LES DONNÉES NUMÉRIQUES DE LA RECHERCHE

JOURNÉE D'ÉTUDE MEDIADIX/URFIST

E. DZALE YEUMO

10 octobre 2014

_01

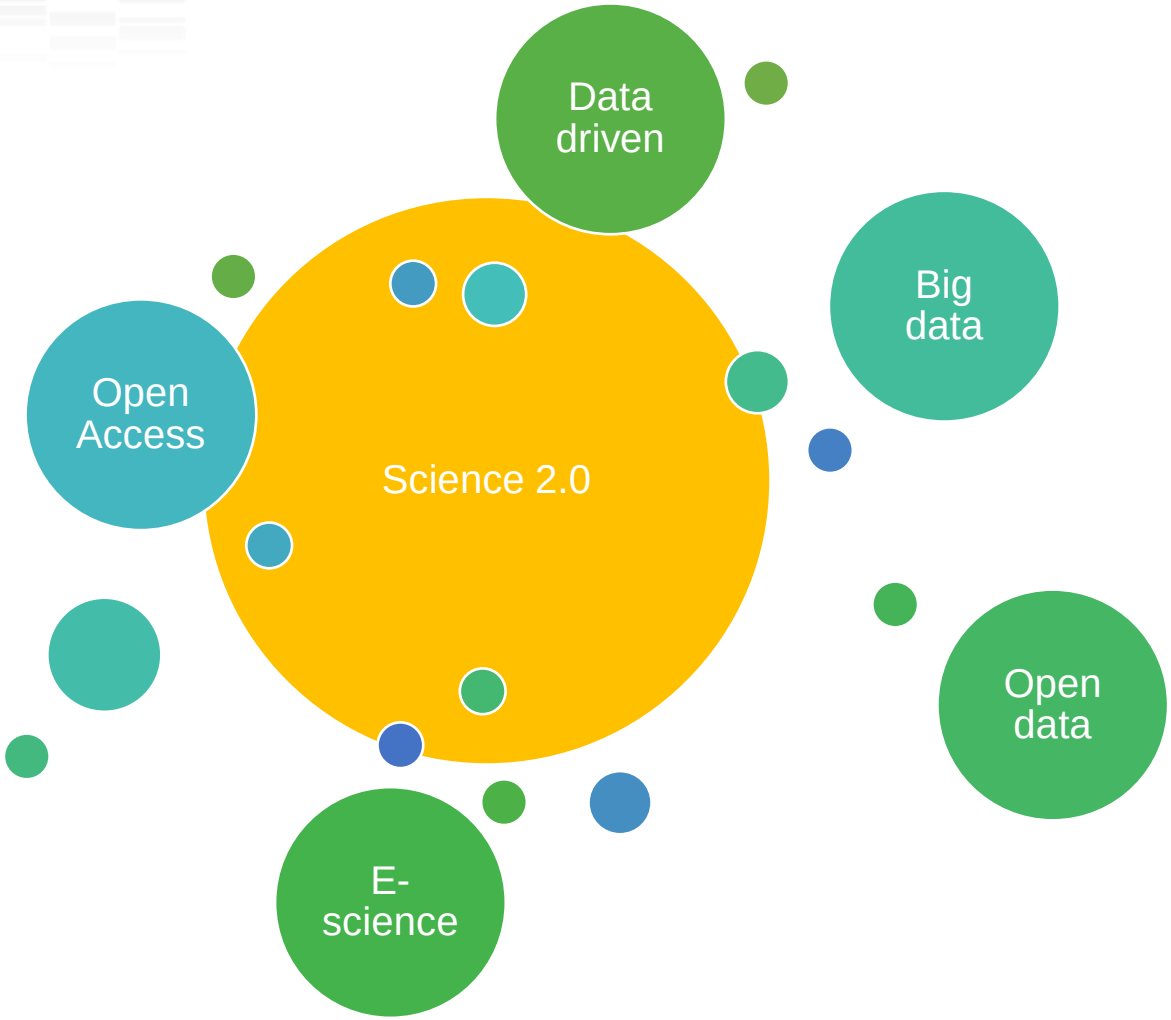
CONTEXTE, DÉFINITIONS

Science 2.0

- ‘Science 2.0’ describes the on-going evolution in the modus operandi of doing research and organising science. These changes in the dynamics of science and research are enabled by digital technologies and driven by the globalisation of the scientific community, as well as the need to address the Grand Challenges of our times. They have an impact on the entire research cycle, from the inception of research to its publication, as well as on the way in which this cycle is organised.

http://ec.europa.eu/research/consultations/science-2.0/consultation_en.htm

Science 2.0


Data driven
Big data
Open Access
Science 2.0
Open data
E-science

Les données de la recherche : définitions

- Les « données de la recherche » renvoient à des choses différentes selon les disciplines scientifiques
- Le périmètre des données de la recherche varie selon les politiques :
  - 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 autres, y compris les documents administratifs.
- Le contexte est important (quand, quelle question scientifique, pour quoi, etc.)
  - Les images d'une ville préhistorique deviennent des données pour un chercheur qui étudie l'histoire de cette ville.

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**.

