Rechercher, gérer et valoriser les données pour sa thèse

Annabelle Filatre (AgroParisTech)

Hanka Hensens (IRD)

Université de Montpellier– CIST Agropolis

Introduction aux Données de la recherche

Définitions

- Pour l'OCDE (2006) :
  « **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 les résultats** de la recherche. »
  (repris par le *Plan National pour Science Ouverte* - 2018)

- Pour la Royal Society of London :
  « Informations **qualitatives ou quantitatives** […] qui sont factuelles. Ces données peuvent être **brutes** ou **primaires** (directement issue d'une mesure), **ou dérivée de données primaires**, mais ne sont pas encore le produit d'analyse ou d'interprétation autres que le calculs. »

THE ROYAL SOCIETY

Définitions

- Pour la Commission Européenne (H2020) :

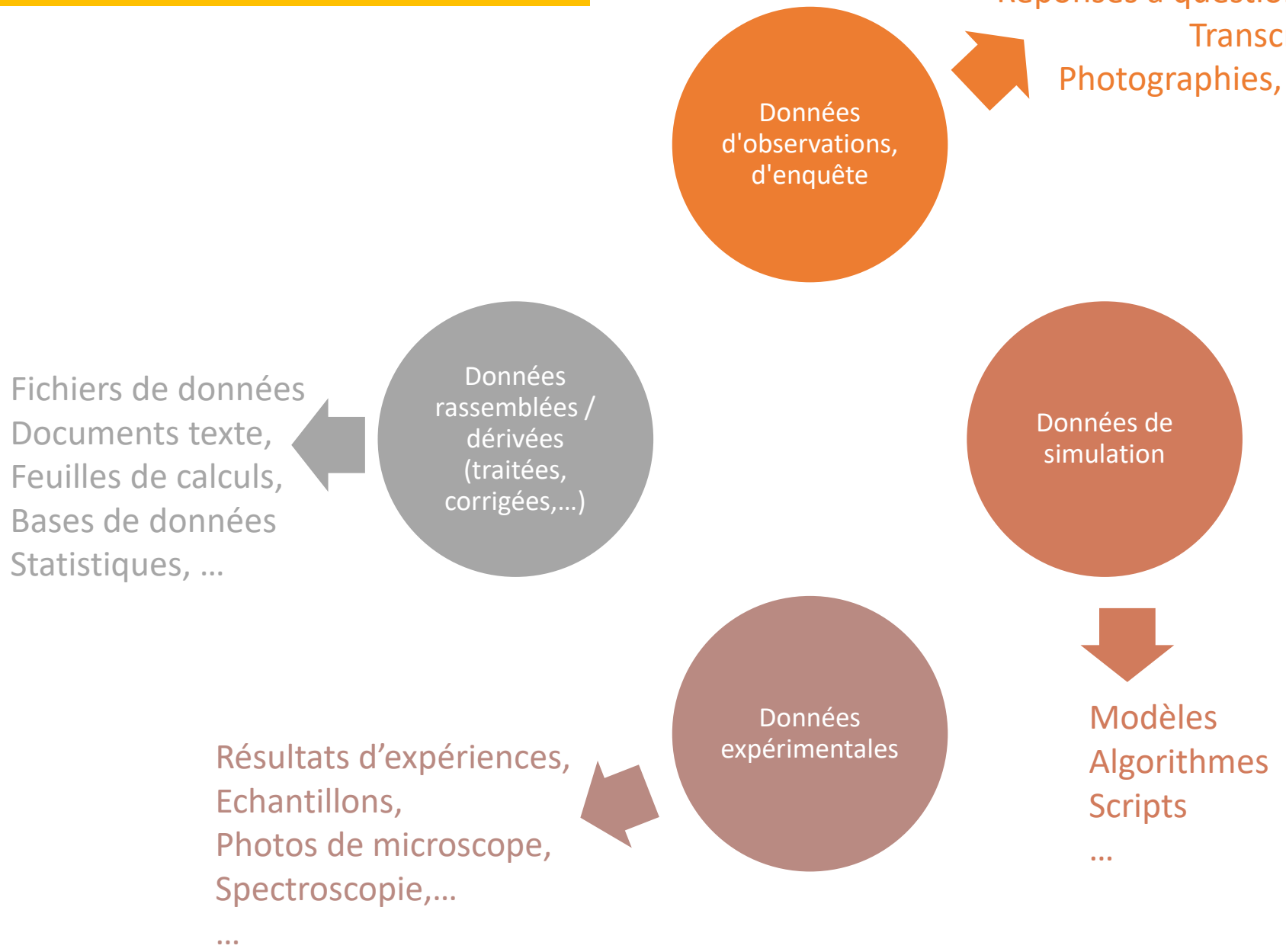
"1. Les **données sous-jacente** (les données **nécessaires pour valider les résultats présentés dans les publications scientifiques**), **incluant les métadonnées associées** (c'est-à-dire les métadonnées décrivant les données de recherche déposées)

