsored)*
- **13%** General-purpose data repositories *(e.g. Dryad, figshare)*

A typical *Life Science researcher* says she would be motivated to share more of her data in the future if she was guaranteed proper credit.

Where Physical Scientists share their work:

- **69%** As supplementary material in a journal
- **41%** Personal/institutional/lab webpages
- **28%** Institutional data repositories *(i.e. university or institute-sponsored)*
- **10%** Discipline-specific data repositories
- **3%** General-purpose data repositories *(e.g. Dryad, figshare)*

A typical *Physical Science researcher* says she would be motivated to share her data in the future because it is standard practice within her research community and because it increases the impact and visibility of her work.

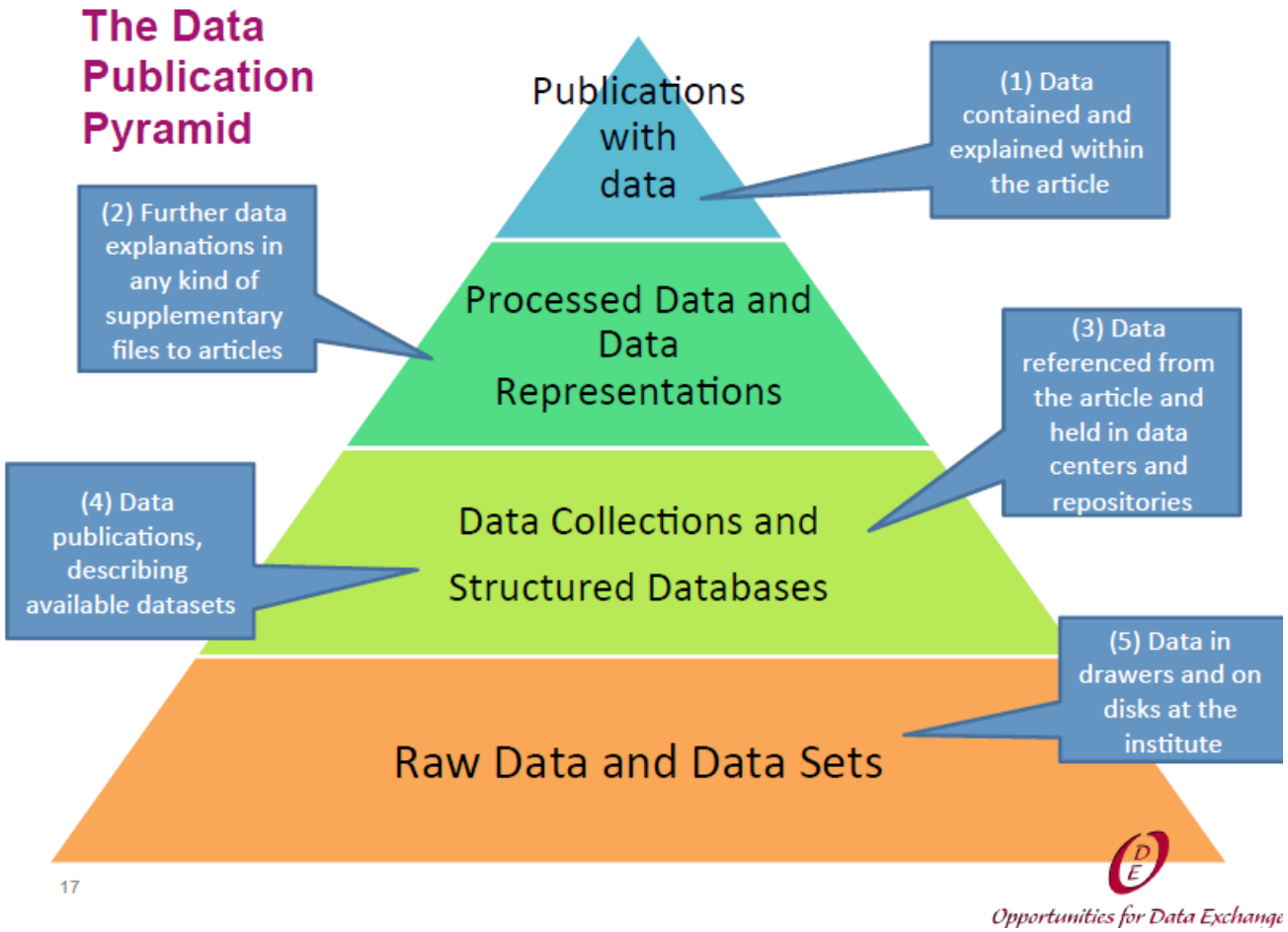
Where Social Scientists share their work:

- **52%** As supplementary material in a journal
- **51%** Personal/institutional/lab webpages
- **25%** Institutional data repositories *(i.e. university or institute-sponsored)*
- **3%** General-purpose data repositories *(e.g. Dryad, figshare)*
- **2%** Discipline-specific data repositories

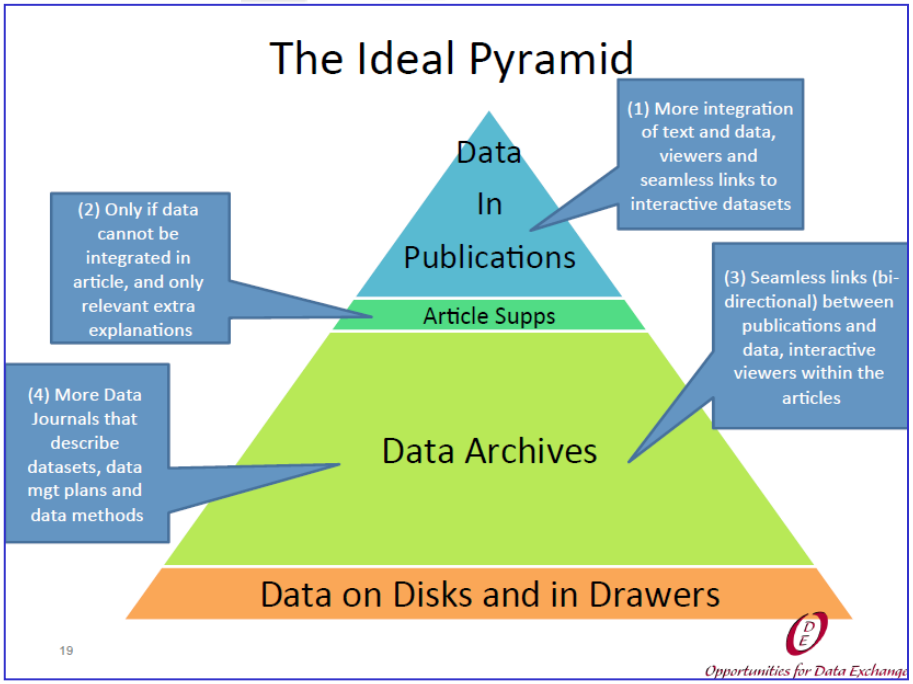
A typical *Social Science and Humanities researcher* says she would be motivated to share her data in the future if it increased the impact and visibility of her work or if she was required to by her funder.

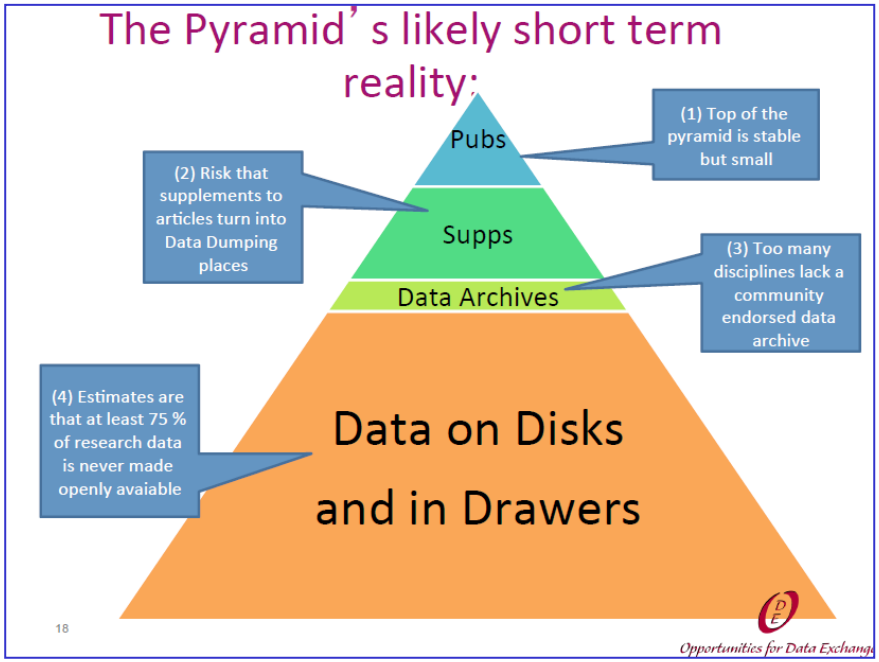
(Ferguson, 2014)

http://exchanges.wiley.com/blog/wp-content/uploads/2014/11/Researcher-Data-Insights-Infographic-FINAL-REVISED-2.jpg



(Callaghan, 2013) D'après : (Reilly, Schallier, Schrimpf, Smit, & Wilkinson, 2011)





(Callaghan, 2013)

Bibliographie – Partage des données (1)

