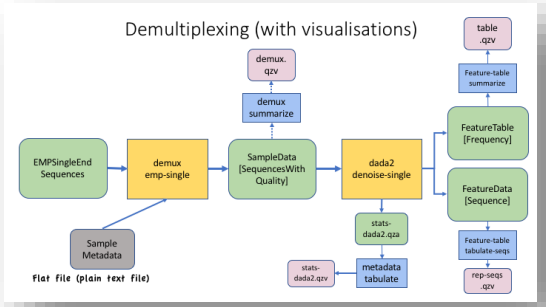
LA GESTION DES DONNÉES DE LA RECHERCHE

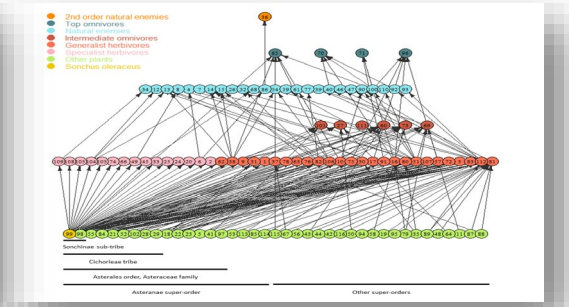
DEVONS-NOUS VRAIMENT TOUT CHANGER ?

JEAN-FRANÇOIS MARTIN - JEUDI 25 MARS 2021

LE CONTEXTE

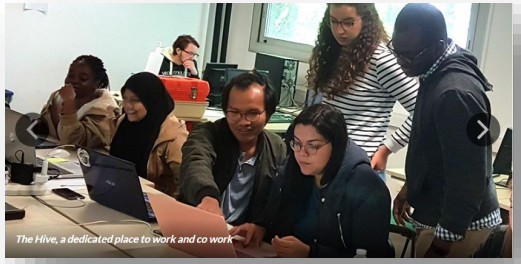
POURQUOI SUIS-JE ICI AUJOURD'HUI ?

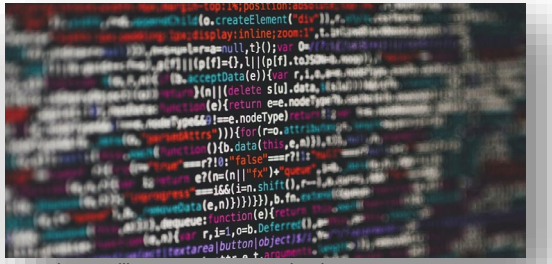




https://qiime2.org/

Des volumes de données multipliés par 10x tous les trois ans





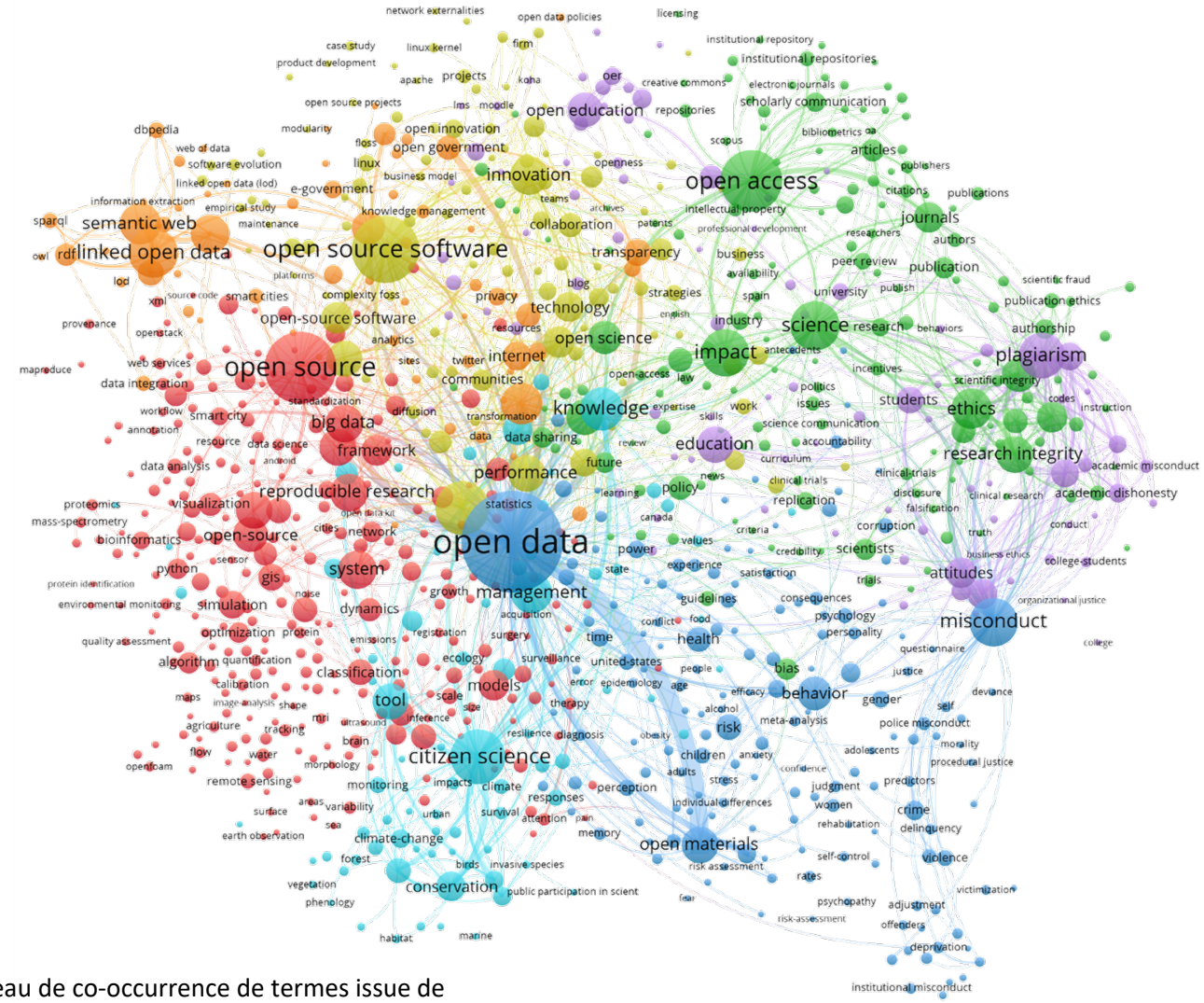
Research Data Alliance - RDMF19 - Costing data management | RDA