2.Toute autre donnée (par exemple les données conservées qui ne sont pas directement attribuables à une publication, ou les données brutes), y compris les métadonnées associées.
(…) L'accent est mis sur les données de recherche **disponibles sous forme numérique**. »

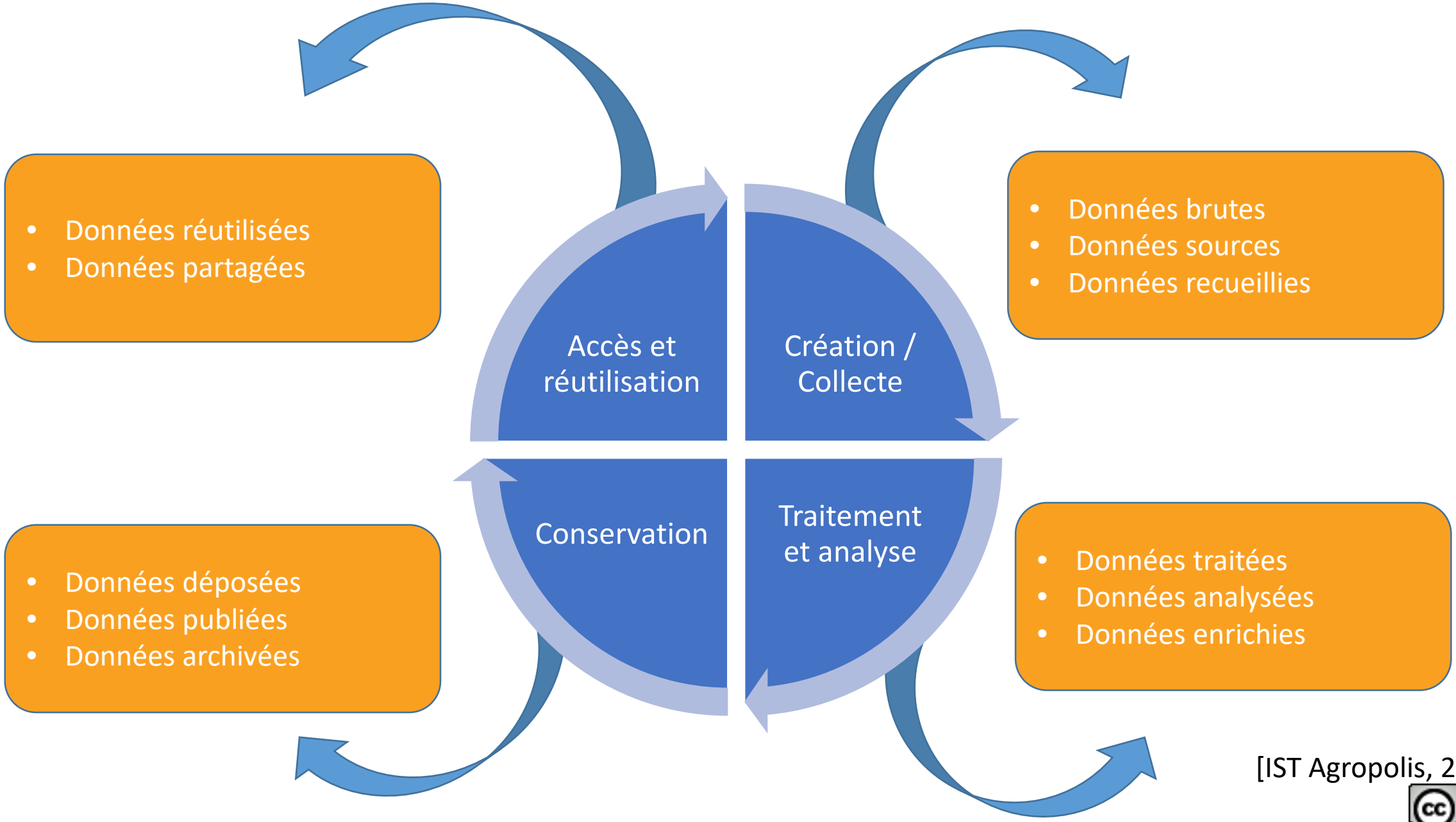
Commission Européenne

HORIZON 2020

Typologie / Méthode d'obtention



Typologie / Cycle de vie des données



[IST Agropolis, 2016]

Ne sont PAS des données de la recherche

Selon l'OCDE :

Les documents non achevés :

- les carnets de laboratoire
- les analyses préliminaires, les projets de documents scientifiques et les programmes de travaux futurs
- les examens par les pairs
- les communications personnelles avec des collègues, …

Les objets matériels :

- les souches bactériennes
- les animaux de laboratoire,…

Les productions scientifiques

- les publications scientifiques, communications à congrès,
- les supports de formation, …

Les données administratives non intégrées dans un corpus de recherche...