http://www.oecd.org/fr/science/sci-tech/38500823.pdf

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

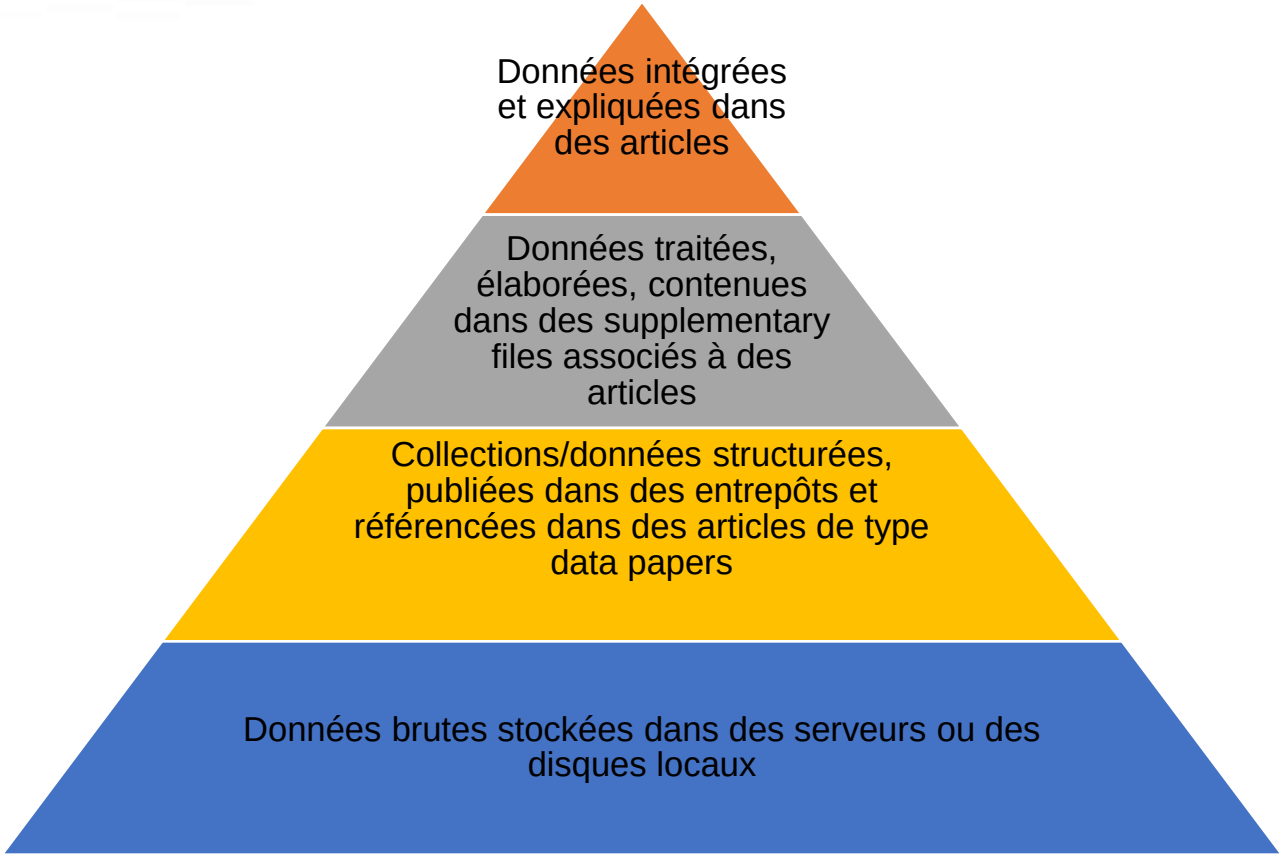
J. Sécurité

K. Efficience

L. Responsabilité de rendre compte

M. Pérennité

États de la donnée



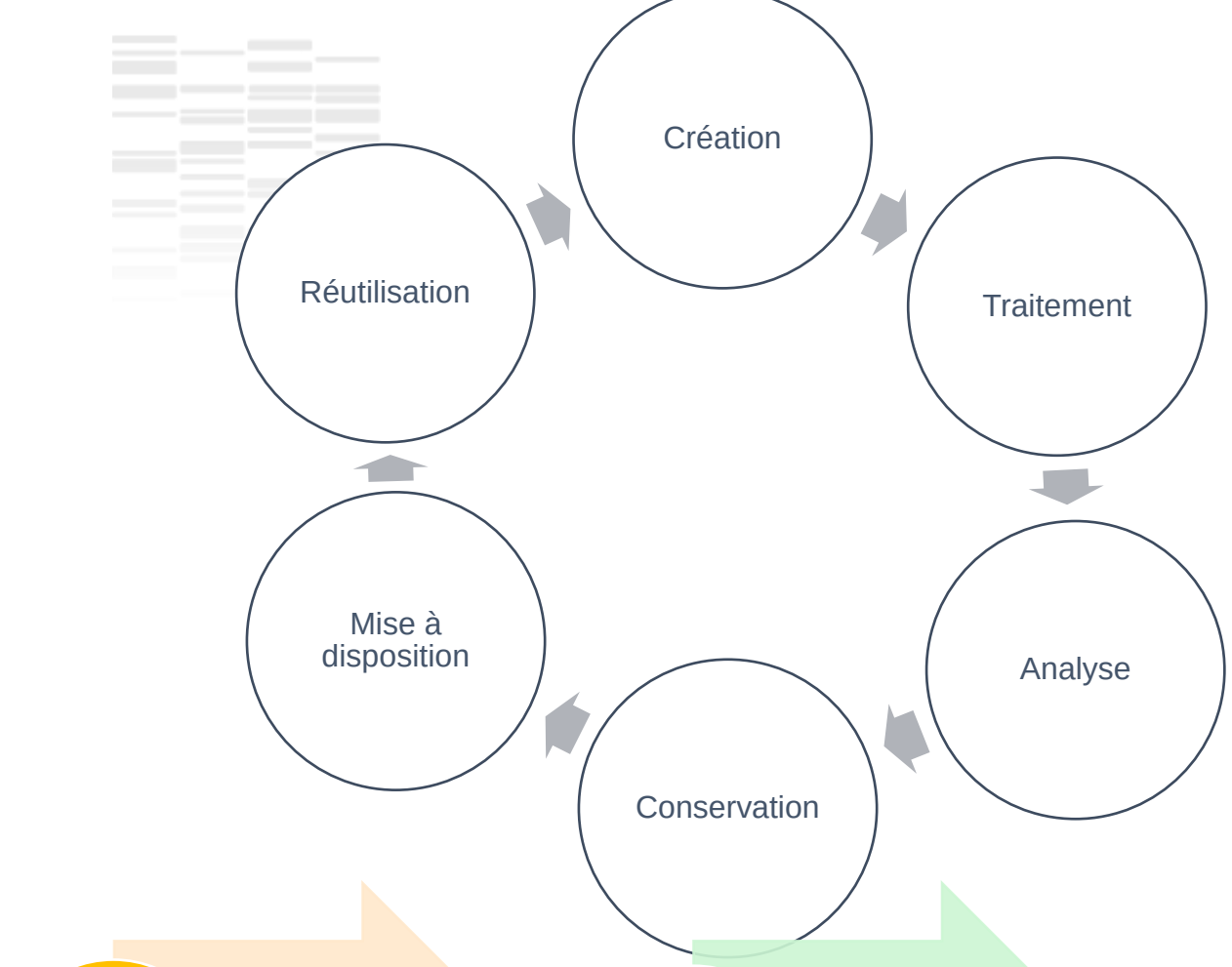
Pyramide des données.
D'après Reilly S., W. Schallier, S. Schrimpf, E. Smit and M. Wilkinson, 2011.
Report on integration of data and publications. Opportunities for Data Exchange (ODE): 1-87.
http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf

Exemples de données selon leur état

	Génétique/génomique	Expérimentation/Observation
Brutes	• Séquences lues ADN-ARN • Génotypes SNP • Données d'expression (arrays, qPCR)	• Imagerie • Sorties des systèmes de mesure • volt, ampère, ohm, fréquence, kg, °C, etc…
Élaborées	• 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…