- Battisti, M. (2010). Libérons les données ! De quelques aspects juridiques. Retrieved from http://www.paralipomenes.net/archives/622
- Colcanap, G., & Perales, C. (2014). *CSPLA - Mission relative au data mining (exploration de données) : l'analyse de Couperin et de l'ADBU* 1-10. http://adbu.fr/wp-content/uploads/2014/04/Audition_CSPLA_TDM_2014_04_04_final.pdf
- Callaghan, S. (2013). *Datasets : from creation to publication or "A tale of two datasets"*. Paper presented at the LCPD13 Workshop, Valetta, Malta. http://lcpd2013.research-infrastructures.eu/slides/Callaghan.pdf
- Couture, M., 2010 La diffusion et le partage de l'information scientifique. In: Couture, M., Dubé, M., and Malissard, P., eds. *Propriété intellectuelle et université. Entre la libre circulation des idées et la privatisation des savoirs. Québec: Presses de l'Université du Québec. pp. 1-24.* http://archipel.uqam.ca/3460
- Commission open data en santé. (2014). *Rapport remis à Madame Marisol TOURAINE, Ministre des Affaires sociales et de la Santé, Le 9 juillet 2014* 1-63. http://www.drees.sante.gouv.fr/IMG/pdf/rapport_final_commission_open_data-2.pdf
- Dekkers, M.; Loutas, N.; De Keyzer, M.; Goedertier, S., 2013. Licences pour les données et les métadonnées. Module de formation 2.5. http://fr.slideshare.net/OpenDataSupport/licences-pour-les-donnes-et-les-mtadonnes
- European Commission, 2011. Communication from the Commission to the European parliament, the council, the European Economic and Social Committee and the Committee of the regions. Open data : An engine for innovation, growth and transparent governance 1-13 p. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF
- Gaillard, R., 2014. De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ?". Université de Lyon, ENSSIB, Lyon. 1-104 p. http://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf

Bibliographie – Partage des données (2)

- Ferguson, L. (2014). How and why researchers share data (and why they don't). Retrieved from http://exchanges.wiley.com/blog/2014/11/03/how-and-why-researchers-share-data-and-why-they-dont/
- Gaspin, C., & Pontier, D. (2012). *Rapport du groupe de travail sur la gestion et le partage des données* 1-62. http://www.pfl-cepia.inra.fr/uploads/gdp_docs/Rapport-GestionDonnees-web.pdf
- Hole, B. (2013). *From Open Access to Open Data*. DPC Workshop. http://fr.slideshare.net/brianhole/from-open-access-to-open-data
- Hrynaszkiewicz, I., & Shintani, Y. (2014). Scientific Data : An open access and open data publication to facilitate reproducible research. *Journal of Information Processing and Management, 57*(9), 629-640. http://dx.doi.org/10.1241/johokanri.57.629
- Kuipers, T., & Van der Hoeven, J. (2009). *Insight into digital preservation of research output in Europe. Survey Report PARSE.Insight: INSIGHT into issues of Permanent Access to the Records of Science in Europe* 1-83. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf
- Loi n° 78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal (1978). http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068643&dateTexte=20080929
- Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (1978). http://www.cnil.fr/documentation/textes-fondateurs/loi78-17/

Bibliographie - Partage des données (3)

- Monastersky, R. (2013). The library reboot. As scientific publishing moves to embrace open data, libraries and researchers are trying to keep up. *Nature, 495*, 430-432 http://www.nature.com/news/publishing-frontiers-the-library-reboot-1.12664
- Moriceau, A. (2013). *Quel droit pour les données de la recherche ? la délicate question de leur mise à disposition*. Paper presented at the Gestion et valorisation des données de la recherche. FRéDoc 2013 Aussois - France. http://renatis.cnrs.fr/IMG/pdf/Moriceau_droit_Fredoc2013.pdf
- Publishing Research Consortium. (2010). *Access vs. importance*. http://publishingresearch.net/index.php?view=download&alias=19-prc-access-vs-importance&option=com_docman&Itemid=815
- Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). *Report on integration of data and publications* 1-87. http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf
- Tabor, A. (2014). Le guide Creative Commons du partage scientifique. Puneet Kishor de Creative Commons explique pourquoi les chercheurs ont intérêt à partager leurs travaux sous licence CC, et surtout comment s'y prendre. Retrieved from https://www.mysciencework.com/news/11680/le-guide-creative-commons-du-partage-scientifique
- Trojette, M. A., & Lombard, R. (2013). *Ouverture des données publiques. Les exceptions au principe de gratuité sont-elles toutes légitimes ? Rapport au premier ministre* 1-121. http://www.modernisation.gouv.fr/sites/default/files/fichiers-attaches/20131105-rapporttrojetteannexes.pdf
- Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., Heymann, D., & Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC Public Health, 14*. http://dx.doi.org/10.1186/1471-2458-14-1144

06

(Ré)utilisation des données de la recherche ?

Trouver des données - Utiliser des annuaires

Evaluer la qualité des entrepôts, la qualité des données ?

Exemples de réutilisation

INRA SCIENCE & IMPACT

6.1- (Ré)utilisation des données

Trouver des données (annuaires et entrepôts)

Trouver des données ?

- Les sources bibliographiques classiques : bases de données, archives ouvertes, réseaux sociaux, moteurs de recherche n'intègrent actuellement quasiment pas ou très peu les datasets.

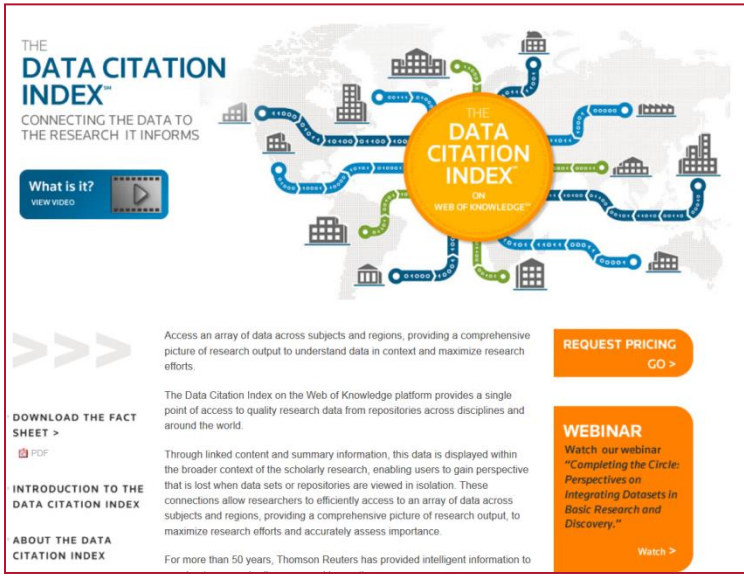
- En pratique, les données sont donc le plus souvent encore repérées par l'article de recherche qui possède le lien (ou non) vers les données déposées.
  - L'article joue le rôle de «métadonnées » permettant de retrouver les données.

- Conseil : utiliser les annuaires et répertoires d’entrepôts (DataCite, Re3data, Databib, …)

Comment trouver un entrepôt ?

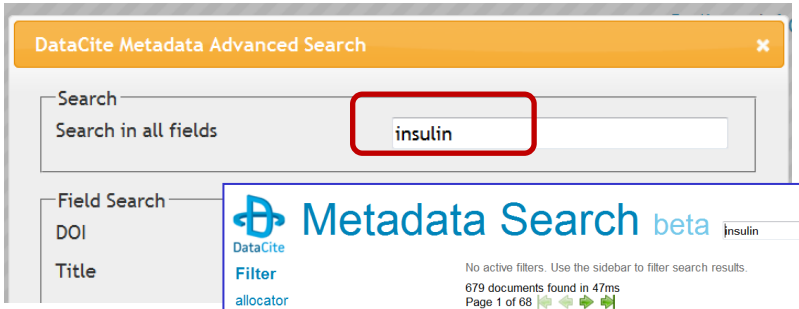
- Multiplication et hétérogénéité des entrepôts
- Conseil : utiliser des annuaires ou répertoires

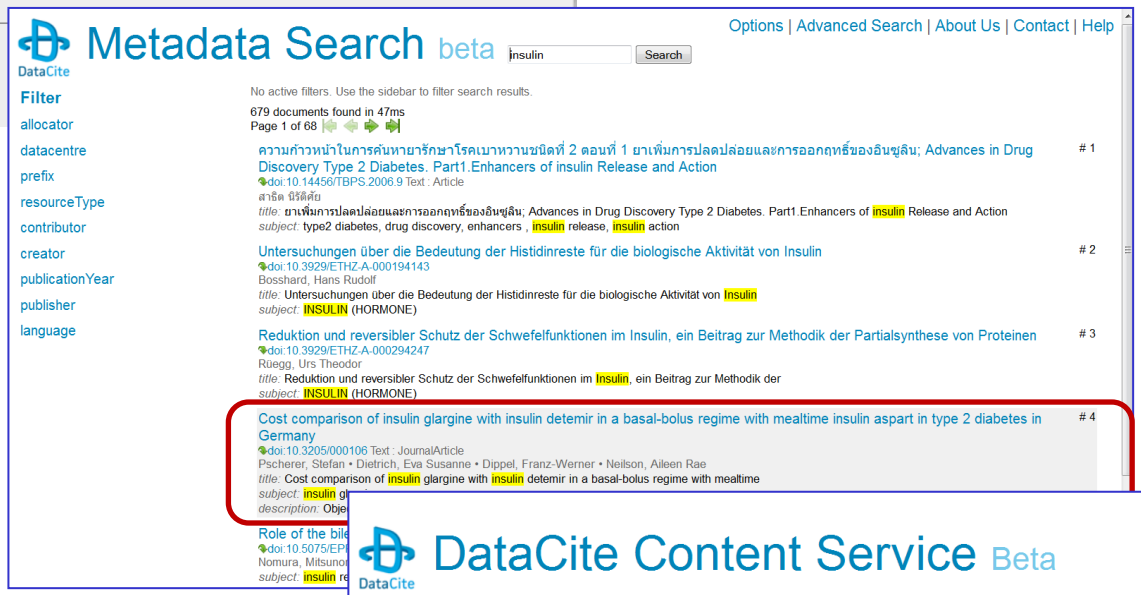
Accès gratuit



Accès payant

Exemple : Datacite





6.2- (Ré)utilisation des données

Qualité des données

Qualité des entrepôts

L'Europe se donne un cadre pour l'audit et la certification des entrepôts numériques (*European Framework for Audit and Certification of Digital Repositories*).
3 niveaux :

- **certification de base**
  - accordée au entrepôts ayant obtenu le Data Seal of Approval (DSA) Accréditation attribuée aux entrepôts numériques ayant mis en place des procédures d'assurance qualité garantissant un accès aisé et à long terme aux données stockées. Demande par procédure d'auto-évaluation
- **certification "étendue"**
- **certification "formelle"** réalisée par des experts accrédités.