Centralité de l'analyse des données dans les démarches d'ingénierie

ENSEIGNEMENT
RECHERCHE

LE CONTEXTE

POURQUOI SUIS-JE ICI AUJOURD'HUI ?



Réseau de co-occurrence de termes issue de l'analyse avec VosViewer de 25 152 articles publiés dans le WOS au 27/11/2019

Les données – un hub des problématiques de la Science Ouverte

SCIENCE OUVERTE

LE CONTEXTE

POURQUOI SUIS-JE ICI AUJOURD'HUI ?

PREMIER AXE : **GÉNÉRALISER L'ACCÈS OUVERT AUX PUBLICATIONS**

1. Rendre obligatoire la publication en accès ouvert des articles et livres issus de recherches financées par appel d'offres sur fonds publics.
2. Créer un fond pour la science ouverte.
3. Soutenir l'archive ouverte nationale HAL et simplifier le dépôt par les chercheurs qui publient en accès ouvert sur d'autres plateformes dans le monde.

DEUXIÈME AXE : **STRUCTURER ET OUVRIR LES DONNÉES DE LA RECHERCHE**

4. Rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics.
5. Créer la fonction d'administrateur des données et le réseau associé au sein des établissements.
6. Créer les conditions et promouvoir l'adoption d'une politique de données ouvertes associées aux articles publiés par les chercheurs.

TROISIÈME AXE : **S'INSCRIRE DANS UNE DYNAMIQUE DURABLE, EUROPÉENNE ET INTERNATIONALE**

7. Développer les compétences en matière de science ouverte notamment au sein des écoles doctorales.
8. Engager les opérateurs de la recherche à se doter d'une politique de science ouverte.
9. Contribuer activement à la structuration européenne au sein du *European Open Science Cloud* et par la participation à *GO FAIR*.

Source : Plan National science Ouverte

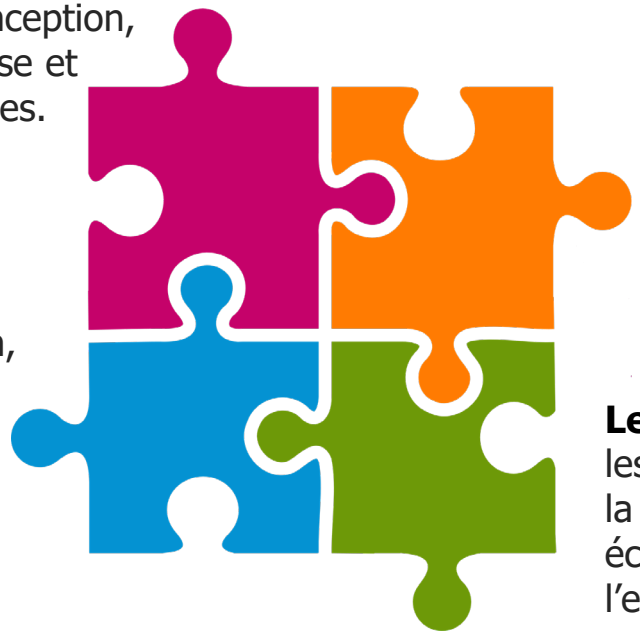
SCIENCE OUVERTE

POURQUOI SUIS-JE ICI AUJOURD'HUI ?

❑ Le respect de valeurs essentielles

La fiabilité dans la conception, la méthodologie, l'analyse et l'utilisation des ressources.

L'honnêteté dans l'élaboration, la réalisation, l'évaluation et la diffusion de la recherche, d'une manière transparente, juste, complète et objective.