_02

CYCLE DE VIE DE LA DONNÉE



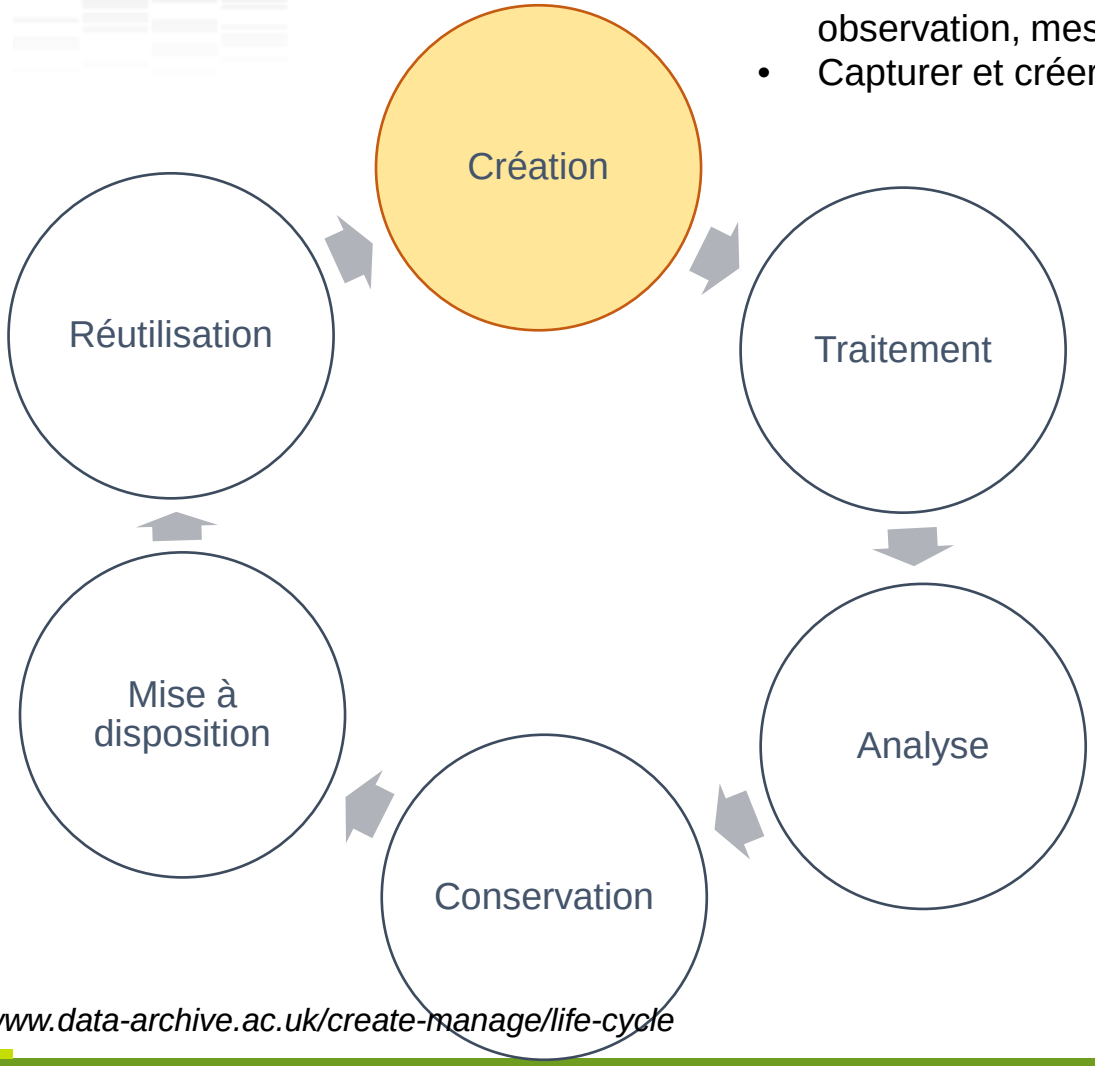
Gestion

Partage

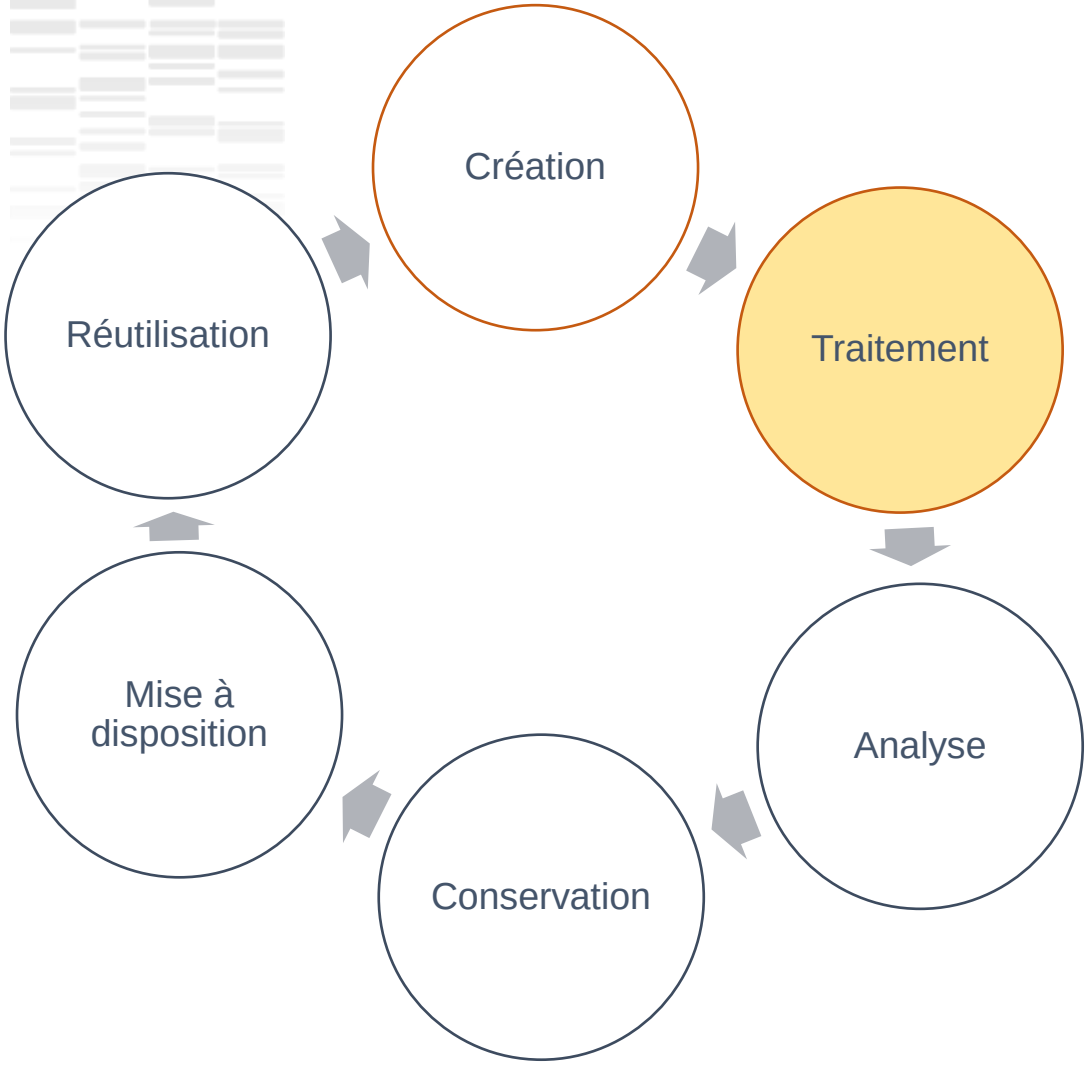
Diffusion
Services

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

- Planifier la gestion des données (formats, stockage etc)
- Définir les contours d'un éventuel partage
- Collecter les données (expérimentation, observation, mesure, simulation)
- Capturer et créer les métadonnées