http://www.trusteddigitalrepository.eu/Welcome.html

Qualité des données (1)

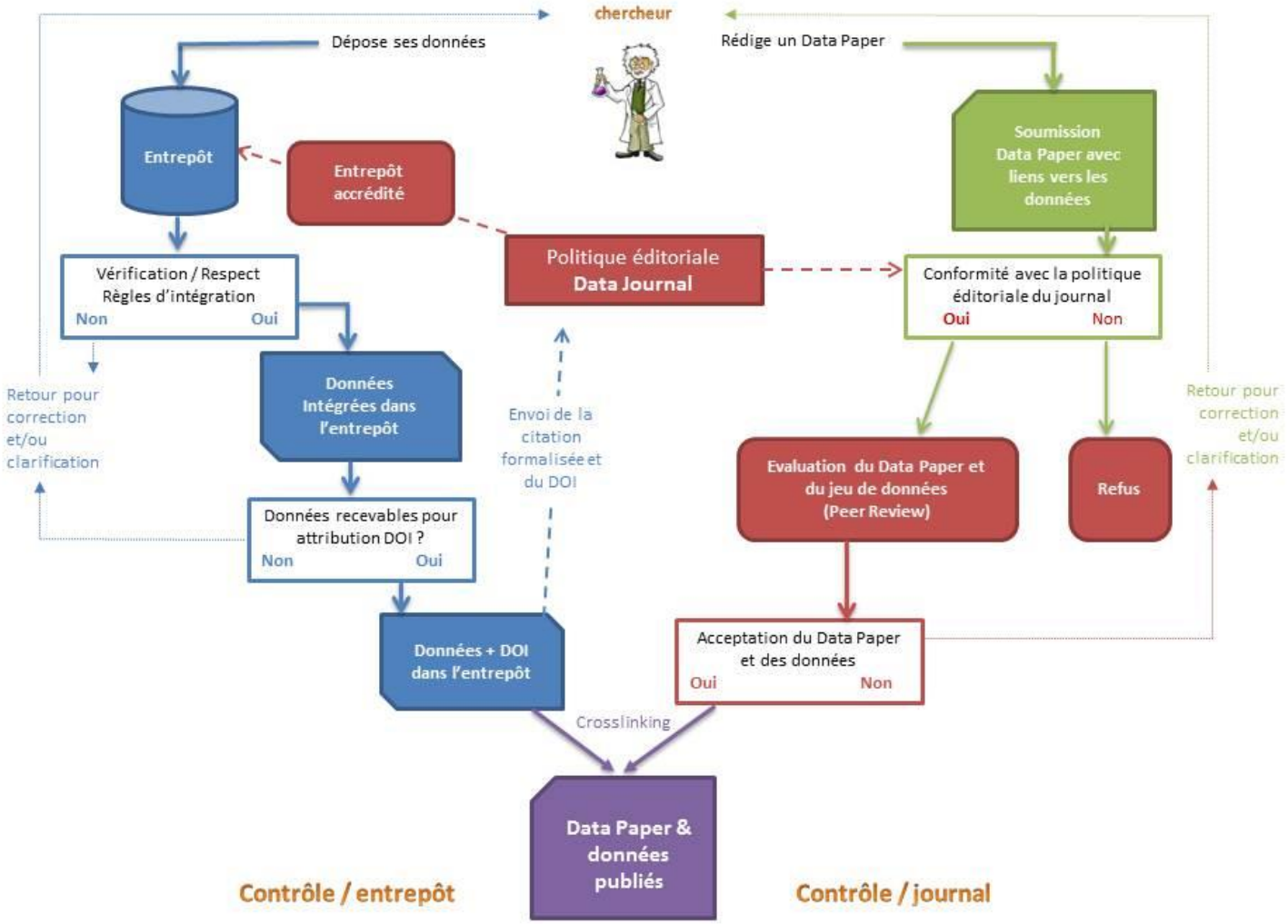
Le rapport de 2008 du Research Information Network (RIN)

- identifie trois catégories de processus impactant la qualité des données et concernant les phases :
  - de **création des données** (méthode de collecte des données, outils utilisés, étalonnage des instruments ...)
  - de **gestion des données** (description fine des données, garantie d'accès pérenne aux données ...)
  - d'**évaluation de la qualité** des jeux de données via un processus de **Peer Review**
- souligne la nécessité d'une réflexion entre les différents acteurs concernés - financeurs de la recherche, entrepôts de données et chercheurs pour définir des processus d'évaluation des données performants, acceptables par les chercheurs, et adaptés au secteur thématique considéré.

(Research Information Network (RIN), 2008)

Qualité des données (2)

- **Plan de gestion**
  - Démarche de premier niveau pour informer sur la qualité
  - De plus en plus souvent exigé par les financeurs de la recherche
- **Peer Review** des données
  - Plus difficile à mettre en place que pour les publications,
  - Des dispositifs innovants, en particulier dans les Data Journals / Data Papers



Traduit et modifié de : (Whyte & Callaghan, 2013)

Exemple : « Geoscience Data Journal » - Wiley

Extrait de : Author Guidelines

Dataset submission

"To publish a Data Paper in Geoscience Data Journal about a dataset the authors must complete the following two-tiered process:

The dataset, along with supporting metadata, must be formally archived in a Geoscience Data Journal approved repository or data centre (and preferably have been assigned a digital object identifier (DOI)). A list of approved institutions can be found below. If the one you have elected to work with does not appear please contact the Editor for consultation.

A paper describing the dataset, giving details of its collection, processing, file formats etc. should be written and submitted using the Geoscience Data Journal online submission system (https://mc.manuscriptcentral.com/geosciencedata).

Subject to satisfactory reviews of both dataset and paper,Geoscience Data Journal will publish the data description paper, along with a link to the underlying dataset (usually by means of the dataset's DOI). "

Exemple : « Scientific Data » - Nature

Extrait de :

http://www.nature.com/sdata/for-authors/editorial-and-publishing-policies#publication-criteria

Peer-review and publication criteria

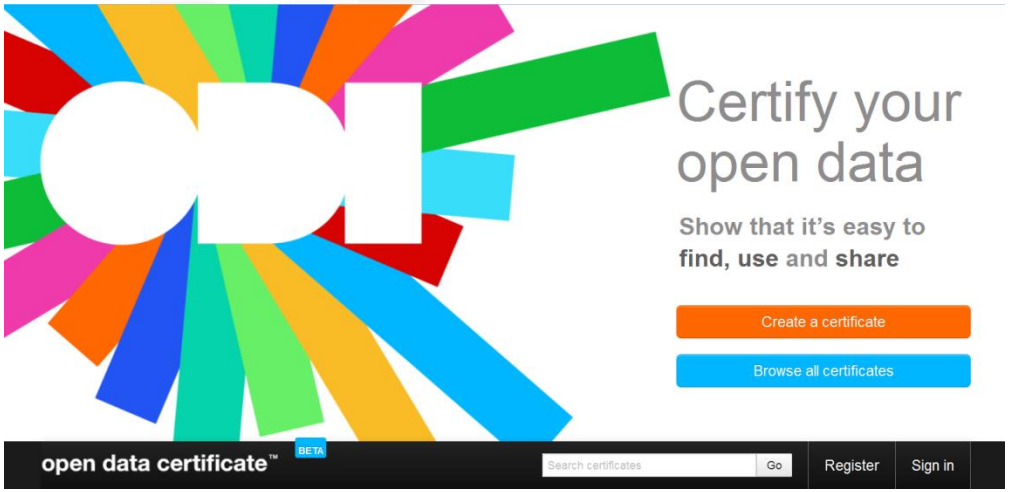
"Referees will evaluate the technical quality of the procedures used to generate the data, the reuse value of the resulting datasets, the completeness of the data description, and alignment with existing community standards. Most Data Descriptors will be reviewed by at least one scientist with expertise in the relevant experimental techniques and one data-standards expert.

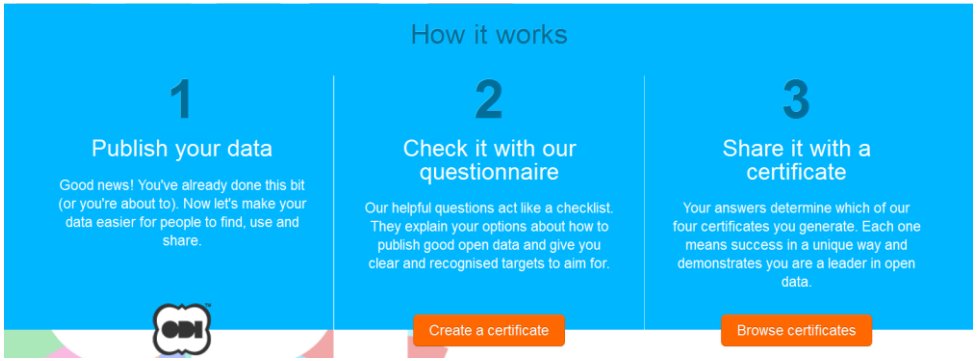
Acceptance will not be based on the perceived impact or novelty of the findings associated with the datasets, and indeed Data Descriptors will not be expected to contain in-depth analyses or new scientific conclusions. Authors will, however, be expected to support the rigour and technical quality of the experiments or procedures used to generate the data, and will be asked to provide evidence of quality-control experiments whenever necessary. Referees may ask for additional supporting experiments when needed to support the data."