La responsabilité pour les activités de recherche, de l'idée à la publication, leur gestion et leur organisation, pour la formation, la supervision et le mentorat, et pour les implications plus générales de la recherche.

Le respect envers les collègues, les participants et participantes à la recherche, la société, les écosystèmes, l'héritage culturel et l'environnement.

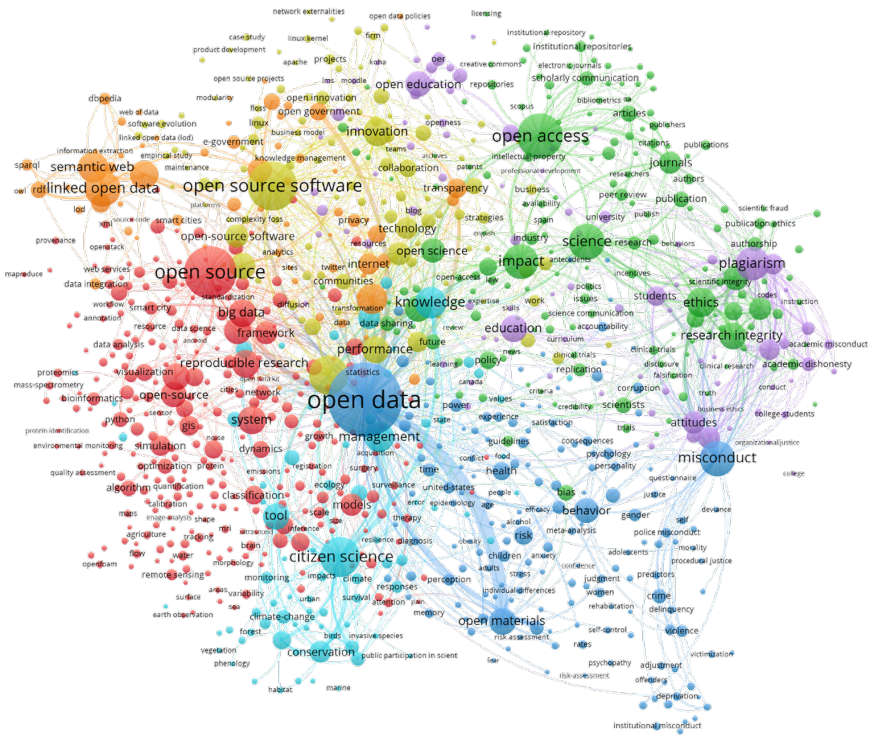
❑ La lutte ferme contre les manquements à l'intégrité scientifique

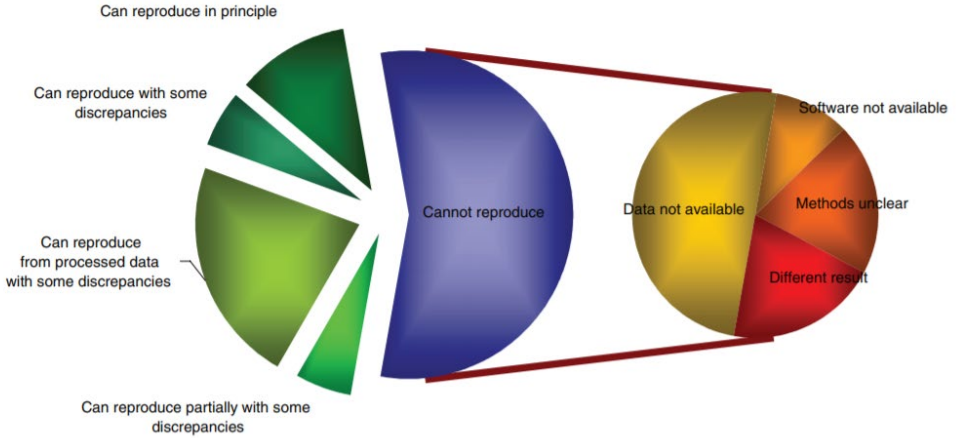
Source utilisée : À propos de l'intégrité scientifique (site Hcères/OFIS, 2019)

INTEGRITE SCIENTIFIQUE

LE CONTEXTE

POURQUOI SUIS-JE ICI AUJOURD'HUI ?





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

Les données – une place centrale dans les bonnes pratiques de la Recherche

Bonnes pratiques — Pratiques contestables — Pratiques frauduleuses

Figure adaptée de René Custers (VIB 2015)