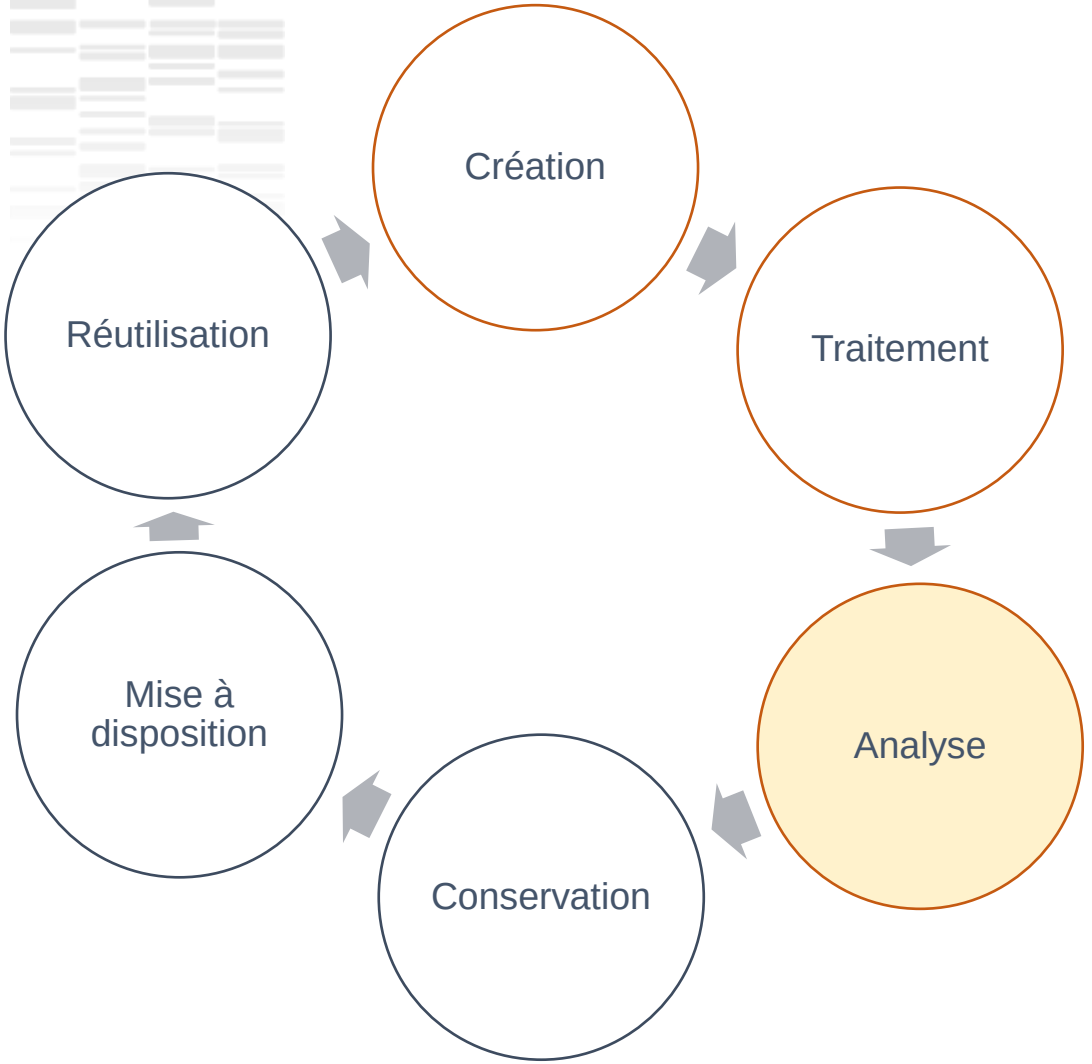


Adapté de http://www.data-archive.ac.uk/create-manage/life-cycle



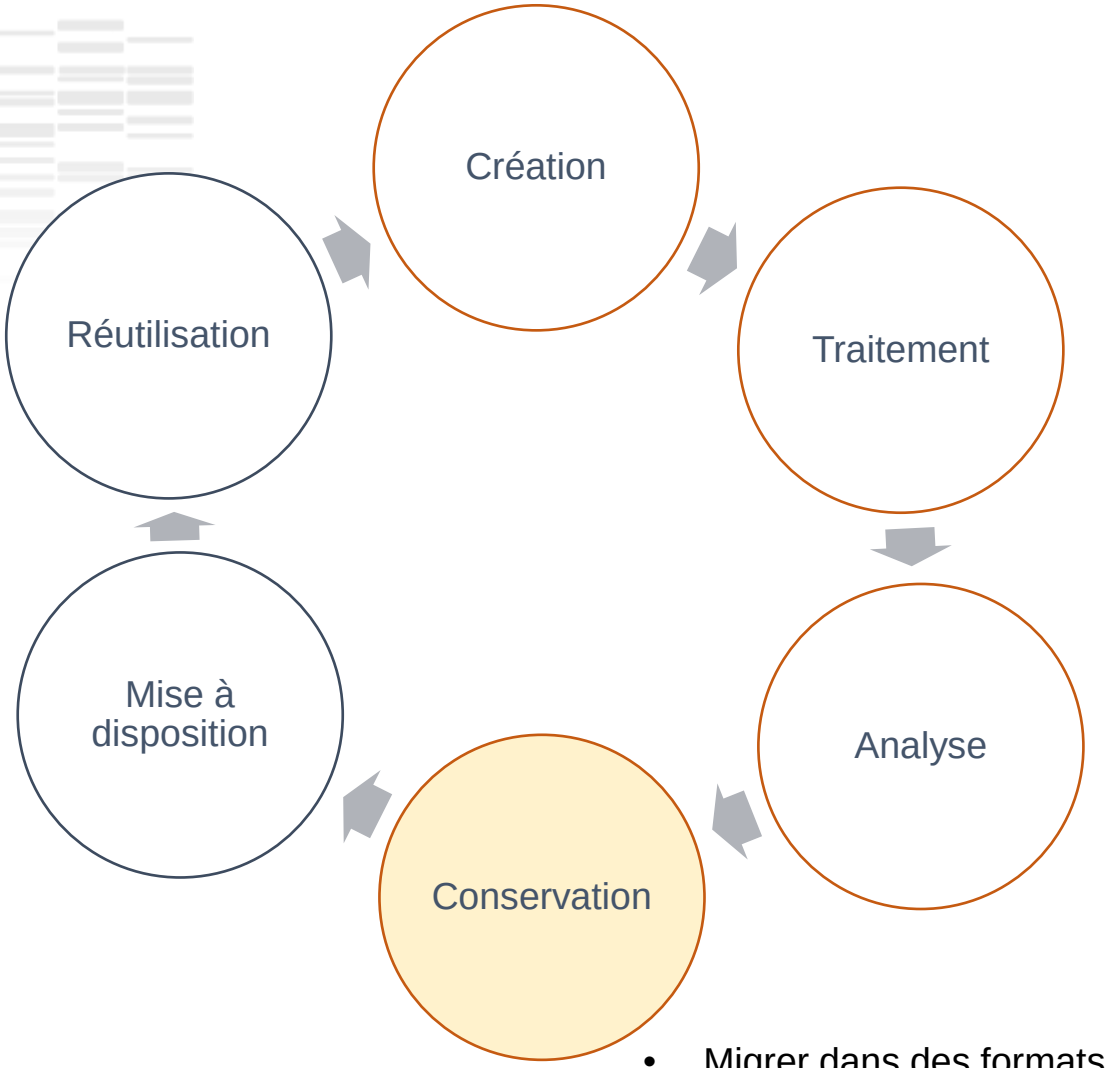
- Saisir, numériser
- Décrire
- Transcrire, traduire
- Vérifier, valider, nettoyer
- Anonymiser