Qualité des données (3)

- **Certification** des données
  - Open Data Certificates – Open Data Institute (ODI)
  - Le classement "5 étoiles" du W3C
- **Reproductibilité** des données
  - Reproducibility Initiative
  - RunMyCode

Certification ODI (Open Data Institute)





https://certificates.theodi.org/

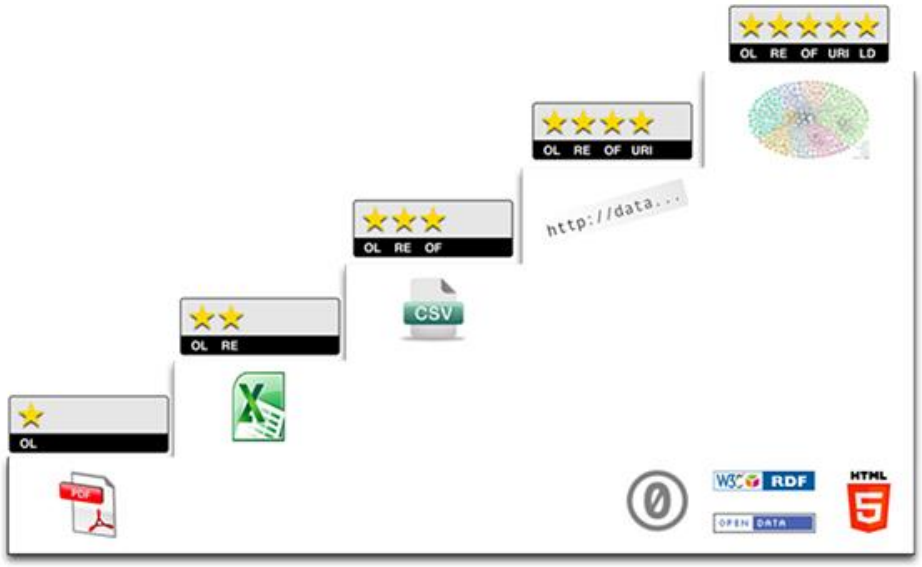
Créé en 2013, en réponse à l'attente des entreprises, gouvernements et citoyens souhaitant un dispositif de qualification des données.

Démarche « auto-certifiante »

Une marque visuelle et différents niveaux de certification (raw, pilot, standard, expert)

Classement
5 étoiles du W3C

Inspiré du dispositif de Tim Berners-Lee,
5 niveaux de qualification des données



★	les données sont visibles, sous licence, mais leur réutilisation peut être assez complexe (***On the web with an open license***) Exemple : des données enregistrées dans un tableau en format html dans une page web, le format pdf ne permettent pas de façon simple d'extraire et ré-utiliser ces données.
★★	les données sont visibles, sous licence, et leur réutilisation est relativement aisée mais pas forcément par tous (***Machine-readable data***) Exemple : les données sont enregistrées selon un format connu, aisément utilisable des programmes informatiques (le format basique csv en est un exemple,le format XML ...)
★★★	les données sont visibles, réutilisables par tous (pas de limite technique liée à l'usage d'un logiciel spécifique) (***Non-proprietary formats***) Le format des données les rend très facilement réutilisables, il ne nécessite aucun logiciel propriétaire
★★★★	les données sont visibles, simples à réutiliser et décrites de façon standardisée (*RDF Standards*) RDF permet de décrire de façon formelle les ressources Web et leurs métadonnées, de façon à permettre le traitement automatique de telles descriptions. Un document structuré en RDF est constitué d'un ensemble de triplets : (sujet, prédicat, objet). ◦ *Le sujet représente la ressource à décrire ;* ◦ *Le prédicat représente un type de propriété applicable à cette ressource ;* ◦ *L'objet représente une donnée ou une autre ressource : c'est la valeur de la propriété.*
★★★★★	les données sont visibles, simples à réutiliser, décrites de façon standardisée. Leur sens est précisé / commenté par des définitions. (***Linked RDF***) Base du web sémantique

Reproducibility Initiative

http://validation.scienceexchange.com/#/reproducibility-initiative

- Projet lancé en partenariat : Science Exchange, PloS One, Figshare, Mendeley, pour garantir la reproductibilité des travaux publiés
- Concrètement, tout chercheur peut s'adresser à Science - Exchange pour faire répéter une expérience dans des conditions similaires par un autre labo compétent et indépendant, et s'assurer ainsi de la reproductibilité des résultats.
- Un service commercial (prestation ~10% coût de l'expérience originale)
- Initiative controversée… http://www.sciencepresse.qc.ca/actualite/2012/08/17/te-paie-pour-me-dire-jai-tort

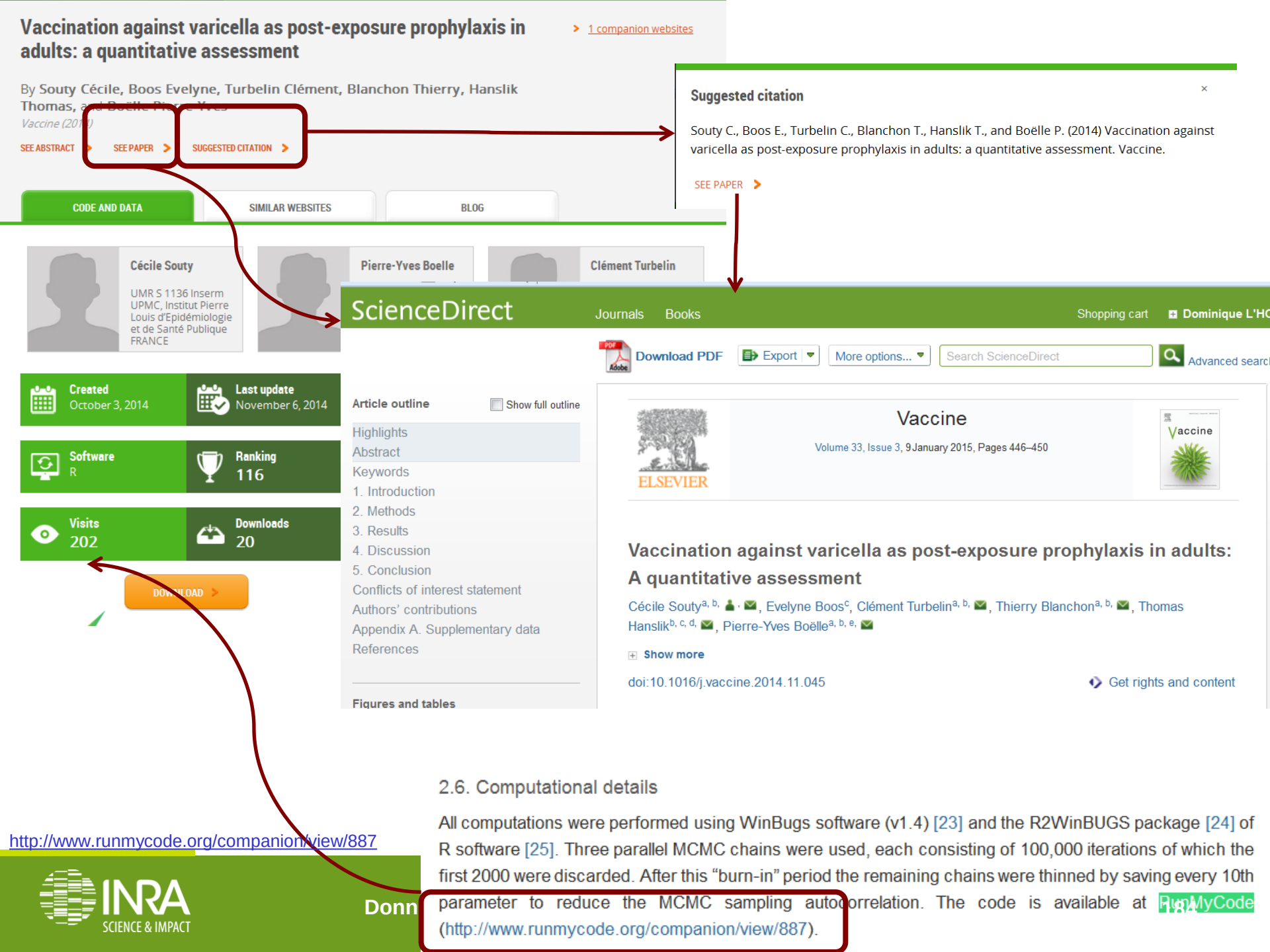
RunMyCode

http://www.runmycode.org

Projet académique international à but non lucratif, initié par des organismes de recherche et des universités. Service gratuit

Principe :

1. Un chercheur a une idée.
2. Il écrit un article basé sur cette idée.
3. Il crée un site Web associé à cet article dans lequel il fournit le code et les données permettant à des tiers de mettre en œuvre la méthodologie et de reproduire les résultats présentés.


Vaccination against varicella as post-exposure prophylaxis in adults: a quantitative assessment
1 companion websites
By Souty Cécile, Boos Evelyne, Turbelin Clément, Blanchon Thierry, Hanslik Thomas,
Vaccine (201
SEE ABSTRACT
SEE PAPER
SUGGESTED CITATION
CODE AND DATA
SIMILAR WEBSITES
BLOG
Cécile Souty
UMR S 1136 Inserm UPMC, Institut Pierre Louis d'Epidémiologie et de Santé Publique FRANCE
Pierre-Yves Boelle
Clément Turbelin
Created
October 3, 2014
Last update
November 6, 2014
Software
R
Ranking
116
Visits
202
Downloads
20
DOWNLOAD
Suggested citation
Souty C., Boos E., Turbelin C., Blanchon T., Hanslik T., and Boëlle P. (2014) Vaccination against varicella as post-exposure prophylaxis in adults: a quantitative assessment. Vaccine.
SEE PAPER
ScienceDirect
Journals
Books
Shopping cart
Dominique L'HO
Download PDF
Export
More options...
Search ScienceDirect
Advanced search
Article outline
Show full outline
Highlights
Abstract
Keywords
1. Introduction
2. Methods
3. Results
4. Discussion
5. Conclusion
Conflicts of interest statement
Authors' contributions
Appendix A. Supplementary data
References
Figures and tables
ELSEVIER
Vaccine
Volume 33, Issue 3, 9 January 2015, Pages 446–450
Vaccine
Vaccination against varicella as post-exposure prophylaxis in adults: A quantitative assessment
Cécile Souty[a, b], Evelyne Boos[c], Clément Turbelin[a, b], Thierry Blanchon[a, b], Thomas Hanslik[b, c, d], Pierre-Yves Boëlle[a, b, e]
Show more
doi:10.1016/j.vaccine.2014.11.045
Get rights and content
2.6. Computational details
All computations were performed using WinBugs software (v1.4) [23] and the R2WinBUGS package [24] of R software [25]. Three parallel MCMC chains were used, each consisting of 100,000 iterations of which the first 2000 were discarded. After this "burn-in" period the remaining chains were thinned by saving every 10th parameter to reduce the MCMC sampling autocorrelation. The code is available at RunMyCode (http://www.runmycode.org/companion/view/887).
http://www.runmycode.org/companion/view/887
INRA
SCIENCE & IMPACT
Donn

6.3- (Ré)utilisation des données

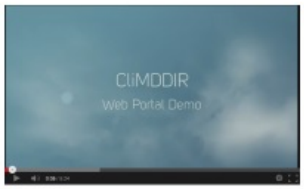
Exemples – Cas d'usage

INRA SCIENCE & IMPACT

Exemples de « Success Stories »

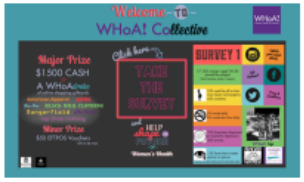
Where in the world?