Image de Pixabay

Les enjeux de la (bonne) gestion des données scientifiques

Enjeux scientifiques

- Diminution de la perte de données
- Continuité de la recherche
- Recherche multidisciplinaire / systèmes complexes
- Big Data et Data Science

Enjeux économiques

- Bonne gestion budgétaire
- Innovation, valorisation
- Conditions de financement des bailleurs
- Evolution de la publication scientifique

Enjeux sociétaux

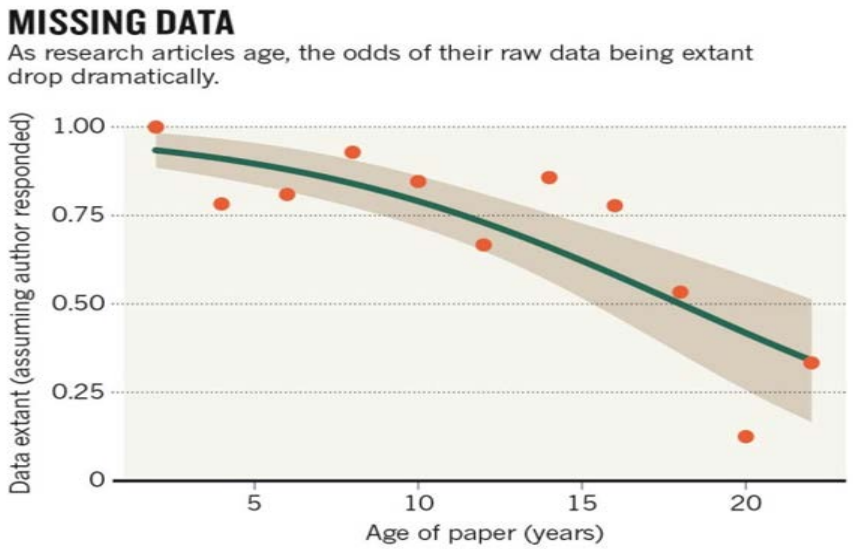
- Transparence et reproductibilité
- Participation citoyenne
- Accroissement de l'impact science / société
- Interopérabilité des données

Enjeux : Perte des données scientifiques

VINES Timothy H., et al. The Availability of Research Data Declines Rapidly with Article Age, Current Biology, 2014.

20 ans après publication,

80 % des données sont perdues

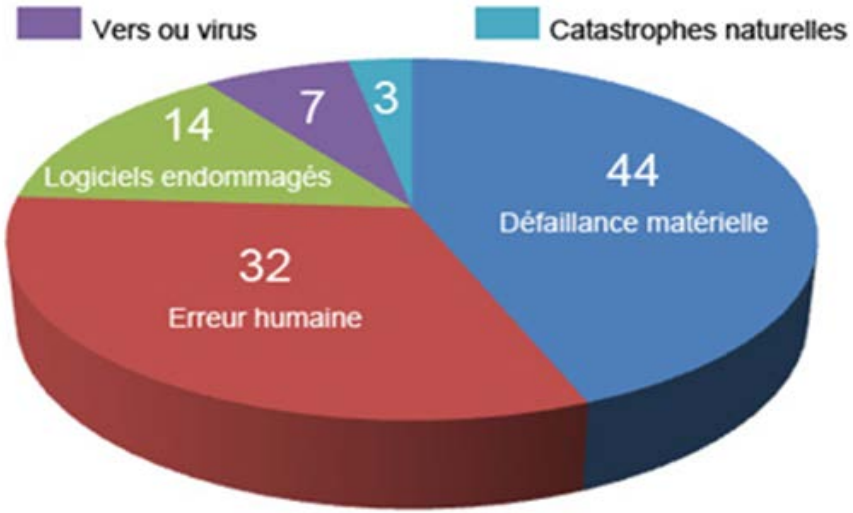


Causes

- ✓ Destruction des supports, virus
- ✓ Obsolescence matérielle ou logicielle
- ✓ Lieu de stockage indéfini
- ✓ Erreurs humaines, départs de personnel