- Organiser, stocker

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



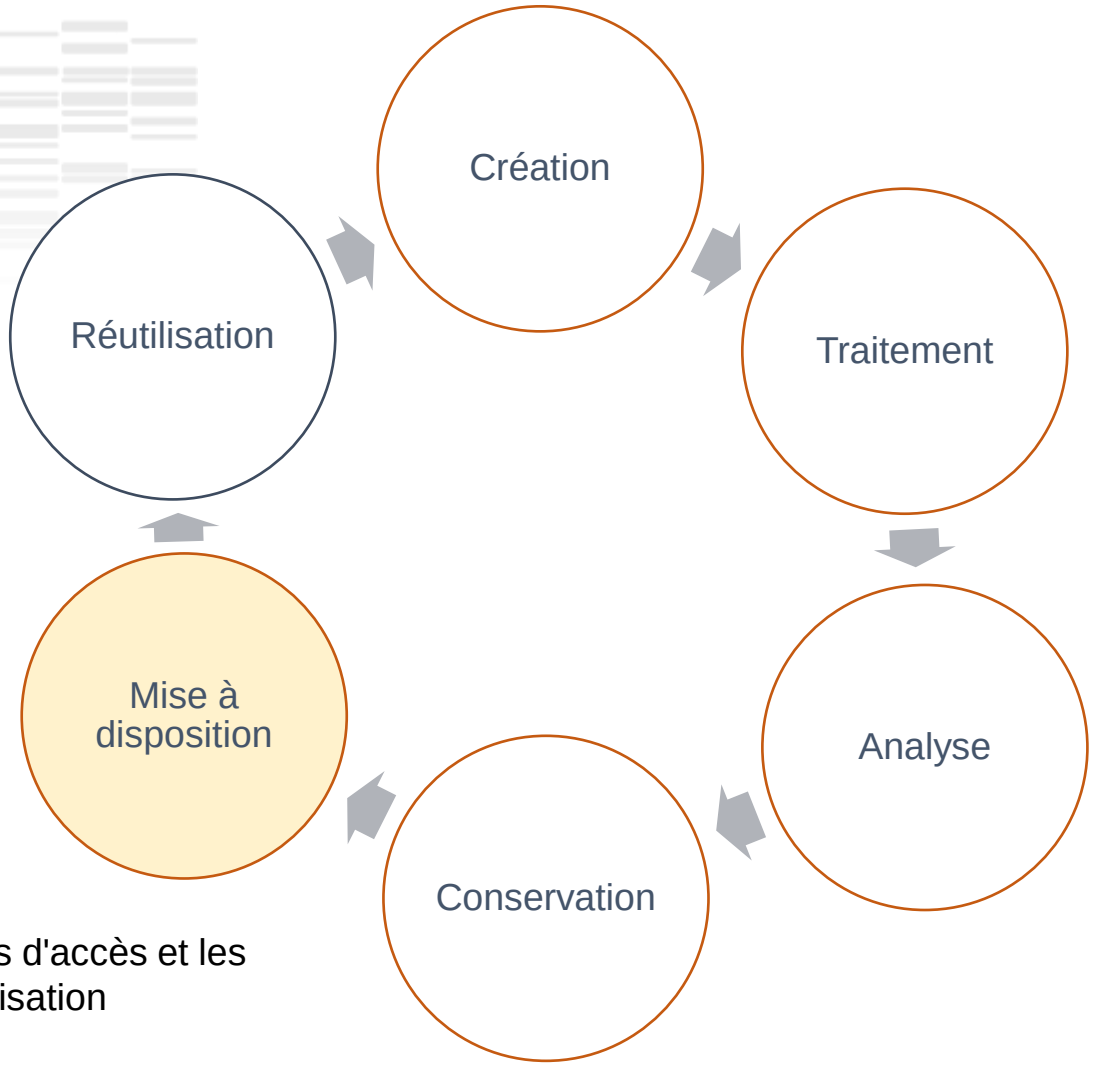
- Interpréter, dériver
- Produire des résultats de recherche
- Préparer les données pour la conservation
- Publier