An ecological story by Trevor Booth and Michael Hutchinson
Predicting where particular plants and animals are likely to be found is one of the most important problems in ecology. Reuse of publically funded research data enables species distribution modelling packages to determine the range of climatically suitable conditions for a particular species and analyse how particular species will react to changes in climatic variables. READ MORE >>



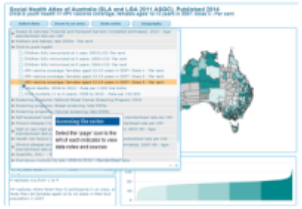
Maximising the benefits of high-resolution climate modelling - Part I

A story about building bridges by Ian Macadam
How will climate change affect the things that we care about? Will water be scarcer? Will our crops be affected? Will our ecosystems be damaged? Will our health suffer? These are all questions that we need to answer if we are to adapt to a changing climate. READ MORE >>



Sensitive data sharing: Benefiting women's health

A story about data sharing as a fundamental national research resource by Sarah Olesen, Gita Mishra, & Leigh Tooth
The Australian Longitudinal Study on Women's Health (ALSWH) has been Government-funded and gathering data on the mental, physical, and social health of over 50,000 women since 1995. If you haven't heard of ALSWH, you've probably read a report, paper, or news article based its data. You may even use a public health service or program guided by its findings. The survey data are used by over 650 collaborators....
READ MORE >>

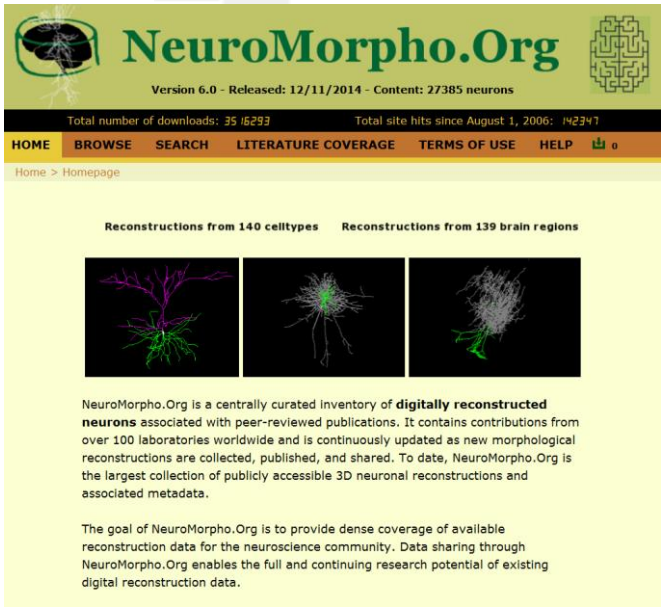


Health policy needs data sharing

A story about of the benefits of sharing Government-commissioned, publicly-funded data for policy development, by John Glover, Diana Hetzel & Sarah Olesen
Researchers, practitioners and communities must have access to public, health-relevant information through data-sharing mechanisms to continue improving the health of all Australians. Access to such data enables these groups to contribute to the evidence base that guides policy and program development, and monitors its progress. .... READ MORE >>

http://www.ands.org.au/discovery/reuse.html#stories

Exemples



NeuroMorpho.org :
collection de reconstructions numériques de neurones, résultant de la contribution de nombreux laboratoires à travers le monde. Les reconstructions morphologiques sont collectées, publiées et partagées. Le but est de fournir un accès gratuit à toutes les données de reconstruction neuronales disponibles dans la communauté des neurosciences
http://neuromorpho.org/neuroMorpho/index.jsp.

Data in use http://ukdataservice.ac.uk/use-data/data-in-use.aspx

"How our data are being used to advance research, inform policy and improve teaching"

Research Data Sharing without barriers

RESEARCH DATA ALLIANCE

Data re-use, share your experiences

Group details

Chair(s): Odile Hologne

"The RDA vision is researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society."

Most of the RDA groups, the interest groups or the working groups, are focused on technical issues, even those linked to disciplinary fields such as wheat data, marine data or biodiversity are dealing with interoperability concerns : linked data, identifiers, metadata ….. And most of the RDA members are also IT specialists (computer science, librarians, data managers …). Sometimes this work looks like a "message in a bottle" : we don't know who will read it … we don't know how the data will be re-used …

We feel the needs to imply more scientists coming from different disciplines and especially from life sciences where the data sharing culture is emerging in the context of huge societal challenges (health, climate change, nutrition …). Their involvement is necessary to reach the RDA vision

That's why we'd like to create an interest group on "data re-use" to give the opportunity to scientists to share their experiences, express their needs not in technical terms but from a user point of view.

The interest group could help to (to be discussed) :

- identify data re-use use cases and practices
- analyze those use cases : what works, what doesn't and why
- create a directory of "success stories" to communicate
- identify new technical problems to be addressed by the technical groups
- identify the benefits
- give a place to the scientist to express their needs

How :

- During the RDA plenaries with scientists testimonies
- Survey
- Interviews

https://rd-alliance.org/groups/data-re-use-share-your-experiences.html

Bibliographie – (Ré)utiliser des données

- Research Information Network (RIN). (2008). *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. Main report.* Retrieved from http://eprints.soton.ac.uk/266742/1/Published_report_-_main_-_final.pdf
- White, E.P.; Baldridge, E.; Brym, Z.T.; Locey, K.J.; McGlinn, D.J.; Supp, S.R., 2013. Nine simple ways to make it easier to (re)use your data. *PeerJ Preprints.* *10.7287/peerj.preprints.7v2*
- Whyte, A., & Callaghan, S. (2013). *Perspectives on the Role of Trustworthy Repository Standards in Data Journal Publication*. Paper presented at the IASSIST, Cologne - Allemagne. http://fr.slideshare.net/angusawhyte/iassist-preparde-whyte

Sites des annuaires

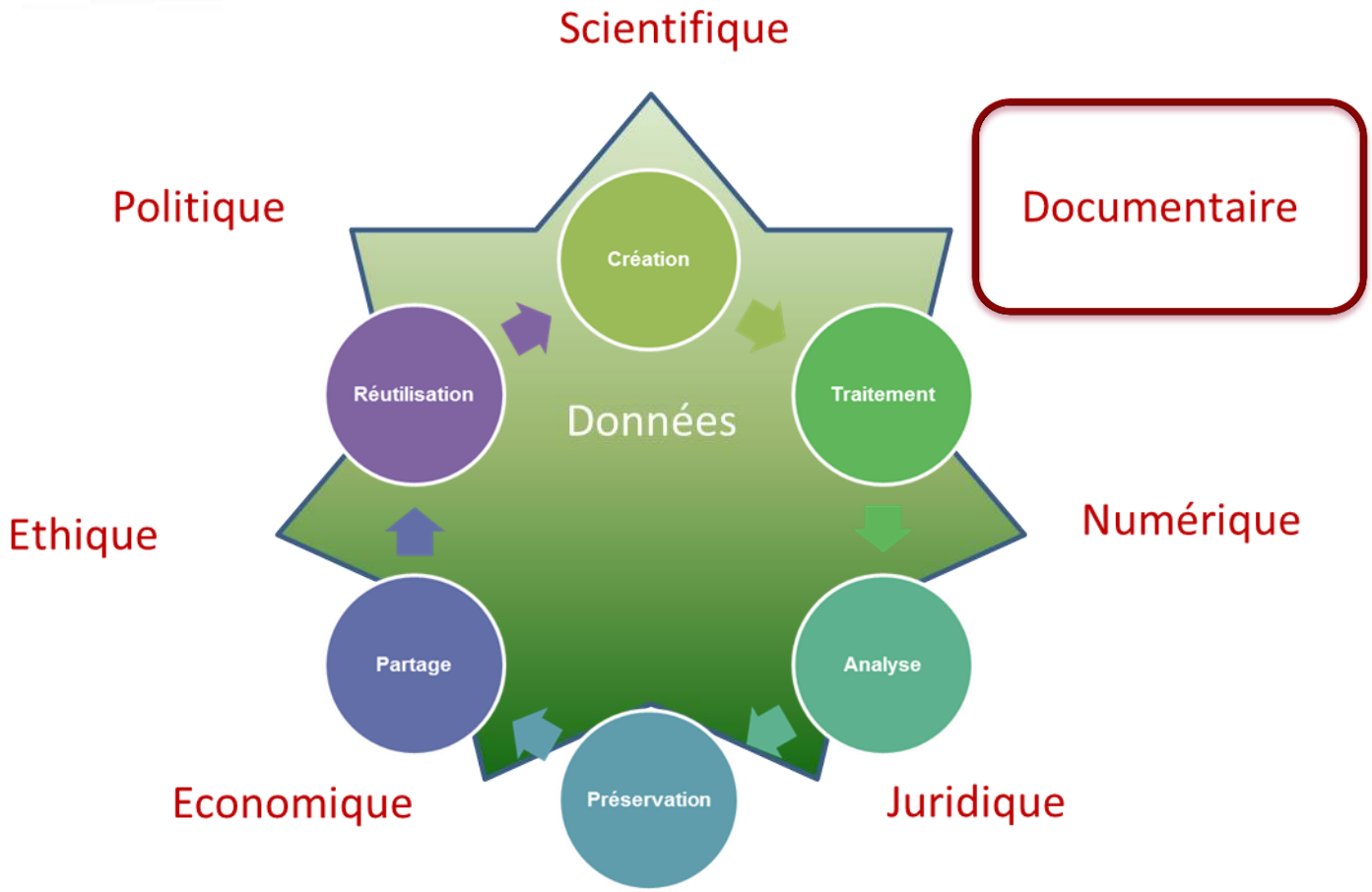
- http://www.re3data.org/
- http://databib.org/
- https://www.datacite.org/
- Data Citation Index – Thomson : http://thomsonreuters.com/data-citation-index/
  - Torres-Salinas, D., Martin-Martin, A., & Fuente-Gutierrez, E. (2013). An introduction to the coverage of the Data Citation Index (Thomson-Reuters): disciplines, document types and repositories. *EC3 Working Papers, 1-8* *http://arxiv.org/ftp/arxiv/papers/1306/1306.6584.pdf*
  - Torres-Salinas, D., Jimenez-Contreras, E., & Robinson-Garcia, N. (2014). *How many citations are there in the Data Citation Index? Paper presented at the STI Conference Leiden.* *http://arxiv.org/abs/1409.0753*

_07

Données de la recherche et IST ?

