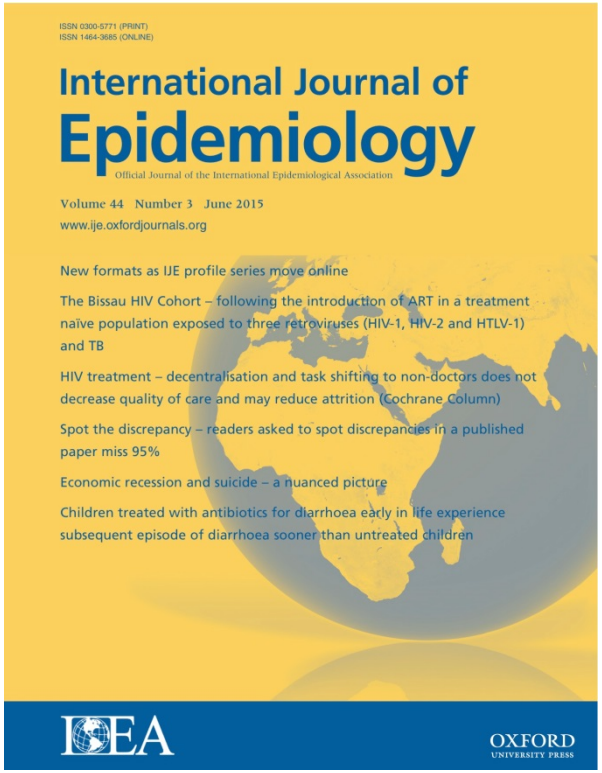
L'ouverture des données de la recherche en 2015 : définitions, enjeux, dynamiques

Kenya Example: How Re-Analyzing Scientific Research Data Can Change The Findings



Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial

Calum Davey[1], Alexander M Aiken[1,*], Richard J Hayes[2] and James R Hargreaves[1]

+ Author Affiliations

« *Re-analysis is a powerful tool in the review of important studies, and should be supported with data made available by researchers and with adequate funding. The publication of the replication results by the* International Journal of Epidemiology *is a watershed. To our knowledge it is very unusual for an international journal to publish a re-analysis of an already published trial.* »

UPMC SORBONNE UNIVERSITÉS

Des impulsions politiques récentes, à l'échelle nationale et européenne

- 17 juillet 2015 : ***Concordat on Open Research Data*** **en Grande-Bretagne**, à l'initiative des agences de financement et des établissements de recherche
- 18 juin 2015 : présentation du rapport *Ambition numérique* de la CNNum et de la **stratégie numérique du gouvernement français**
- 16 décembre 2013 : lancement du **projet-pilote de la Commission européenne sur les données de la recherche** produites dans le cadre d'Horizon 2020

UPMC SORBONNE UNIVERSITÉS

L'exemple du *Concordat on Open Research Data* - juillet 2015

« In order to make data open, it must be not only accessible, but also discoverable and useable. This can be achieved in a number of ways, such as **through peer review, adherence to community-specific data formats and standards, deposition in specific repositories and through appropriate descriptions, or dedicated data articles** in journal publications. However, these methodologies may vary according to subject and disciplinary fields, types of data, and the circumstances of individual projects. Hence **the exact choice of methodology should not be mandated.** »

« Production of open research data should be formally acknowledged as a legitimate output of the research process and should be recognised as such by employers, research funders and others in contributing to an individual's professional profile in relation to promotion, research assessment and research funding decisions. Such formal recognition should be accompanied by the development and use of metrics that allow the collection and tracking of data use and impact. **In general, data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications**. »

« To prevent (…) negative outcomes, **researchers who generate original data must have reasonable right of first use for a short and well-defined period**, which may vary by subject and disciplinary area. This should include an understanding that researchers first need to verify newly-obtained data (generally by repeating measurements) before they themselves can use the data for publications or other outcomes. »

UPMC SORBONNE UNIVERSITÉS

Le périmètre du projet-pilote et son déploiement

H2020: Pilot on Open Research Data

Types of data concerned:

- Data needed to validate the results presented in scientific publications ("underlying data")
- Other data as specified in DMP (=up to projects)

Beneficiaries participating in the Pilot will:

- Deposit this data in a research data repository of their choice
- Take measures to make it possible to access, mine, exploit, reproduce and disseminate free of charge
- Provide information about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (where possible, provide the tools and instruments themselves)

Source : *Open research data: present and planned EC policy* - Jean-Claude Burgelman (mai 2015)

UPMC SORBONNE UNIVERSITÉS

Le périmètre du projet-pilote et son déploiement

European Commission

H2020: Pilot on ORD: take-up in first calls

Basis: 3054 Horizon 2020 proposals

Calls in core-areas: **opt out 24.2%** (442 of 1824 proposals) – range from 9,1-29,1%

Other areas: voluntary **opt in 27.2%** (334 of 1230 proposals) – range from 9 to 50%

→ Preliminary but encouraging results

UPMC SORBONNE UNIVERSITÉS

La donnée et son cycle de vie : de la production à la dissémination

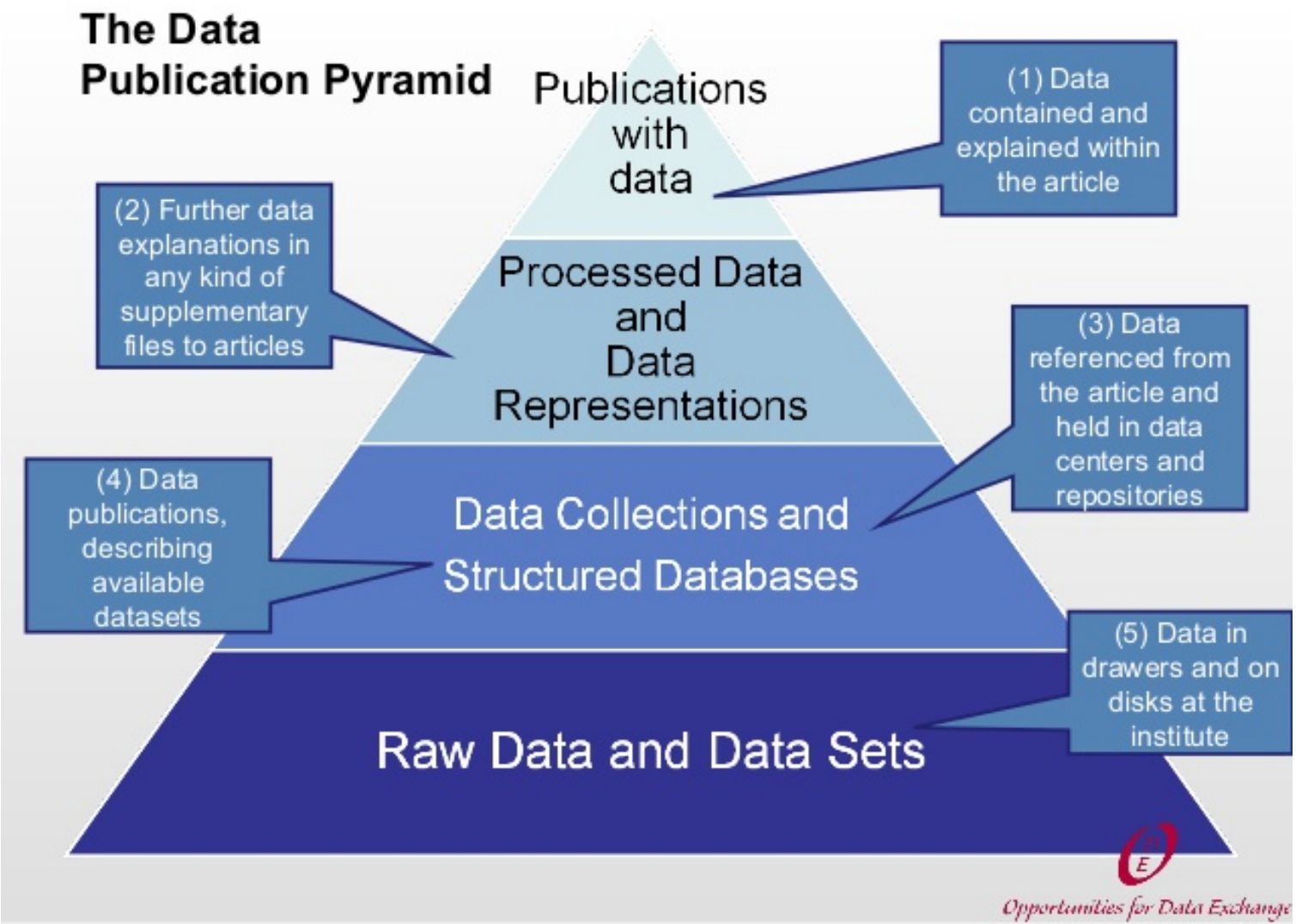
« La donnée de recherche est définie comme **l'enregistrement factuel** couramment considéré dans la communauté scientifique comme nécessaire à la validation des résultats de la recherche » (OCDE, 2007)