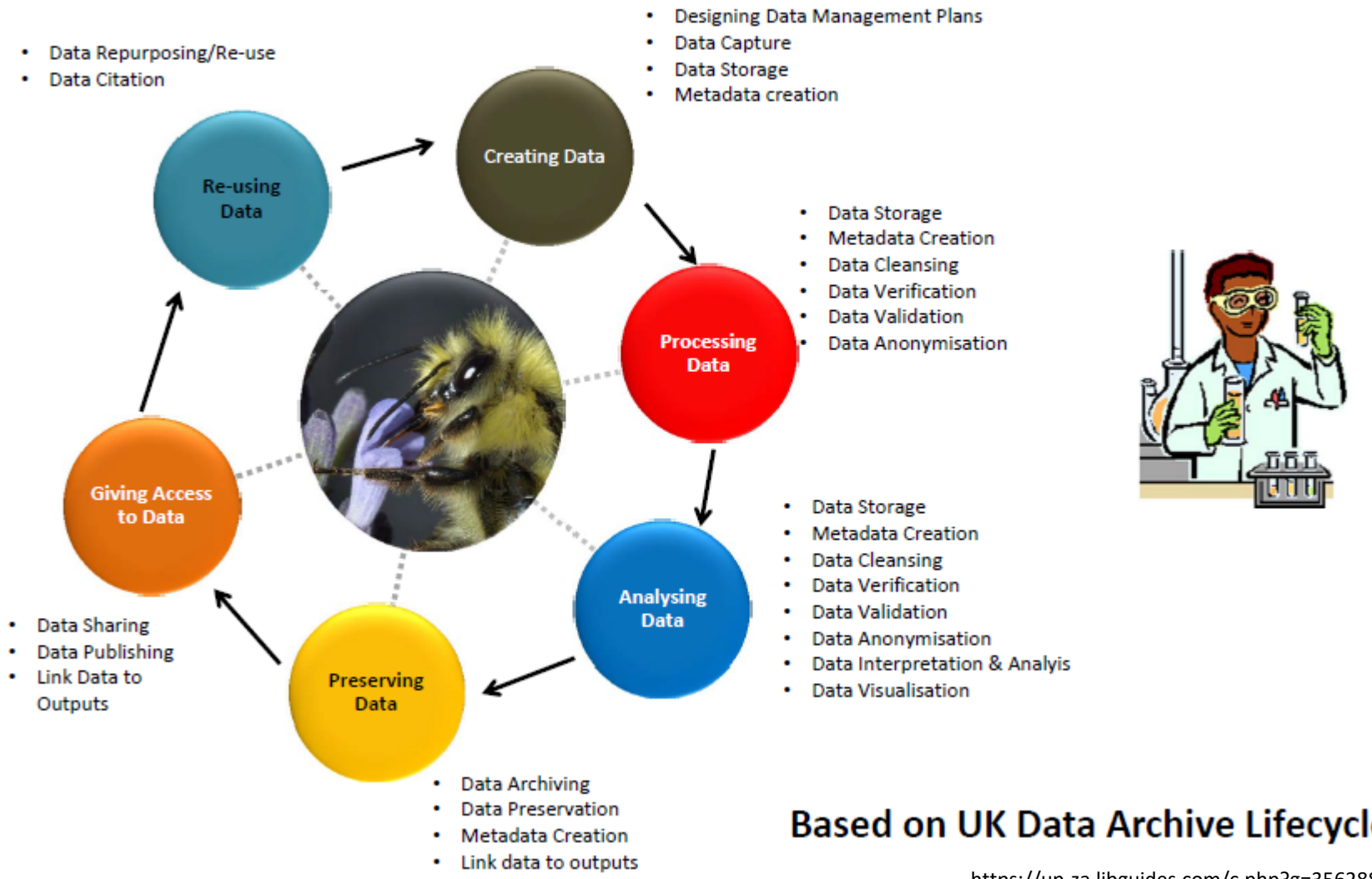
INTEGRITE SCIENTIFIQUE

ASSURER LA QUALITÉ DANS LA GESTION DES DONNÉES

PAR OÙ COMMENCER ?

LE CYCLE DE VIE

DANS LA PRATIQUE C'EST QUOI LA GESTION DES DONNÉES DE LA RECHERCHE?



https://up-za.libguides.com/c.php?g=356288&p=2412654

QUELS OUTILS POUR ACCOMPAGNER LA GESTION DES DONNÉES ?

- Une approche scientifique/méthodologique de la gestion des données
  - Prévoir l'usage
  - Minimiser le risque de perte
  - Documenter les données
  - Analyser la pertinence
  - Évaluer l'efficacité
  - Faire évoluer sa pratique
- Une approche pragmatique
  - Simple à comprendre
  - Simple à mettre en place
  - Simple à évaluer
  - Simple à faire évoluer

} Document texte → Plan de Gestion des Données (PGD)
Data Mangement Plan (DMP)