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



- Migrer dans des formats durables
- Migrer dans des supports adéquats, durables
- Créer la documentation et les métadonnées
- Archiver

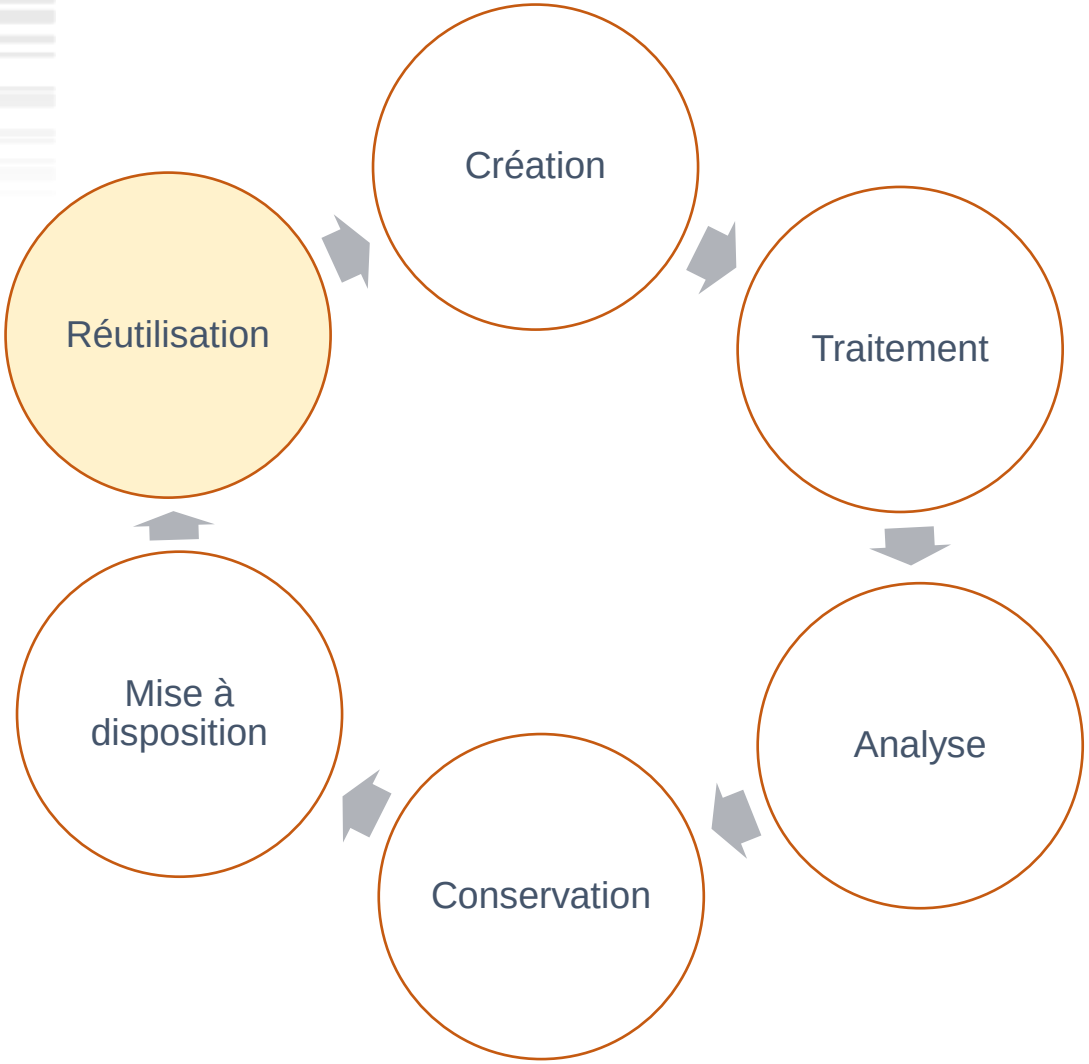
Adapté de http://www.data-archive.ac.uk/create-manage/life-cycle



- Partager
- Définir les droits d'accès et les conditions d'utilisation
- Promouvoir

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

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

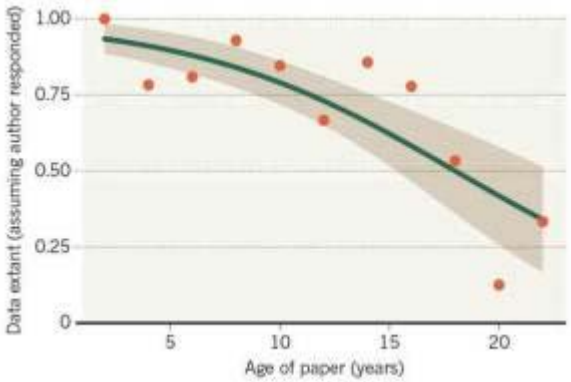


Adapté de http://www.data-archive.ac.uk/create-manage/life-cycle

Enjeux de la gestion

- Coût de récupération des données
- Coût de la perte des données
- répétition des expérimentations

2 ans après la publication d'un article, **les chances d'accéder aux données scientifiques chutent de 17% par an (liens brisés, stockage défaillant, etc.)**



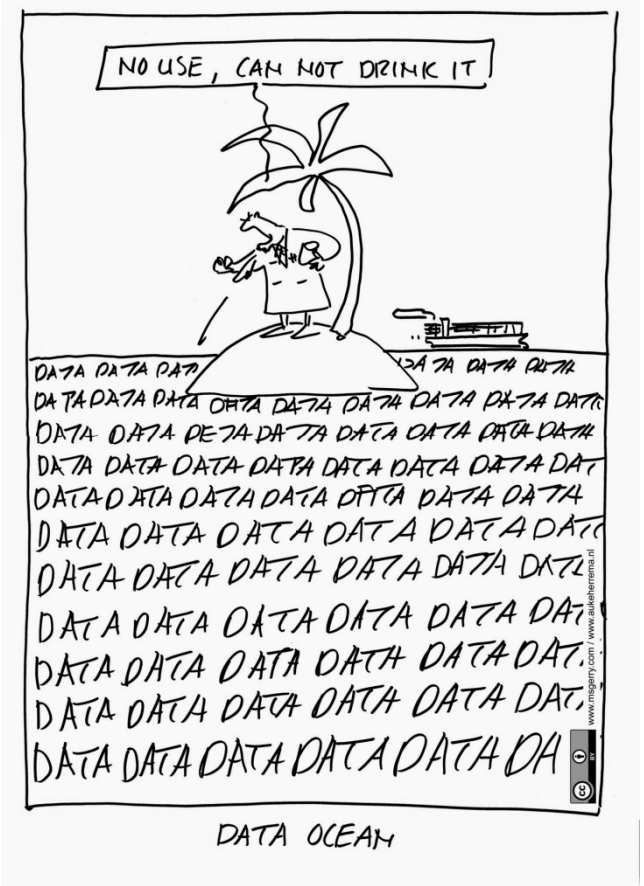
Availability of Research Data Declines Rapidly with Article Age. (Vines TH et al., Current Biology 2014)