« Les données de la recherche sont l'ensemble des informations et matériaux produits et reçus par des équipes de recherche et des chercheurs. Elles sont collectées et documentées à des fins de recherche scientifique. À ce titre, elles constituent **une partie des archives de la recherche** » (AAF, Section AURORE, 2014)

« Les données de la recherche sont **les informations quantitatives ou les relevés qualitatifs collectés par les chercheurs au cours de leurs expérimentations, observations, entretiens ou suivant toute autre méthode**. Les données peuvent être brutes ou primaires (directement issues d'une mesure ou d'un enregistrement) ou dérivées de données primaires à des fins d'analyse ou d'interprétation (une fois nettoyées ou extraites d'un jeu de données plus large). L'objectif de l'ouverture des données de la recherche est de fournir l'information nécessaire à la validation des constatations, des conclusions ou des résultats d'un projet de recherche » (*Concordat*, 2015)

UPMC SORBONNE UNIVERSITÉS

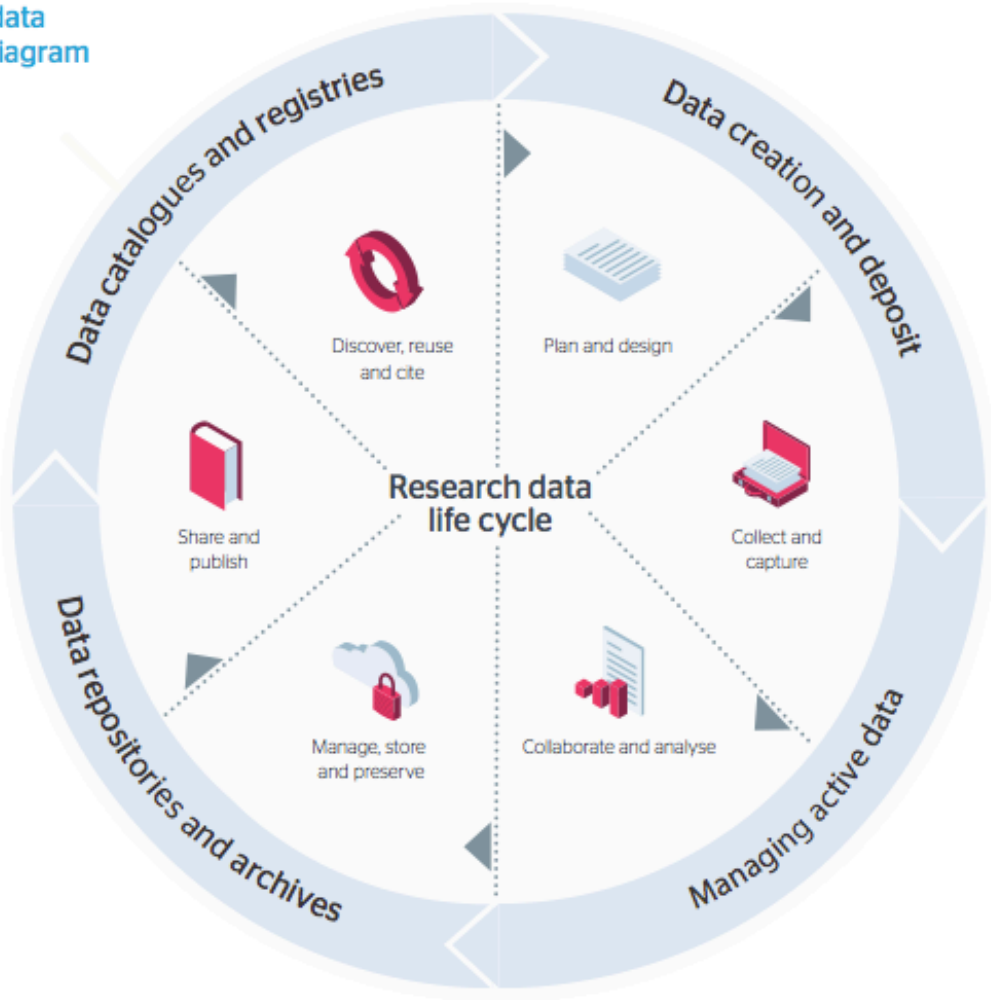
La donnée et son cycle de vie : de la production à la dissémination (I)