DMP OPIDoR : LA DÉMO

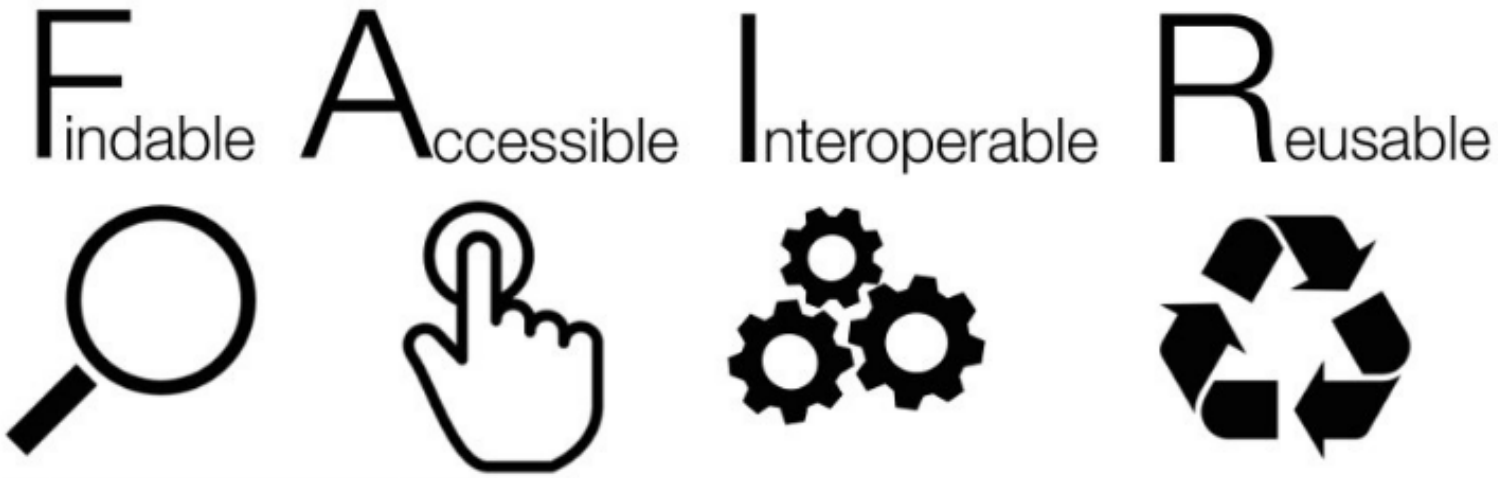
https://dmp.opidor.fr/

MAXIMISER LA RÉUTILISATION DES DONNÉES

QUELLES PRÉCAUTIONS PRENDRE ?

LES PRINCIPES FAIR

MAXIMISER L'IMPACT DES DONNÉES



GO FAIR

LES PRINCIPES FAIR

MAXIMISER L'IMPACT DES DONNÉES

Findable

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

ASSOCIER UN IDENTIFIANT PERSISTENT

PIDS

Une référence durable à un document, un dossier, une page web ou un autre objet : URL, DOI, URI, Accession Number

- En règle générale, un tel identifiant est non seulement persistant, mais il peut aussi faire l'objet d'une action : vous pouvez l'utiliser dans un navigateur web et être dirigé vers la source identifiée.PMID: 27151636 (non-actionable)
  - http://identifiers.org/pubmed/27151636
- DOI:10.1016/j.neuron.2016.04.030
  - http://dx.doi.org/10.1016/j.neuron.2016.04.030
- ***La persistance est un contrat social***

https://en.wikipedia.org/wiki/Persistent_identifier

DOCUMENTER LES DONNÉES : LES METADONNÉES

Les métadonnées sont des « *données qui décrivent des données* » :

- **Information** structurée associée à un "objet", un document ou un jeu de données
- **Documentation** qui permet à l'utilisateur de comprendre, de comparer et d'échanger le contenu du jeu de données décrit

Un objet sans étiquette n'est connu que de son auteur

Il existe des **standards** de métadonnées :

- Standards minimaux (ex : Dublin Core)
- Standards métiers (ex : EML, DDI…)

Il est **conseillé** de produire les métadonnées au **moment de la collecte ou de la création** des données plutôt qu'*a posteriori*. Les métadonnées seront **complétées tout au long du cycle de vie** des données.

LES PRINCIPES FAIR

MAXIMISER L'IMPACT DES DONNÉES

Findable

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

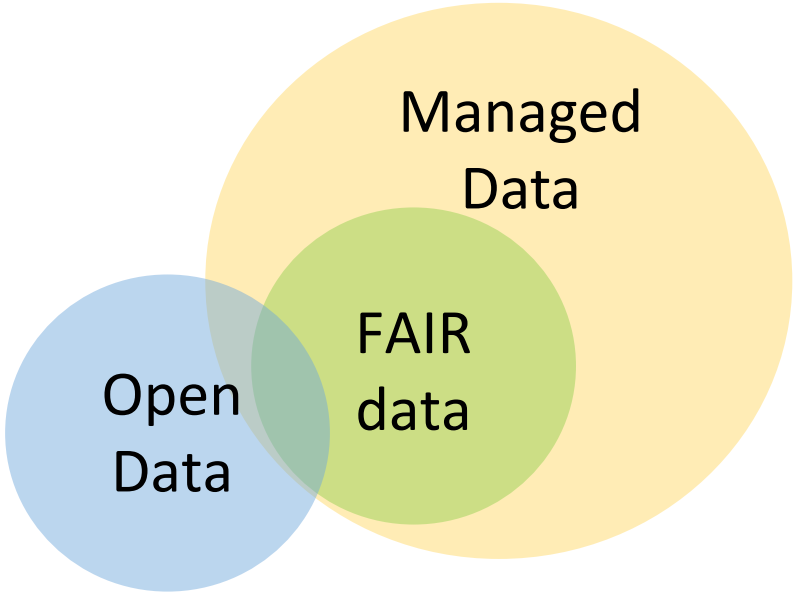
R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