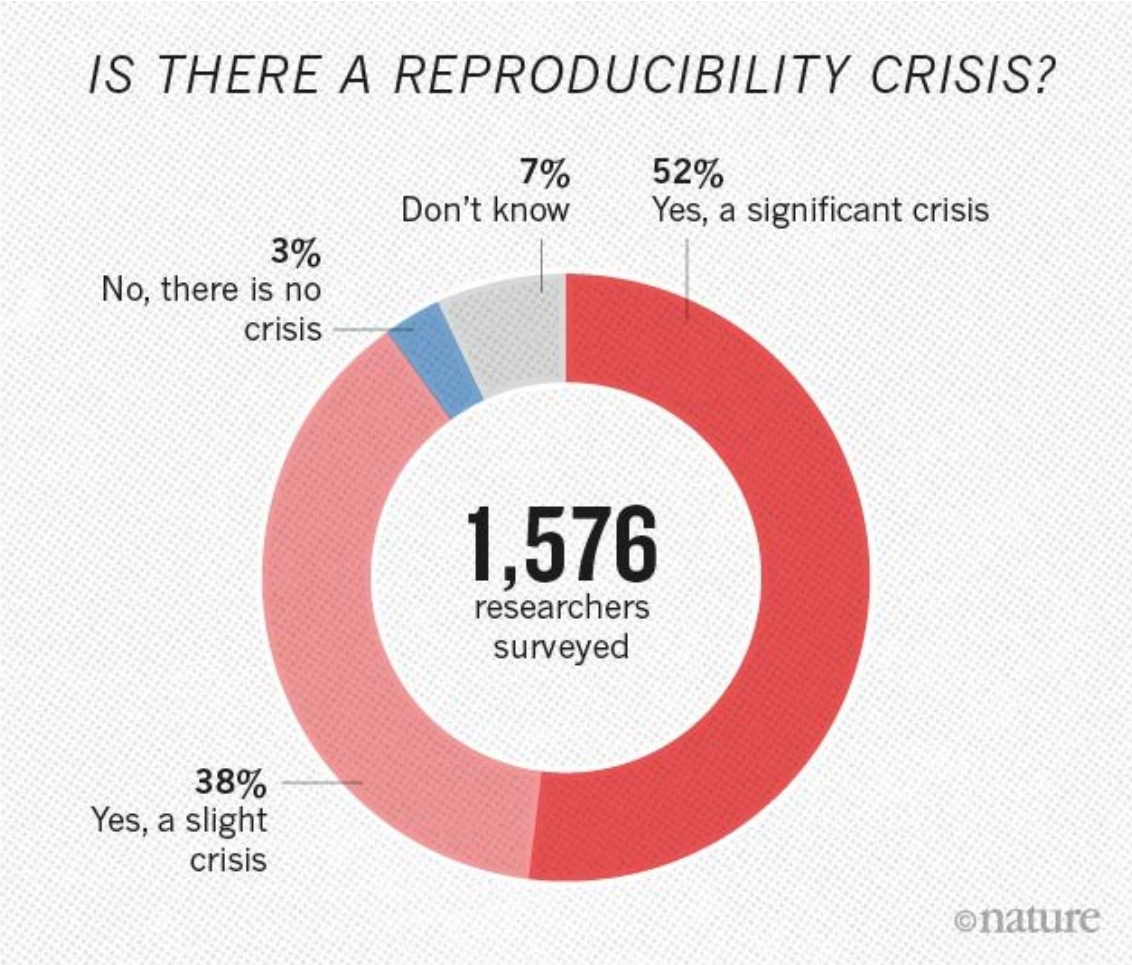
Conséquences

- ✓ Perte de temps, d'argent, de fonds publics
- ✓ Pas de vérification des résultats
- ✓ Pas de comparaison dans le temps ou l'espace
- ✓ Pas de réutilisations par d'autres publics ou pour d'autres fins



Enjeux : Crise de la reproductibilité scientifique

En 2016, 1500 chercheurs répondent à *Nature* :