UPMC SORBONNE UNIVERSITÉS

La donnée et son cycle de vie : de la production à la dissémination (II)

Research data
life cycle diagram



UPMC
SORBONNE UNIVERSITÉS

Le partage des données : un jeu d'acteurs complexe (I)

- Initiative des communautés de chercheurs, culture disciplinaire ayant conduit à la structuration progressive de standards communs, de métadonnées normées, d'infrastructures disciplinaires partagées (*ex.* : le GIS réseau Quetelet, DDI norme de description de données en SHS)

- Mandats institutionnels des agences de financement (*ex.* : RCUK, AHRQ)

	Policy Coverage		Policy Stipulations					Support Provided			
Research Funders	Published outputs	Data	Time limits	Data plan	Access/ sharing	Long-term curation	Monitoring	Guidance	Repository	Data centre	Costs
AHRC	●	●	●	●	●	◐	○	●	○	◐	◐
BBSRC	●	●	●	●	●	●	●	●	●	◐	●
CRUK	●	●	●	●	●	●	●	◐	●	○	○
EPSRC	●	●	●	◐	●	●	●	◐	○	○	●
ESRC	●	●	●	●	●	●	●	●	●	●	◐
MRC	●	●	●	●	●	●	○	◐	●	○	◐
NERC	●	●	●	●	●	●	●	●	●	●	◐
STFC	●	●	●	●	●	●	●	◐	●	◐	◐
Wellcome Trust	●	●	●	●	●	●	●	●	●	◐	●

UPMC SORBONNE UNIVERSITÉS

Le partage des données : un jeu d'acteurs complexe (II)

- Politique d'ouverture des données des éditeurs de revues scientifiques (*ex.* : *BMJ*)

« The BMJ requires sharing of individual patient data for all clinical trials. **This means that trials will be considered for publication only if the authors agree to make the relevant anonymised patient level data available on reasonable request. The BMJ is the first general medical journal to require data sharing for all trials**, extending its initial policy on sharing data for trials of drugs or devices » (juillet 2015)

- Apparition de nouveaux « produits » éditoriaux dédiés au partage et à la valorisation des données : entrepôts de données, valorisation des données au sein des publication, *Data journals* et *data-base paper*