GO FAIR

ACCESSIBLE N'EST PAS SYNONYME D'OUVERT

FAIR Data is NOT Open Data



Source: S. Venkataraman (Digital Curation Centre, Univ. of Edinburgh)

As open as possible, as closed as necessary

Grantees have the right to opt-out

- At any time
- But they need to **say why**

Top three reasons for opt-out

1. Intellectual property rights
2. Privacy
3. Might jeopardise project's main objective

Guidelines of FAIR Data Management in H2020

LES PRINCIPES FAIR

MAXIMISER L'IMPACT DES DONNÉES

Findable

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

Reusable

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

GO FAIR

LA VIE DES DONNÉES
PENDANT LE PROJET DE RECHERCHE

QUELLES PROBLÉMATIQUES ?

QUELLES PROBLÉMATIQUES ?

Quelle est votre pratique en ce qui concerne :

1. Leur stockage
2. Leur sauvegarde
3. Leur accès
4. La collaboration autour de ces données

EXEMPLE : STOCKER ET SÉCURISER

Comparatif de systèmes de stockage des données

Support de stockage	Sécurité	Accès	Coût	Remarque d'utilisation
Ordinateur professionnel	★★☆☆ Sujet au piratage informatique, aux détériorations et pannes	★☆☆☆ Pas adapté au partage, nécessite l'utilisation d'un support externe ou d'Internet (mail, cloud…)	★★★★ Pas de coût supplémentaire ou coût peu important	- Pour un stockage temporaire - Nécessité de crypter les données confidentielles et sensibles
Support externe	★☆☆☆ - Sujet au vol, à la perte du support - Durée de vie limitée (dégradation du matériel)	★★★☆ Facilement transportable, il permet de transférer les données vers un autre ordinateur	★★★★ Pas de coût supplémentaire ou coût peu important	- Pour un stockage temporaire - Nécessité de crypter ou de sécuriser physiquement les données confidentielles et sensibles
Serveur institutionnel	★★★★ Stockage fiable, durable et sécurisé (contre le vol, le piratage, les incendies…)	★★☆☆ La connexion au serveur institutionnel ne facilite pas le travail avec des personnes extérieures	★★☆☆ Coût assez important mais pas forcément répercuté sur l'usager	- Pour un stockage plus pérenne - Adapté pour le stockage de données sensibles et des versions « stables » de vos données - Toutes les institutions ne proposent pas ce service
Dropbox Google Drive OneDrive Serveur Cloud	★★☆☆ On ne sait pas vraiment où sont stockées les données, ni ce qu'elles deviennent	★★★★ Permet un travail synchronisé avec toutes les personnes ayant été autorisées au partage	★★☆☆ Payant à partir d'une certaine limite de stockage	- Pour un partage avec des personnes externes à l'institution - Ne pas y mettre de données sensibles ou confidentielles - Pas de contrôle sur la procédure de sauvegarde des données

Tableau tiré de http://doranum.fr/le-stockage-des-donnees/

cirad

DIFFUSER LES DONNÉES DE LA RECHERCHE

QUELS ENJEUX ?

QUELLES STRATÉGIES DE DIFFUSION?

LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique (1)

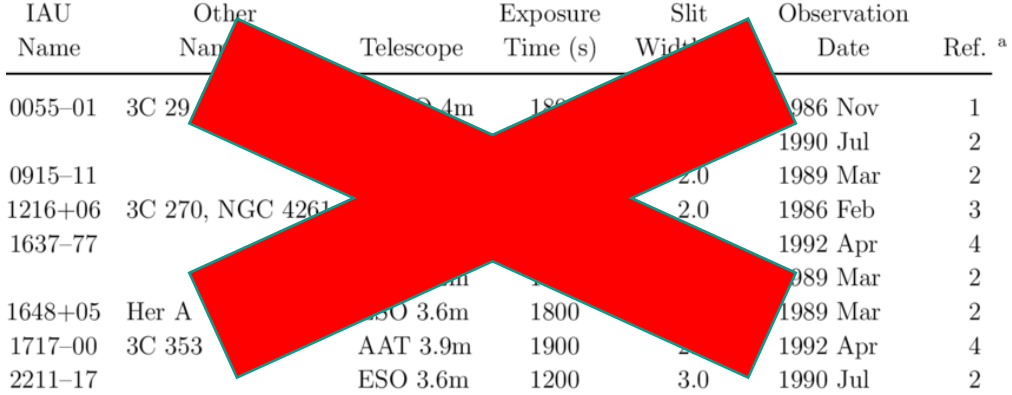
Article 30

« Dès lors que les données issues d'une activité de recherche financée au moins pour moitié par des dotations de l'Etat, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne ne sont pas protégées par un droit spécifique ou une réglementation particulière et qu'elles ont été rendues publiques par le chercheur, l'établissement ou l'organisme de recherche, leur réutilisation est libre. »