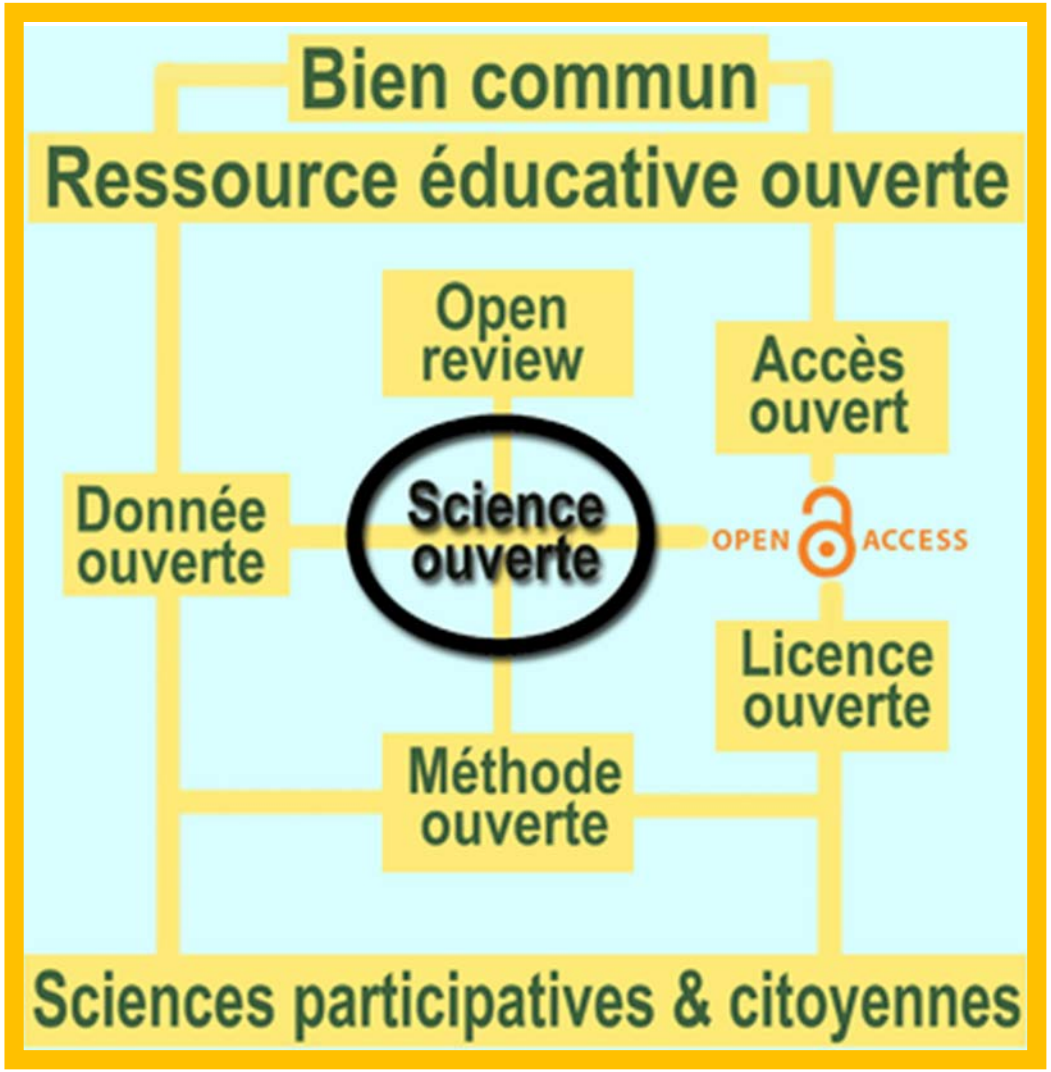
"More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments"

Nature may 2016 : https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Dégât collatéral du « Publish or Perish »…

Enjeux : Vers une Science Ouverte

- ✓ Vise **la transparence et le partage de l'ensemble du processus de recherche**,
- ✓ de la formulation de l'hypothèse à la diffusion des résultats, en passant par les méthodes, données, protocoles, ainsi que l'évaluation, la publication, etc.
- ✓ **Pour susciter analyses, critiques et discussions (publiques) dans le but de le valider et l'améliorer**



Notion de *Science ouverte* et thèmes/enjeux connexes, Wikipedia

Contexte international : Libre Accès et Science Ouverte

Open Data

- Loi Cada (1978)
- Directive PSI (2003)
- Directive Inspire (2007)
- Charte du G8 (2013)
- Loi Valter (2015)

Open Science

- Horizon 2020 (2013-2016)
- Loi pour une République Numérique (2016)
- Plan National pour la Science Ouverte (2018)
- ANR : politique Science Ouverte (2019)

Open Access

- Initiative de Budapest (2002)
- Initiative de Berlin(2003)
- Directive Inspire (2003)
- Déclaration de l'OCDE (2006)
- Plan S (2018)

Contexte National : Plan National pour la Science Ouverte

Juillet 2018

«La France s'engage pour que les résultats de la recherche scientifique soient ouverts à tous, chercheurs, entreprises et citoyens, sans entrave, sans délai, sans payement.» (https://www.ouvrirlascience.fr)

- **Axe 1 Généraliser l'accès ouvert aux publications**
- **Axe 2 : Structurer et ouvrir les données de la recherche**
  - **Obligation de diffusion ouverte des données**
  - Créer la fonction Administrateur des données dans chaque établissement
  - **Données ouvertes associées aux articles scientifiques**
- **Axe 3 : s'inscrire dans une dynamique durable, européenne et internationale**