figshare

Data for this Article

caNanoLab
Nanomaterial composition and characterization

ZooKeys

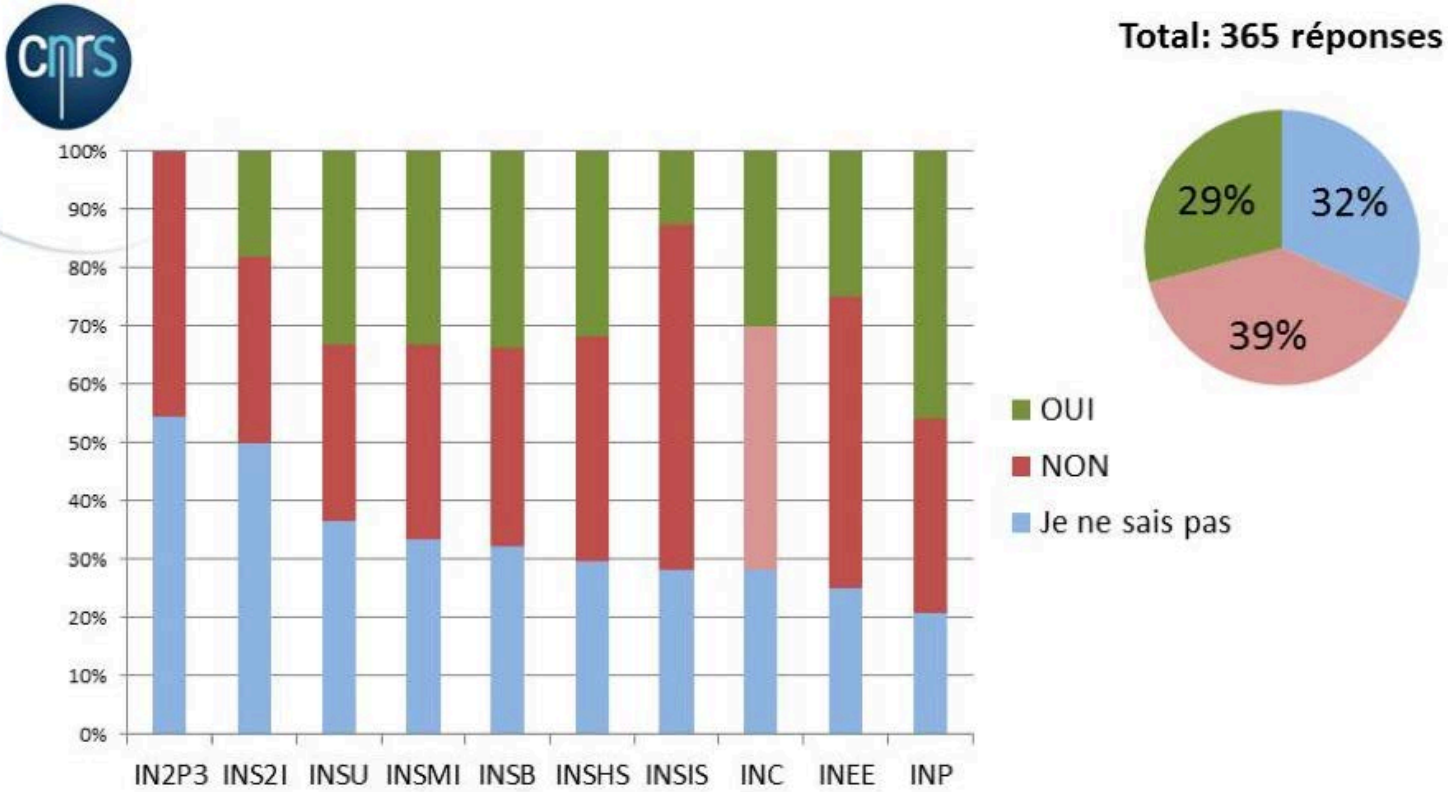
- Initiatives institutionnelles - universités, établissements de recherche…

UPMC SORBONNE UNIVERSITÉS

Une difficulté essentielle : le statut juridique de la donnée

=> Résultats de l'enquête sur les usages et besoins IST des Unités de recherche, réalisée auprès des Directrices et Directeurs d'Unités du CNRS (mars 2015)

"PAP 3 59- Pensez-vous que ces données soient libres de droits?"



Une difficulté essentielle : le statut juridique de la donnée

« Information recueillie dans le cadre du travail d'un chercheur, **susceptible de protection par la propriété intellectuelle** » (Audrey Ego, juriste et doctorante au laboratoire CRDP de l'université de Lille 2)

=> Protection par le droit d'auteur ?

=> Protection par le droit *sui generis* des bases de données ?

=> « Monopoles de fait » ?

=> Contraintes liées à la protection des données personnelles ?