INRA
SCIENCE & IMPACT

La gestion des données mobilise différents acteurs et compétences

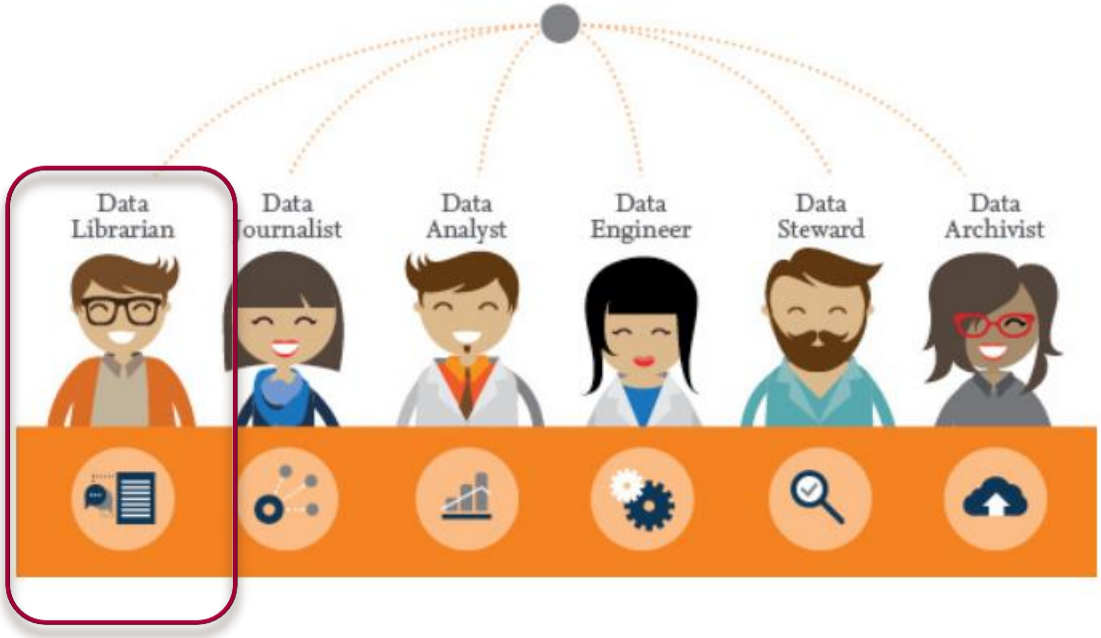


7.1- Données de la recherche et IST

Rôles possibles

INRA SCIENCE & IMPACT

DATA SCIENCE ROLES



Liz Lyon, University of Pittsburgh
http://libraryconnect.elsevier.com/articles/2014-10/learning-about-research-data-lab-pitt-ischool#sthash.y0l3370N.dpuf
Data Job description
http://www.ands.org.au/guides/dmframework/dmskills-information.html

An exciting opportunity ?

"Researchers need help to manage their data. This is a really exciting opportunity for libraries…."

(Lyon, 2012)

Getting Onboard the Data Train

How Librarians Fit In

Diane Clark,
Staff Development &Training Librarian | Agriculture, Food & Nutritional Science Librarian, University of Alberta Libraries

(Clark, 2014)

Linking data and publications — a new task for scientific libraries

(Vlaeminck & Sven, 2013)

Vision ANDS

- Data is the next great challenge for scholarly communication
- And so, it should be the next great challenge for libraries

Challenges…

- S’engager dans de nouvelles activités
- Etre crédible dans un nouveau domaine d'expertise
- Acquérir de nouvelles compétences, « *getting techie* »
- «Essuyer les plâtres» sur de nouveaux objets (Data, Datasets …)

7.2- Données de la recherche et IST

Compétences mobilisables
Compétences à développer

INRA
SCIENCE & IMPACT

Les atouts de l'IST …

- Rôles de leadership
  - Open Access et recherche des données existantes
  - gestion des entrepôts de publications
- Des compétences en :
  - gestion de l'information,
  - création et utilisation des métadonnées
  - Information et formation auprès des publics scientifiques
- Rôle de conseil possible sur les aspects juridiques, PI …
- Des relations privilégiées avec les chercheurs et d'autres acteurs impliqués (publishers …)
- Une communauté engagée « Data » active (pays anglosaxons surtout)

➡ Un ensemble de compétences pertinentes par rapport au cycle de gestion des données

Apports possibles de l'IST

- Contribuer à l'émergence d'une politique institutionnelle
- Recenser et valoriser les sources de données
- Aider à définir les stratégies de dépôt
- Intervenir dans la gestion des DOI et des métadonnées associées
- Sensibiliser les scientifiques à la valorisation des données
- Etre le relais des bonnes pratiques…

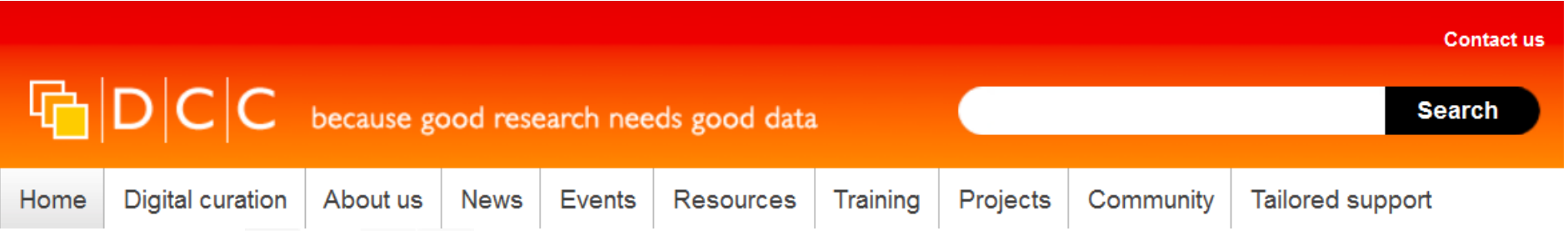
Grâce à quelles compétences ?

- Traditionnelles
  - Compréhension des voies de publication (Data Journals …)
  - Connaissance des entrepôts de données
  - Mécanismes de citation et d'évaluation (métries)
  - Connaissances juridiques (CC, PI, BdD)
  - Maintenance de référentiels
- Technologiques
  - Identifiants (DOI …)
  - Métadonnées
  - Interopérabilité (OAI-PMH et autres)
  - Linked Open Data, ontologies et standards associés

7.3- Données de la recherche et IST

Engagement et formation des communautés

INRA
SCIENCE & IMPACT



How are libraries engaging in RDM?

The library is leading on most DCC institutional engagements.

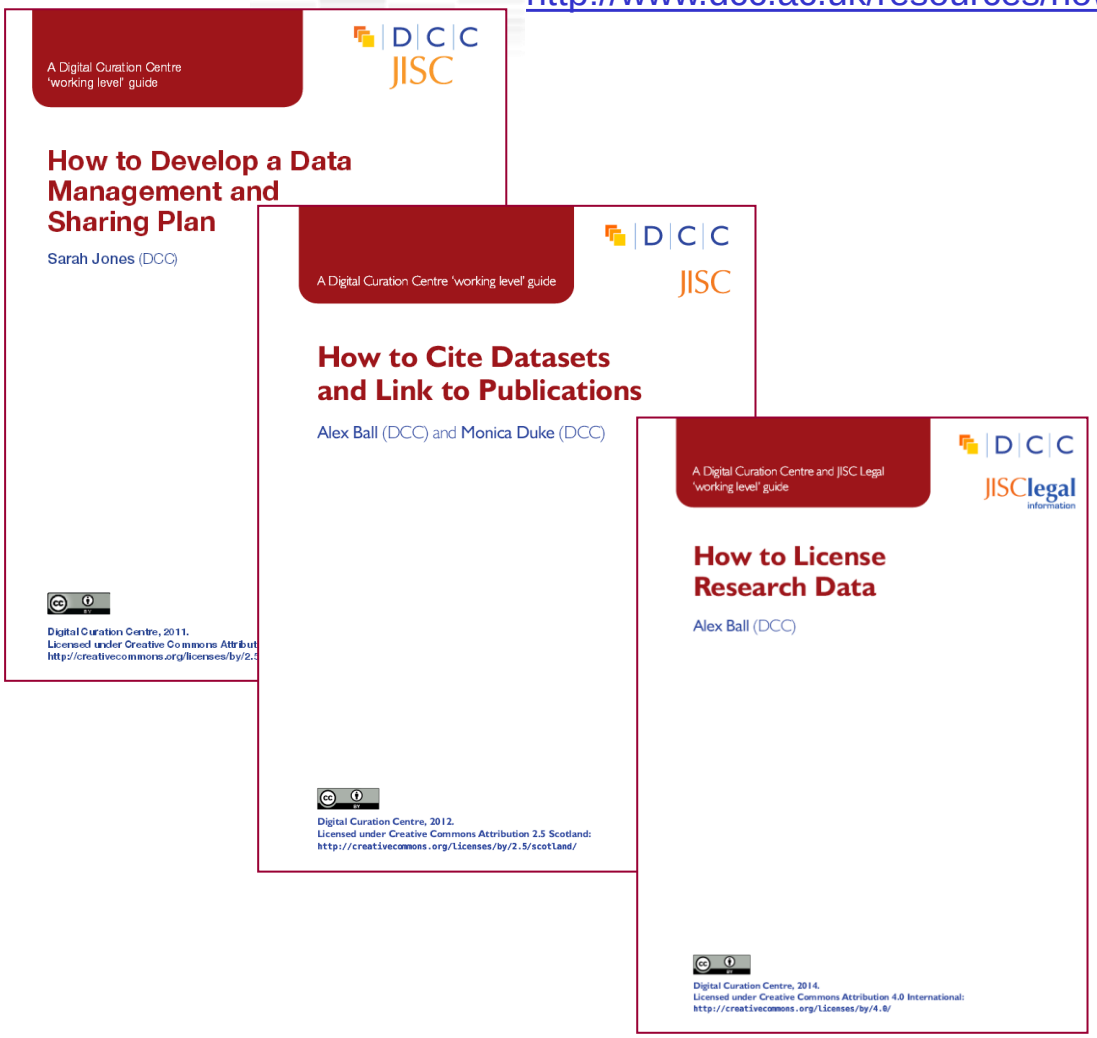
www.dcc.ac.uk/community/institutional-engagements