Contexte national : Politique Science Ouverte de l'ANR

L'ANR DEMANDE LE DÉPÔT des publications issues des projets financés, dans une archive ouverte telle que HAL ou une archive institutionnelle locale

L'ANR DEMANDE L'ÉLABORATION D'UN DMP, plan de gestion des données, pour les projets financés à partir de 2019

https://anr.fr/fr/lanr-et-la-recherche/engagements-et-valeurs/la-science-ouverte/

Droit des données : le cadre juridique français

- D’une manière générale, **les données sont réputées « de libre parcours »** : elles ne sont pas protégées par le droit d’auteur
- **Les données issues de la recherche sont considérées comme des documents administratifs,** si elles ont été produites dans le cadre d'une mission de service public ou majoritairement grâce à des fonds publics :
  - Elles sont donc **communicables à la demande** (sauf exceptions légales), si elles sont "**achevées**".
  - Elles sont soumises à un principe d’**ouverture par défaut.**
  - Elles doivent être diffusées **gratuitement et librement réutilisables**

Cf. Loi pour une République Numérique (2016)

Droit des données : Exceptions 1- Publication sous condition

1. Les données **personnelles** (RGPD), en particulier les données **sensibles** (CNIL), dont les **données de santé**
2. Les données protégées par le **droit d'auteur** (œuvres originales)
3. Les données qui impliquent **un partenaire étranger ou privé** (Droit sui generis des bases de données)
4. Les données concernant les **ressources génétiques et « connaissances traditionnelles » associées : *Protocole de Nagoya***
5. Les informations pouvant avoir un impact sur la **conservation de la biodiversité** *Code de l'Environnement*
6. Les données présentant des risques pour la protection du potentiel scientifique et technique de la Nation

Droit des données : Exceptions 2-Publication

- Données relatives à la **sécurité publique, sureté de l'Etat et sécurité des l'établissements** : biens, personnes, informatique, …

- **Secret professionnel** : secret des procédés, **secret médical**, secret de l'instruction, secret bancaire, …

- **Secret défense**

Droit des données : Des licences pour préciser les droits de diffusion

En France, selon Décret n° 2017-638 du 27/04/2017 relatif aux licences de réutilisation à titre gratuit des informations publiques et aux modalités de leur homologation,
2 choix possibles seulement :

- **l'ODbL (Open Database License version 1.0),**
  pour contrôler les redistributions et les travaux dérivés, ou pour une diffusion internationale
- **la Licence Ouverte Etalab,**
  si le suivi du devenir des données n'est pas recherché et si les données sont essentiellement distribuées en France