=> Captations éditoriales et problème de la fouille de données (*data mining*)

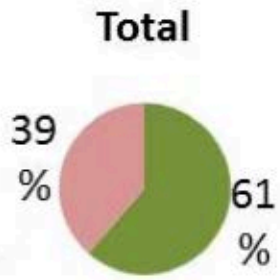
=> Un levier essentiel et « un impératif juridique de simplification » (Pierre Naegelen)

=> « Le chercheur doit rester titulaire de droits sur les données qu'il produit ou qu'il analyse comme sur ses publications, s'il souhaite les partager ou en autoriser la réutilisation. Dans ce cas, il lui est vivement conseillé de mettre ses données protégeables sous une licence libre » (CNRS, avis du COMETS, mai 2015)

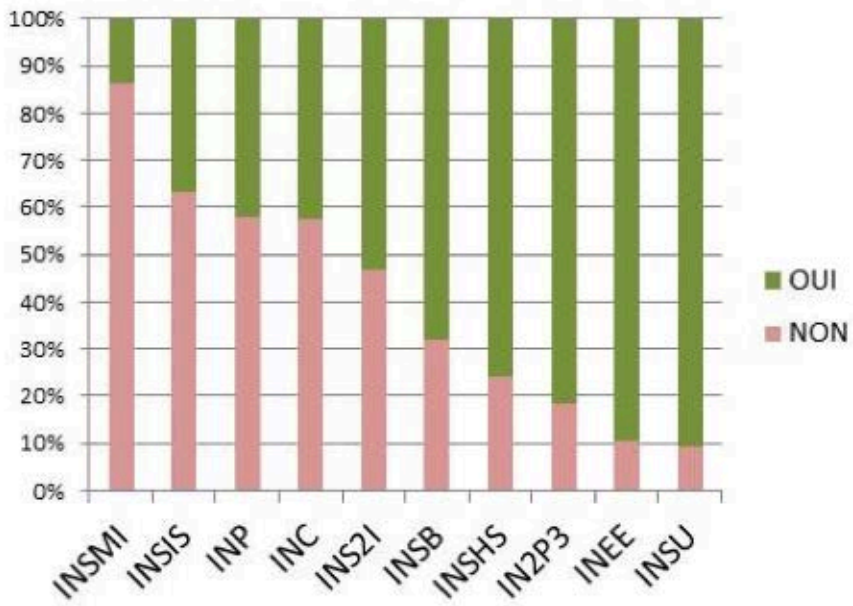
UPMC SORBONNE UNIVERSITÉS

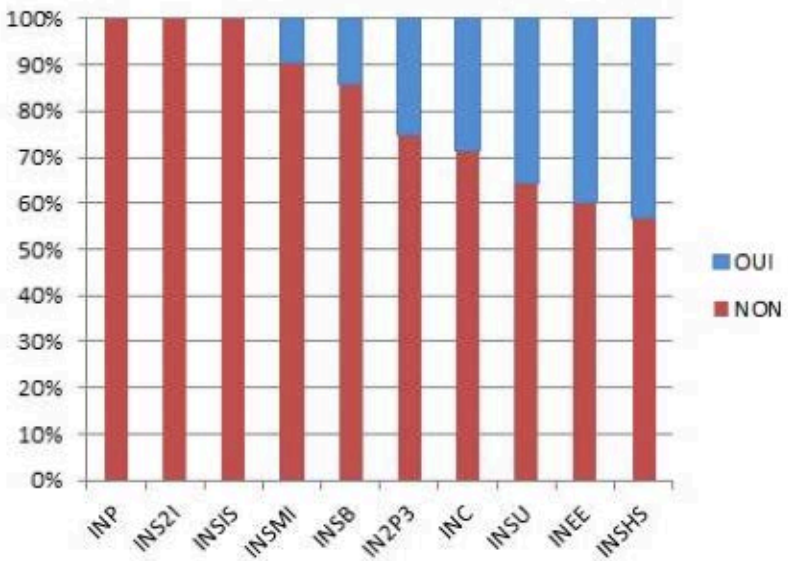
Les établissements de recherche face aux données

"PAP 3 47- Les recherches conduites dans votre laboratoire produisent-elles des données de la recherche nécessitant des pratiques de gestion"



Enquête DIST: avez-vous une activité relative aux données de la recherche ?
OUI:33% NON : 67%





Quelles pistes en France ?