They are involved in:

- defining the institutional strategy
- developing RDM policy
- delivering training courses
- helping researchers to write DMPs
- advising on data sharing and citation
- setting up data repositories…

DCC because good research needs good data

http://www.dcc.ac.uk/resources/how-guides



Training

- Curation webinars
- Digital Curation 101
  - DMP workshop at UCT
  - LIASA RDM workshop
  - UCT Research Data Management Policy and Strategy Workshop
- Materials for Trainers
  - Career profiles and related data management skills
  - DC 101 training materials
  - Disciplinary RDM training
  - Skills frameworks
- Data management courses and training
- Tools of the Trade training
- RDM for librarians

Information specialists and data librarians

Who are information specialists and data librarians?

In the context of eResearch, information specialists such as librarians, archivists, curators and records managers may play key roles. Data librarians are professional library staff engaged in managing research data, using research data as a resource, or supporting researchers in these activities.

http://www.ands.org.au/guides/dmframework/dmskills-information.html

What do information specialists do?

The role may include supporting researchers or institutional initiatives in the following areas:

- Data management
  - data management planning
  - issues such as copyright, intellectual property, licensing of data, embargoes, ethics and re-use, privacy
  - storing and managing data during the research project (curation)
  - depositing data in archives at the end of the project, determining retention and disposal
  - open access and publishing of data
  - research organisation policies affecting data
- Metadata management
  - creating and maintaining metadata
  - developing and applying metadata standards
- Using data (data as a resource)
  - finding or obtaining data for re-use
  - citing data
  - data analysis tools and support services
  - data literacy (an extension of information literacy to include the ability to "access, assess, manipulate, summarize and present data". Read more)

The role may also include:

- developing, delivering or arranging
  - resources such as data management checklists
  - training sessions, on topics such as data management planning, data literacy, use of statistical and analytical tools
  - awareness sessions or materials
- referral to sources of information and advice, either within or external to the organisation

Reading list for information specialists beginning data management projects [PDF 76.9KB]

Canadian Community of Practice for Research Data Management in Libraries

http://data-carl-abrc.ca/

Gestion des données - Project ARC

Project ARC réunit les initiatives menées par des bibliothèques de recherche dans le domaine de la gestion des données afin de coordonner les activités et la formation pour les connaissances et les habilitées requises pour l'archivage des données de la recherche.

Exemple de cours développé autour des données
http://datalib.library.ualberta.ca/rdmi/

Getting Onboard the Data Train

How Librarians Fit In

Diane Clark,
Staff Development &Training Librarian | Agriculture, Food & Nutritional Science Librarian, University of Alberta Libraries

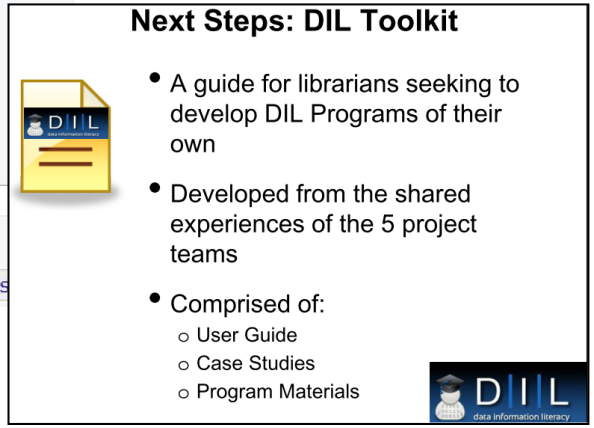
(Clark, 2014)

Tools, Services, Expertise

- Tools
  - DMPOnline | Dataverse | ERA
- Services
  - teach | consult & recommend | refer | store & preserve
- Expertise
  - metadata | remedial data help | data cleansing | preservation |

Se former pour ensuite pouvoir sensibiliser – accompagner – former…

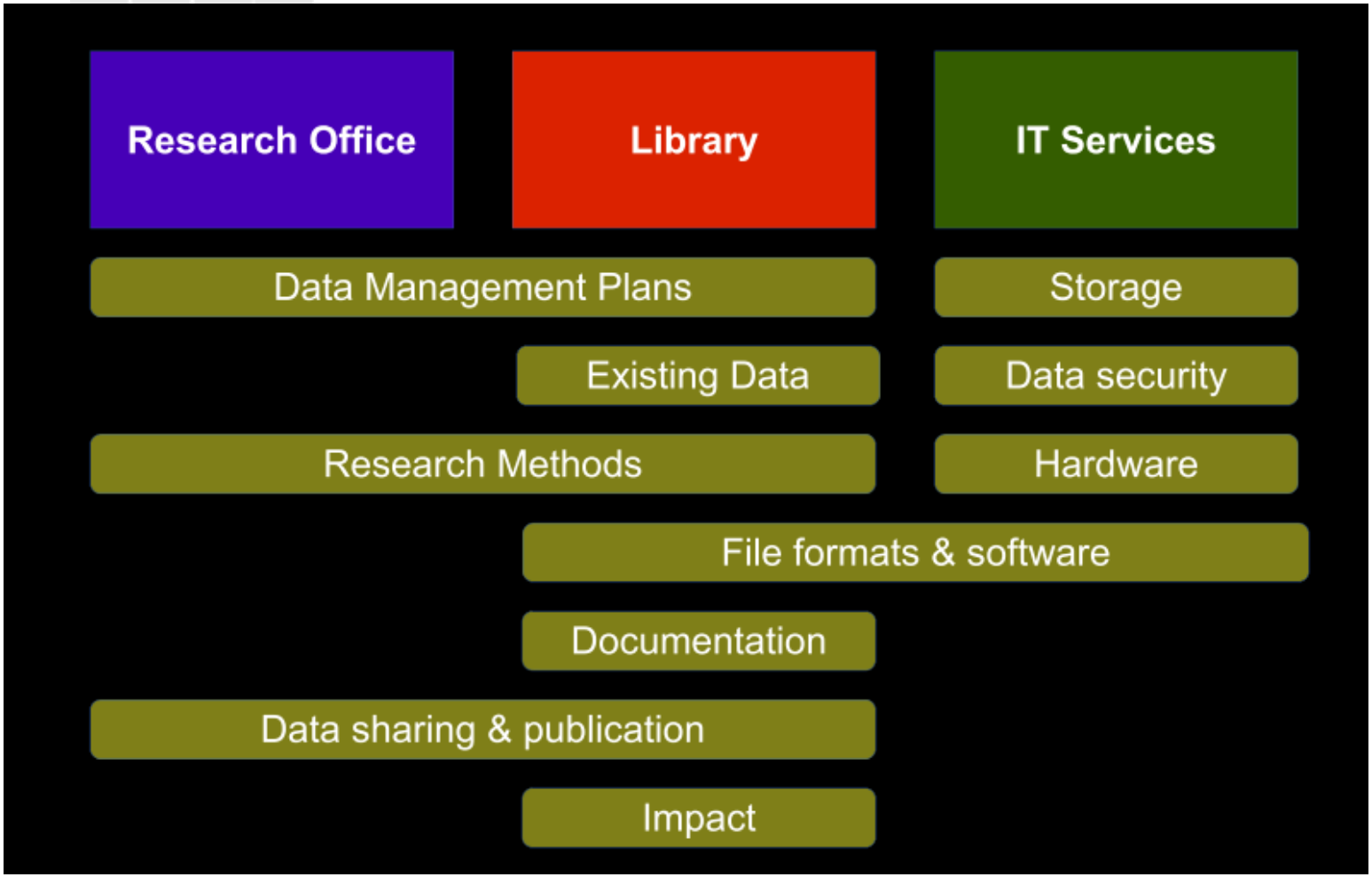
http://esciencelibrary.umassmed.edu/lib_roles



(Carlson, 2014)

http://datalib.edina.ac.uk/mantra/libtraining.html

(Jones, Guy, & Pickton, 2013)


Research Office
Library
IT Services
Data Management Plans
Storage
Existing Data
Data security
Research Methods
Hardware
File formats & software
Documentation
Data sharing & publication
Impact

(Ball, 2013)

Et en France ?