ODbL
stands for
Open Data Commons Open Database License

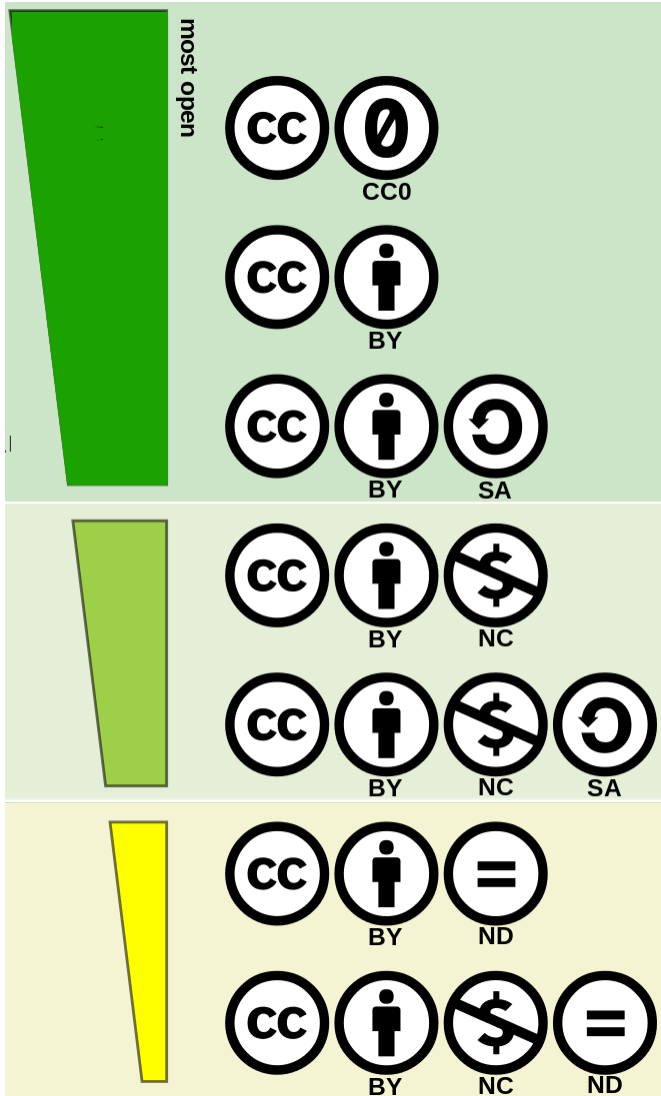
LICENCE OUVERTE
OPEN LICENCE

Droit des données : Licences Creative Commons

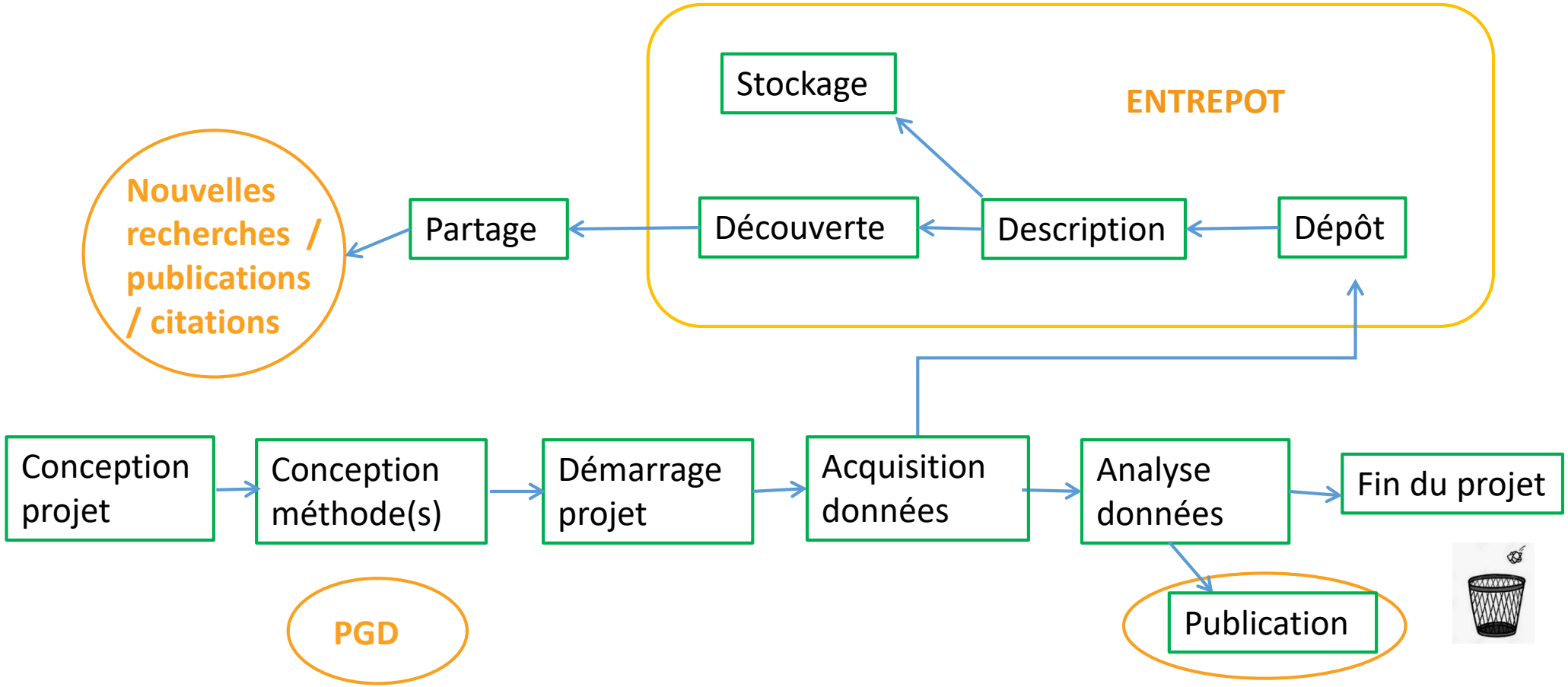
Alternative internationale possible

- **Pour choisir : https://creativecommons.org/choose/?lang=fr**
- **Pour comprendre : https://creativecommons.org/licenses/?lang=fr-FR**
- **5 icônes = 5 droits combinables ⇨ 7 licences**