=> **Définition de politiques de gestion des données** à l'échelle des établissements

=> **Promotion de nouveaux modes de valorisation** des données (*ex.* : *Annals of Forest Science*)

=> ***Rewarding* des actions de « curation » des données par les chercheurs** : « le travail de mise à disposition de données utilisables à partir de données brutes doit être reconnu dans l'évaluation et les décisions de promotion des personnels qui s'y impliquent. Pour faciliter cette reconnaissance, le COMETS préconise qu'une rubrique sur ces activités soit ajoutée dans le rapport d'activité et la fiche annuelle d'activité des chercheurs » (CNRS, avis du COMETS, 2015)

UPMC SORBONNE UNIVERSITÉS

Une première étape décisive : systématiser les plans de gestion des données

Data Management at Harvard

- Overview
- Harvard Guidelines and Policies
- Funding Agencies Guidelines and Policies
- Writing a Data Management Plan
- File Naming & Formats
- Sharing Your Data
- Citing Your Data
- File Storage & Costs
- Additional Resources @ Harvard & Beyond

Overview

Why Manage Data?

- **Organize your research so it can be reproduced.** Managing your data makes it easier to understand the details and procedures relating to your data and data collection throughout the lifecycle of the project.
- **Preserve and share your data to get recognition.** The data you collect are the basis of your research. They are your unique contribution, and preserving them means that your work will be recognized by others. It also ensures that your work can support future research and facilitate new discoveries.
- **Satisfy funding requirements and Harvard research policies.** The number of granting bodies requiring that data be preserved and shared is growing. A good data management plan will help you meet the requirements of your funding agency and address preservation, documentation and verification issues. It helps reviewers understand the characteristics of your data.

About This Site

HARVARD LIBRARY

research DATA collaborative

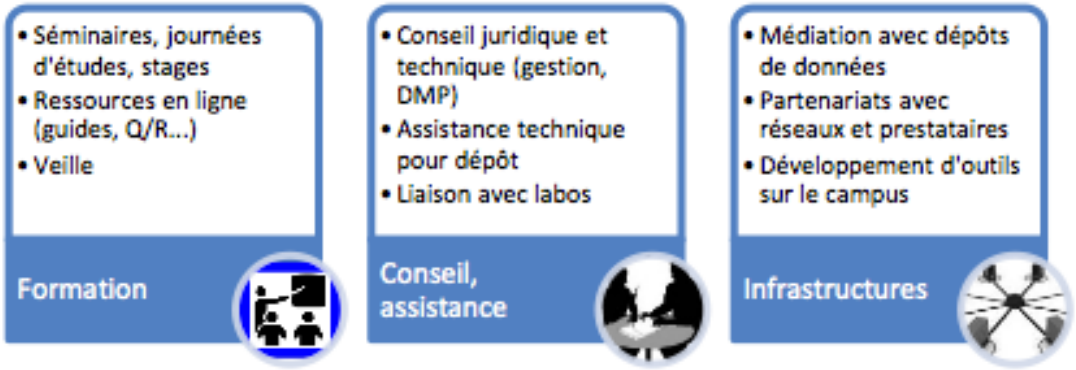
This site has been created by the Harvard Library and the Research Data Collaborative to assist researchers with the complex task of data management. See detailed site credits.

UPMC SORBONNE UNIVERSITÉS

Pour conclure, une initiative récente à Lille 3 : la gestion des données associées aux thèses

Architecture de l'offre

Contribuer à la gestion des données de la recherche dans les thèses de doctorat en SHS de l'Université de Lille 3 est un projet certes limité mais d'une grande valeur symbolique, d'autant plus qu'il témoignera de l'ambition – et de la capacité – de notre Université à relever le défi des données. L'offre de service à mettre en œuvre s'articule autour de trois axes – la formation des doctorants et jeunes chercheurs ; le conseil et l'assistance personnalisée (tout au long du cycle de vie de la donnée, de la préparation d'un *DMP* au partage en ligne) ; la mise à disposition des infrastructures nécessaires pour la sauvegarde, le stockage et le partage des données. C'est une offre modulaire, avec le potentiel de généraliser les services à l'ensemble du personnel scientifique du campus et ouverte aux évolutions et changements à venir.



=> 2015-2016 : élaboration d'un DMP adapté aux SHS ;

=> 2016-2017 : ajuster le circuit de traitement des thèses pour une prise en charge des données liées ;

UPMC SORBONNE UNIVERSITÉS