UTILISER LES ENTREPÔTS DE DONNÉES

« L'éditeur d'un écrit scientifique mentionné au I ne peut limiter la réutilisation des données de la recherche rendues publiques dans le cadre de sa publication. »

« Supplementary Data » chez les éditeurs

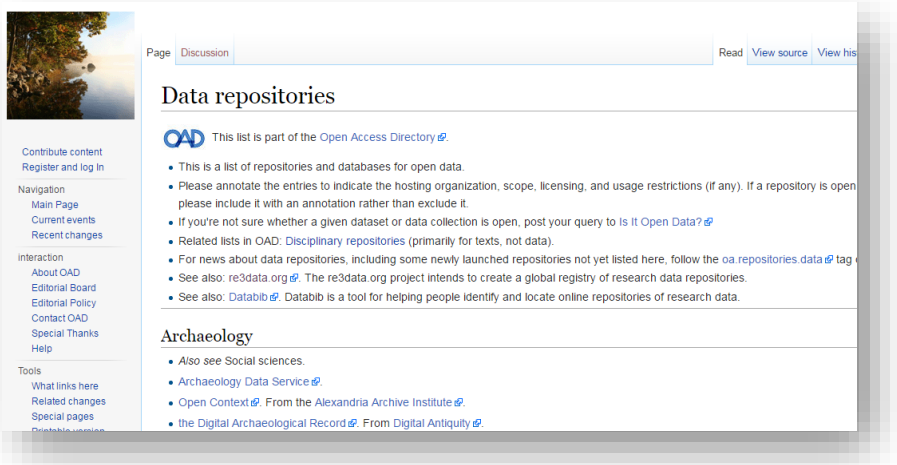


IAU Name	Other Nam	Telescope	Exposure Time (s)	Slit Widt	Observation Date	Ref. [a]
0055–01	3C 29	4m	18		986 Nov	1
					1990 Jul	2
0915–11				2.0	1989 Mar	2
1216+06	3C 270, NGC 4261			2.0	1986 Feb	3
1637–77					1992 Apr	4
		m			989 Mar	2
1648+05	Her A	SO 3.6m	1800		1989 Mar	2
1717–00	3C 353	AAT 3.9m	1900		1992 Apr	4
2211–17		ESO 3.6m	1200	3.0	1990 Jul	2

IDENTIFIER LE « BON » ENTREPÔT

Existe-t-il un entrepôt privilégié dans ma discipline ?

ENA
European Nucleotide Archive

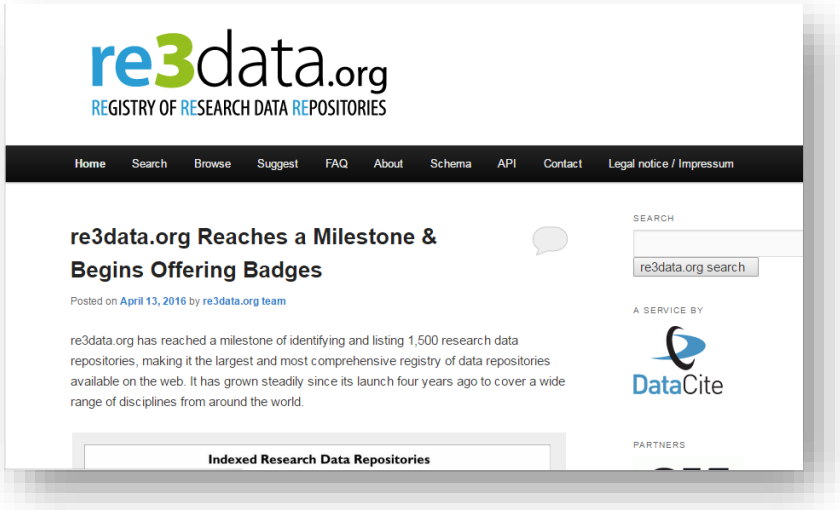


Ais-je accès à un entrepôt institutionel ?