Table 2: Total US Data Loss Estimates	
Types of Loss	**Average Cost of Each Data Loss Incident**
Technical Services	340
Lost Productivity	217
Value of the lost data	3,400
Sub Total	$3,957
Total US Data Loss Costs	$18.2 Billion

http://gbr.pepperdine.edu/2010/08/the-cost-of-lost-data/

Enjeux de la gestion

- La donnée numérique est très fragile
  - Une photo même très abimée peut encore être exploitable, un fichier informatique
- Efficacité, qualité de la recherche
- Collaboration



_03

GESTION DES DONNÉES DE LA RECHERCHE

DIFFICULTÉS, LIMITES, 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
  - La difficulté : quelles données, quelles métadonnées de contexte seront utiles plus tard y compris au producteur de données

Des freins au niveau des chercheurs

- Les chercheurs ne sont pas formés à la gestion des données
- Les chercheurs rechignent à consacrer du temps à documenter leurs données s'ils ne perçoivent pas un impact direct sur leurs recherches et leurs productions
- Beaucoup de chercheurs restent sceptiques quand à l'utilité de leurs données sur le long terme
- Une fois un projet de recherche terminé, les chercheurs manquent de temps et de volonté pour documenter les données produites.

Freins organisationnels et techniques

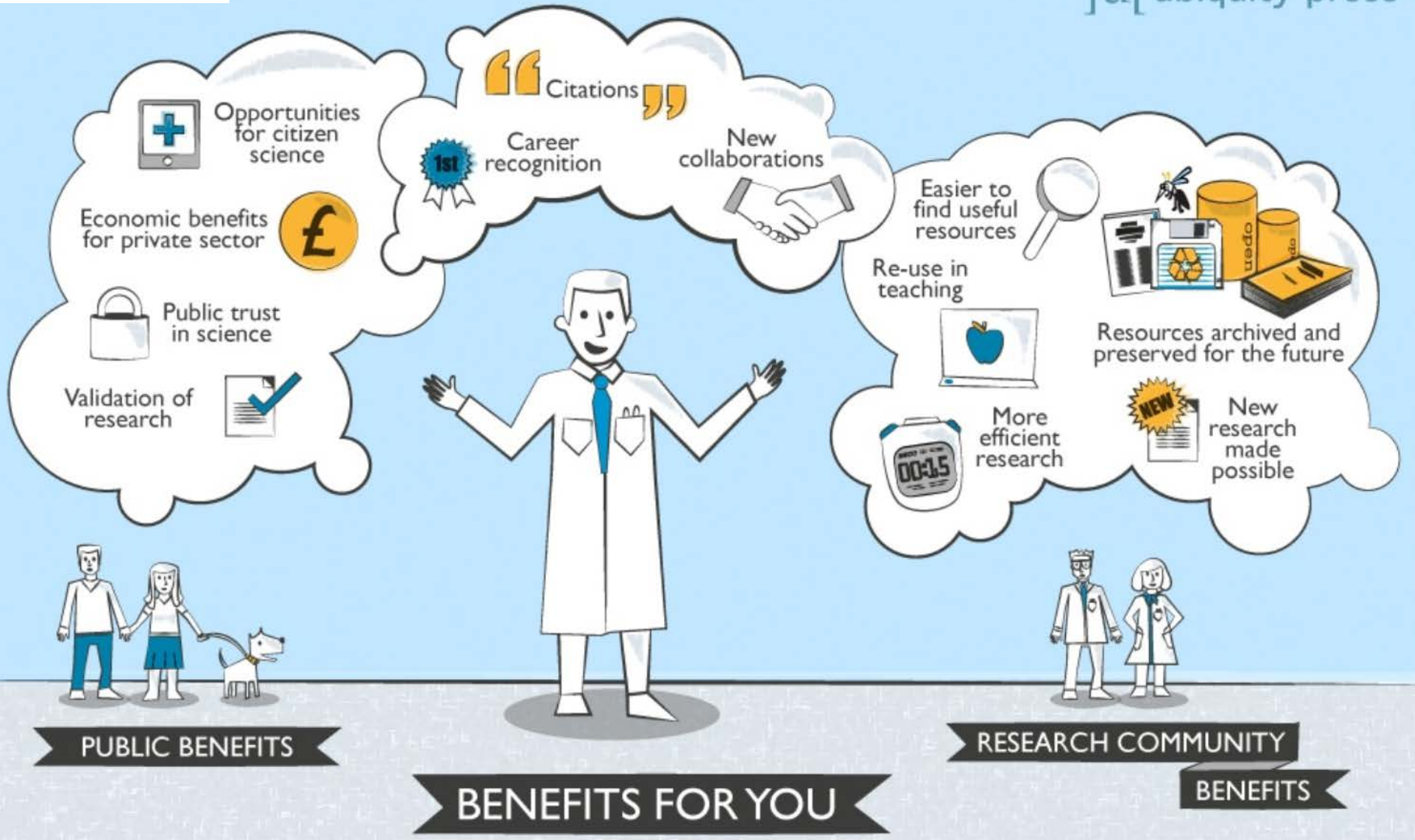
- Manque de politique
- Des pratiques de gestion des données trop diverses (formats, métadonnées)
- Pas de stratégie de préservation
- Manques d'infrastructures adéquats
  - Espaces de stockage sécurisés adéquats
  - Espaces collaboratifs avec une gestion fine des droits d'accès aux données.
- Manque d'incitation
  - Les données ne sont pas encore reconnues dans le processus d'évaluation du chercheur

_04

PARTAGE DES DONNÉES DE LA RECHERCHE

Brian Hole
DPC Workshop, York, 5 July 2013

]u[ubiquity press



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

Pourquoi partager ?

- Par intégrité scientifique
  - Les pairs peuvent vérifier, valider, répliquer, corriger, compléter, etc. vos résultats
- Pour accroître son Impact
  - Publication + données > Publication
  - Réutilisation des données dans d'autres domaines, pays, secteurs
  - Citation par ceux qui utilisent vos données = plus de notoriété
- Pour préserver ses données pour ses propres usages futurs
- Pour contribuer à l'enseignement et la formation

Quand ne pas partager ?