À l'Inra… actuellement

- Chantier « Data Partage »
  - Co-pilotage du chantier / O. Hologne – Directrice DV-IST
  - Implication dans des groupes de travail
    - Données et publications
    - Plan de gestion des données
    - Entrepôt/annuaire
    - Familles de données
    - Communication

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

- Implication dans des initiatives internationales
  - Research Data Alliance (RDA), notamment groupe de travail sur l'interopérabilité du blé
  - Global Open Data for Agriculture and Nutrition (GODAN)
  - Coherence in Information for Agriculture Research (CIARD)
  - Science Europe
- Publication de « linked data »
- Travail sur les ontologies, standards de métadonnées…

A l'INRA. Une vision possible du futur ?

Aide aux chercheurs pour la gestion et la publication des données

Lancement

- Savoir où trouver des données pertinentes pour le chercheur
- Savoir quel entrepôt conseiller
- Aider à la rédaction d'un plan de gestion

Développement

- Connaître la politique institutionnelle sur la gestion et de partage des données
- Comprendre les exigences des entrepôts visés
- Aider le chercheur à se conformer aux exigences des entrepôts visés
- Aider le chercheur à se conformer aux exigences des financeurs

Dissémination

- Informer sur, et aider au choix d'une stratégie de diffusion (data paper, dépôt dans un entrepôt, …)
- Informer sur la politique des données des éditeurs
- Contribuer à une meilleure découverte des données (métadonnées, diffusion dans les média sociaux, etc.)

Un peu de vécu !

Chercheur - Orléans

Je voudrais publier des données statistiques sur les propriétés des sols dans le cadre des programmes du GIS Sol. J'ai repéré le journal Scientific data
http://www.nature.com/scientificdata/?WT.mc_id=EMI_SCIDATA_1403_PLEnvSci

Ils demandent cependant de déposer les données sur un de leurs entrepôts
http://www.nature.com/scientificdata/for-authors/data-deposition-policies/#recommended-repositories

Je ne sais pas lequel choisir. J'ai repéré Pangaea.

Pourriez vous me conseiller sur le journal et l'entrepôt ?
Merci d'avance

« Scientific Data » est un tout nouveau journal édité par le groupe Nature. Il correspond à la notion de « Data Journals » que nous avions décrite dans le wiki (page Données publiées sous forme de data paper)
Les journaux de ce type suggèrent fréquemment des entrepôts spécifiques aux auteurs pour le dépôt de leurs données. A priori, nous ne voyons pas de contre-indication à suivre ces consignes si les entrepôts recommandés remplissent les conditions suffisantes de qualité et de fiabilité pour le traitement, la citation et la conservation pérenne des données.
A priori, Pangeae fait partie des entrepôts connus. Il utilise des standards reconnus et sérieux pour citer les données (le DOI), et pour protéger les données (licences CC) …
Nous en avons donné une courte description dans le wiki et il est pertinent par rapport à votre thématique. Une alternative serait de déposer dans un entrepôt généraliste comme Dryad qui est également dans la liste des entrepôts recommandés par « Scientific Data ».
Vous pouvez éventuellement comparer les caractéristiques des deux systèmes pour orienter votre choix définitif en utilisant par exemple des annuaires comparatifs d'entrepôts (voir la page wiki "Annuaires d'entrepôts"). Un des éléments importants à prendre en compte étant certainement l'orientation "communauté scientifique" de Pangeae par rapport à la couverture plus large et multidisciplinaire de Dryad.

Chercheur – Versailles

Nous nous trouvons maintenant dans la nécessité de fournir des données brutes pour que certaines revues acceptent la publication d'articles.
Les collègues co-auteurs sont d'accord pour rendre publiques les données brutes mais souhaitent avoir une traçabilité et identifier qui chargent les données et ceci sans limite de temps (donc pas un système d'embargo de quelques mois ou années comme le proposent certains éditeurs).
Connaitriez-vous un système de dépôt de données qui permette ainsi de tracer les personnes déchargeant les données ?

En espérant que vous pourrez m'orienter…

Nous n'avons malheureusement pas trouvé d'entrepôt qui pourrait vous alerter à chaque fois que quelqu'un téléchargera votre jeu de données. Nous pensons que les services de traçage peuvent présenter un risque au niveau juridique dans certains pays, et que cela pourrait expliquer leur absence dans les entrepôts de données.
Si toutefois vous voulez absolument savoir qui télécharge vos données, nous vous proposons une démarche qui n'est pas idéale et qui ne peut qu'être provisoire : déposer dans Prodinra un rapport concernant vos données et joindre le jeu de données, puis indiquer un contact pour ceux qui souhaitent obtenir le jeu de données. Il n'est pas certain que la revue que vous visez accepte ce compromis. Elle pourrait considérer que le jeu de données ne sera pas vraiment en accès libre après la période d'embargo convenue, puisqu'il faut d'abord vous écrire, et que rien ne garantit que vous répondrez favorablement à toutes les demandes.
Peut être que tracer les réutilisateurs de vos données n'est pas une finalité en soi pour vous et que votre préoccupation réelle est autre?

Si votre préoccupation est de vous assurer que ceux qui téléchargent vos données les interprètent et les ré-utilisent correctement, nous pensons qu'une description précise des métadonnées et une bonne documentation de vos données lors du dépôt dans Dryad permet de minimiser le risque de mauvaise interprétation/utilisation.
Peut être que votre préoccupation est d'être cités par ceux qui utilisent vos données. Dryad incite ceux qui téléchargent les données à en citer les auteurs et propose des citations prêtes à utiliser (voir copie écran jointe).
De plus Dryad identifie les jeux de données avec un DOI, et plusieurs services sont en train de se développer autour du DOI des données pour permettre de traquer les citations de données et proposer des indicateurs.
Parallèlement au dépôt dans Dryad, vous pouvez aussi publier un data paper dans un data journal tel que Scientific data de Nature (http://www.nature.com/scientificdata/) ou un autre pour augmenter vos chances d'être cités.
Une liste de data journals est accessible à l'adresse suivante :
http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList

A l'INIST - CNRS

- Projet « Ingénierie des connaissances » de l'Inist pour la communauté ESR.
  14 activités identifiées dont :
  « Gestion et valorisation des données de la recherche »
- Projet de formations sur la diffusion des résultats de recherche dans le cadre du programme H2020
  - Vers la communauté scientifique
    - Une première réalisation :
      Institut de l'Information Scientifique et Technique (INIST) (2014). [Une introduction à la gestion et au partage des données de la recherche].
      http://www.inist.fr/donnees/co/Donnees_recherche_web.html
  - Vers la communauté IST ?

Bibliographie – « Données et IST »

- Ball, A. (2013). [Les métiers liés aux données de la recherche : Data Librarian].
- Carlson, J. (2014). [The data information literacy project]. http://www.duraspace.org/esi-logistics
- Clark, D. (2014). [Getting onboard the data training: How librarians fit in]. http://fr.slideshare.net/di_clark/getting-onboard-the-data-train
- Hügi, J., & Prongué, N. (2014). Le virage Linked Open Data en bibliothèque : étude des pratiques, mise en œuvre, compétences des professionnels. *Ressi, (Décembre).* *http://www.ressi.ch/num15/article_100http://www.ressi.ch/num15/article_100*
- Lyon, L. (2012). *The informatics transform : re-engineering libraries for the Data Decade*. Paper presented at the VALA2012 PLENARY 4. http://www.vala.org.au/vala2012-proceedings/vala2012-plenary-4-lyon http://webcast.gigtv.com.au/Mediasite/Play/90077d28660e4c0487f8cf293d639e071d
- Lyon, L. (2012). The Informatics Transform: Re-Engineering Libraries for the Data Decade. *International Journal of Digital Curation, 7*(1), 126-138. http://dx.doi.org/10.2218/ijdc.v7i1.220
- MacMillan, D. (2014). Data Sharing and Discovery: What Librarians Need to Know. *The Journal of Academic Librarianship, 40(5), 541-549. 10.1016/j.acalib.2014.06.011*
- Martin, V. (2014). *Demystifying eResearch. A primer for librarians*. Santa Barbara (USA): ABC-CLIO, 189 p. (ISBN = 978-1-61069-520-6, e-ISBN = 978-1-61069-521-3)
- Tenopir, C.; Sandusky, R.J.; Allard, S.; Birch, B., 2014. Research data management services in academic research libraries and perceptions of librarians. *Library & Information Science Research, 36 (2): 84-90. 10.1016/j.lisr.2013.11.003*

Sitographie

Sites références et rapports par pays

- Australie
  - Australian National Data Service
    http://www.ands.org.au/index.html
- Canada
  - Research Data Canada
    http://rds-sdr.cisti-icist.nrc-cnrc.gc.ca/eng/
  - Canadian Association of Research Libraries
    http://data-carl-abrc.ca/
- Pays Bas
  - Meijer, S. L. (2014). *Regulation of open access to research data. A study about open access to research data and the role of Dutch government.*
  - Rombouts, J. (2014). *Research Data (in the) Netherlands. Landscape, collaboration and "Data People". Carrefours de l'IST, 25/11/2014, Nancy - France*
- UK
  - Digital Curation Center http://www.dcc.ac.uk/
  - Data Archive http://www.data-archive.ac.uk/
  - Knowledge Exchange – Research Data
    http://www.knowledge-exchange.info/Default.aspx?ID=284
  - Nombreux sites universitaires très riches (cf. diapos 52, 80-81)

Sites références par pays (2)

France

Sites

- ✓ Site du gouvernement Data.gouv
  https://www.data.gouv.fr/fr/
- ✓ Plateforme d'information et de veille sur les données de la recherche créée par le ministère de l'Enseignement Supérieur et de la Recherche
  http://www.donneesdelarecherche.fr/
- ✓ Wiki IST INRA (intranet)
  https://wiki.inra.fr/wiki/donneesrechercheist

Documents

- ✓ Gaillard, R. (2014). *De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ?". (Conservateur de Bibliothèque), Université de Lyon, ENSSIB, Lyon. Retrieved from http://www.enssib.fr/bibliotheque-numerique/documents/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche.pdf*

Merci de votre écoute

Esther Dzale Yeumo edzale@versailles.inra.fr

&

Dominique L'Hostis dolhostis@nantes.inra.fr

INRA SCIENCE & IMPACT