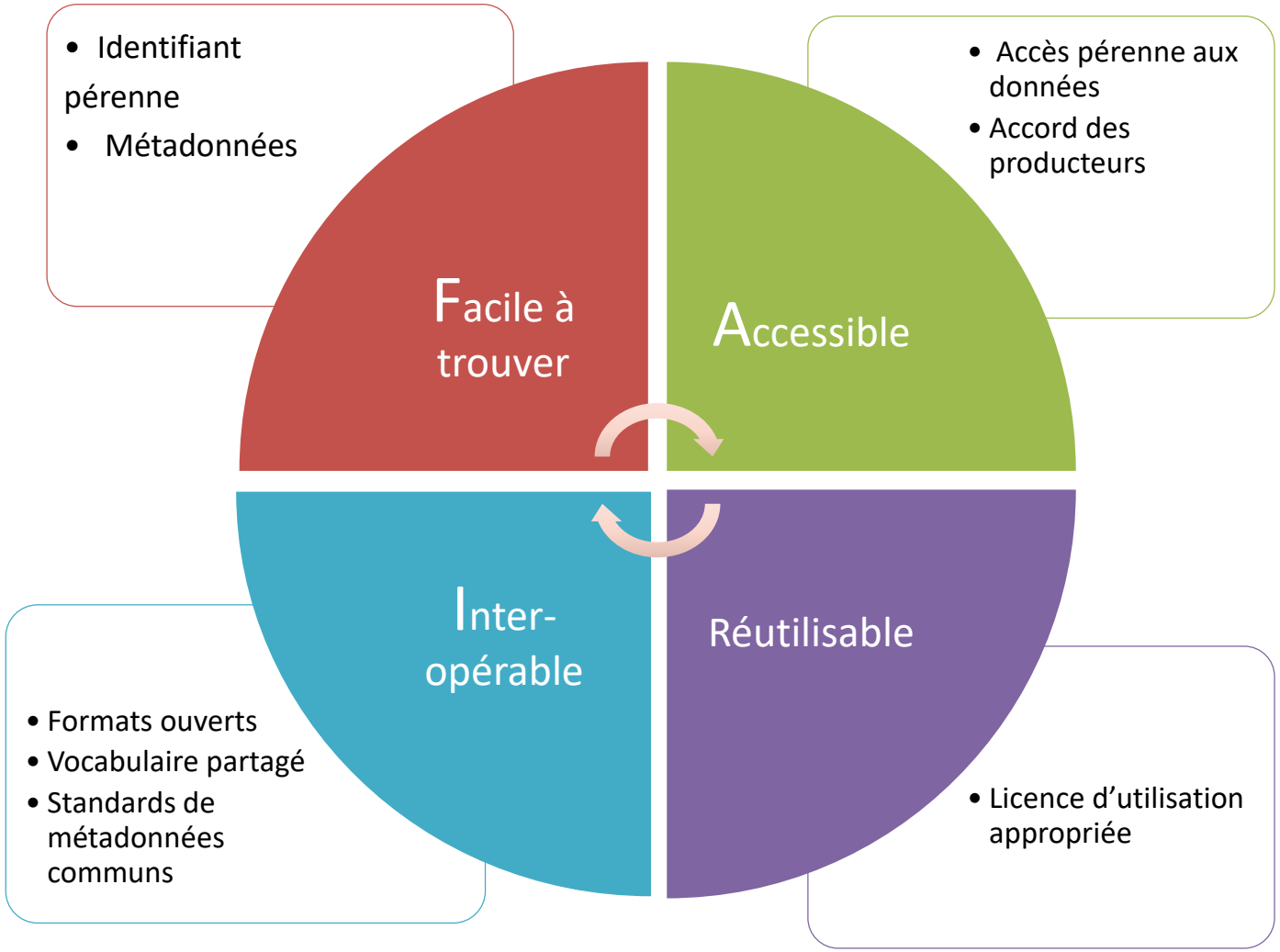
	ZERO	Domaine public
	BY	Attribution = Paternité
	SA	Share Alike = Partage dans les mêmes conditions
	NC	Non Commercial
	ND	No Derivatives = Pas de Modification

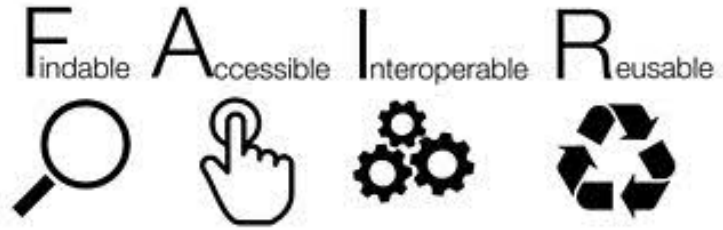


Conclusion: des pratiques à changer…



… dans le respect des principes FAIR





Aussi ouvert que possible, aussi fermé que nécessaire