- Quand les données peuvent porter atteinte à la sureté de l'Etat ou à la sécurité des populations
  - Zones de recherche règlementées
- Quand il s'agit de données personnelles
  - Données liées au secret médical
  - Etc.
- Quand les données sont protégées par le secret industriel ou commercial
- Quand les données sont protégées par le secret statistique

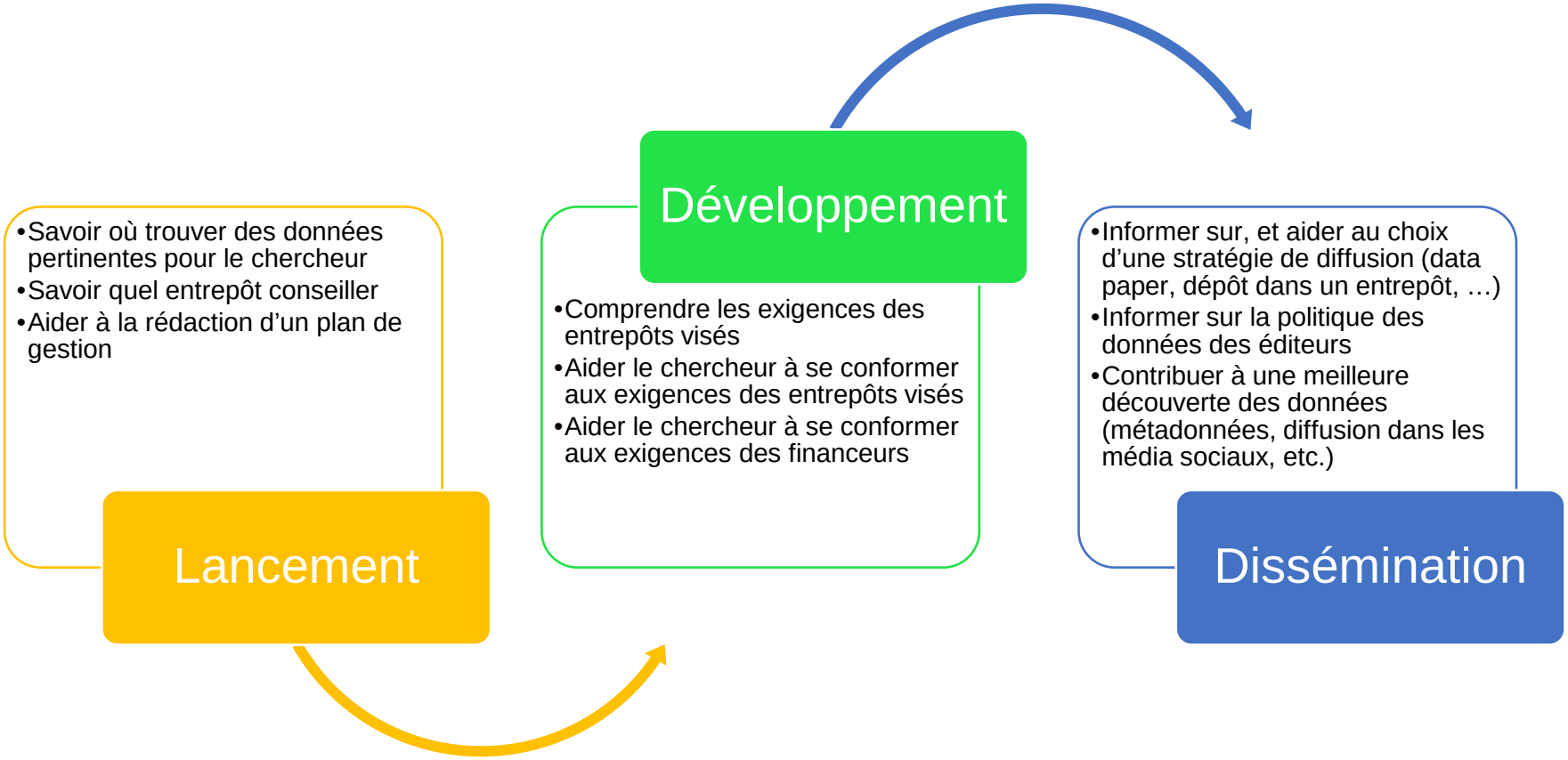
Défis, limites du partage

- Reconnaissance/rétribution : prise en compte des données dans l'évaluation et la carrière des scientifiques.
- Pérennité : mise à 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)
- Références : les données doivent être référençables
- 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.

_05

RÔLE DE L'IST

- Être proactif
- Prendre en compte le cycle de vie de la recherche lors des développements de services, d'outils ou de formations



En pratique

- Tenir à jour une liste d'entrepôts pour son unité/département/centre de recherche. Alerter les chercheurs sur de nouvelles ressources pertinentes, en particulier celles qui ne sont pas spécifiques à leur domaine.
- S'assurer que les entrepôts de son institut sont connus de la communauté scientifique (référencement dans les annuaires pertinents, liens avec des ressources externes similaires, etc.)
- Chercher des données, privilégier les entrepôts thématiques reconnus, les entrepôts internationaux, régionaux, nationaux, etc.
- Se former et former

Exemple de l'IST à l'Inra

- Implication dans des groupes de travail pour la mise en œuvre de la politique de gestion et de partage des données
  - Données et publications
  - Plan de gestion des donnés
  - Entrepôt/annuaire
  - Familles de données
  - Communication
- 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 données de la recherche dans le Web de données
- Ontologies, standards de métadonnées

Références bibliographiques

- MacMillan,D.,DataSharingandDiscovery:WhatLibrariansNeedtoKnow,TheJournalofAcademicLibrarianship(2014), http://dx.doi.org/10.1016/j.acalib.2014.06.011
- Guide gestion des données de la recherche, Université de Bristol : http://data.blogs.ilrt.org/files/2013/11/Research-data-management-overview-V2_0-ccby.docx-1.pdf
- http://datalib.edina.ac.uk/xerte/play.php?template_id=9
- http://policy.monash.edu.au/policy-bank/academic/research/research-data-management-policy.html
- http://research.unimelb.edu.au/integrity/conduct/data/review
- http://www.ed.ac.uk/schools-departments/information-services/research-support/data-library/research-data-mgmt/benefits-sharing-data
- Les images de cartoons utilisés dans les diapositives sont celles de la 4ème plénière de RDA : https://plus.google.com/photos/105390107996739970370/albums/6062285100070235121?banner=pwa&authkey=CM-45JeajoiYxgE
- Module de formation « Une introduction à la gestion et au partage des données de la recherche » proposé par l'Inist : http://www.inist.fr/donnees/co/Donnees_recherche_web.html

MERCI