INRAE

IRD Institut de Recherche pour le Développement FRANCE

cirad



zenodo

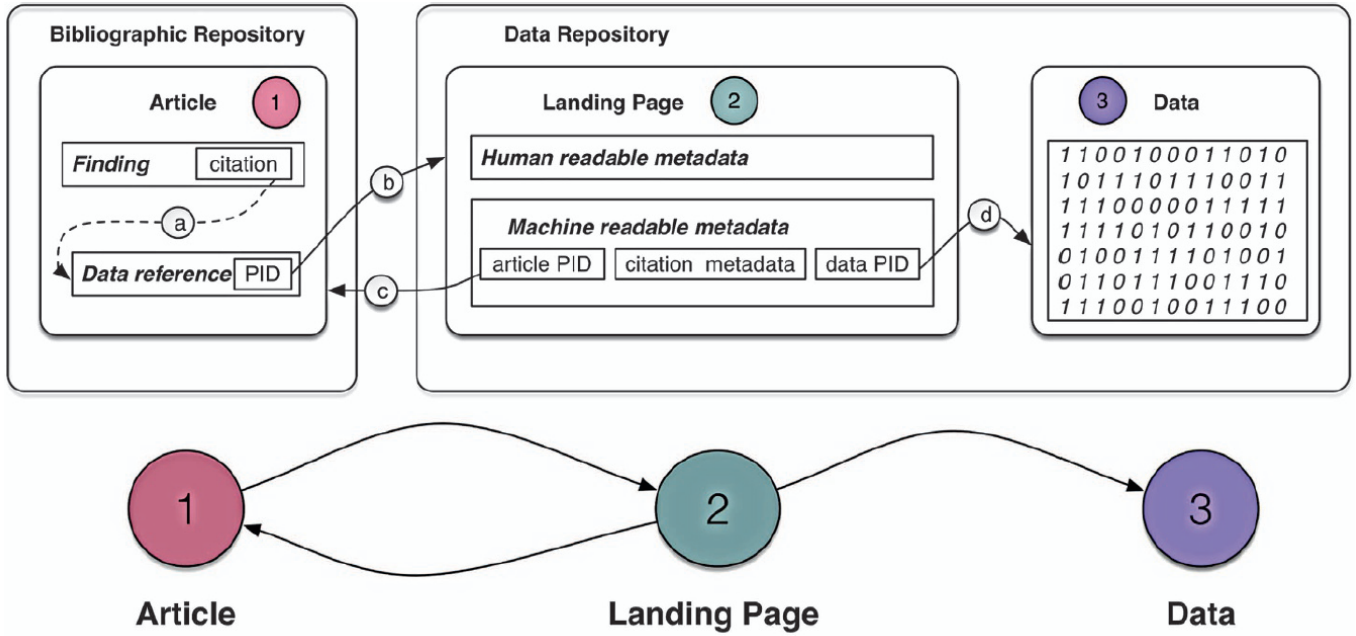
LIER LES DONNÉES AUX PUBLICATIONS

SCIENTIFIC DATA

OPEN

A data citation roadmap for scientific publishers

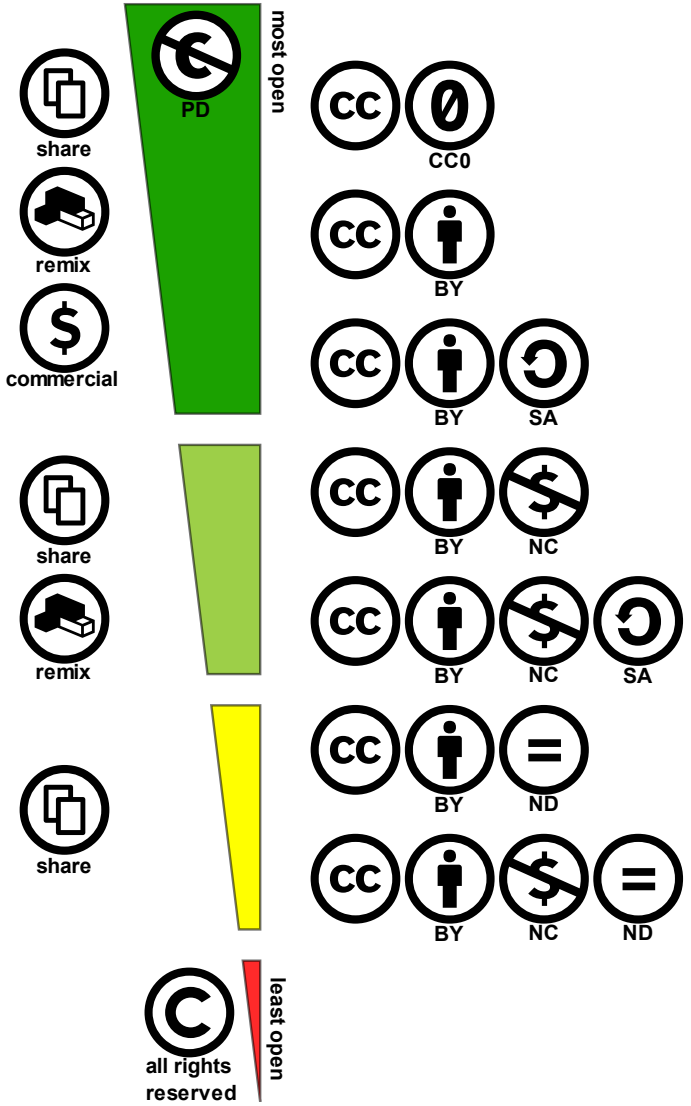
Helena Cousijn[1,*], Amye Kenall[2,*], Emma Ganley[3], Melissa Harrison[4], David Kernohan[5], Thomas Lemberger[6], Fiona Murphy[7], Patrick Polischuk[8], Simone Taylor[9], Maryann Martone[10] & Tim Clark[11,12]



ASSOCIER UNE LICENCE D'UTILISATION

LICENCE OUVERTE
OPEN LICENCE

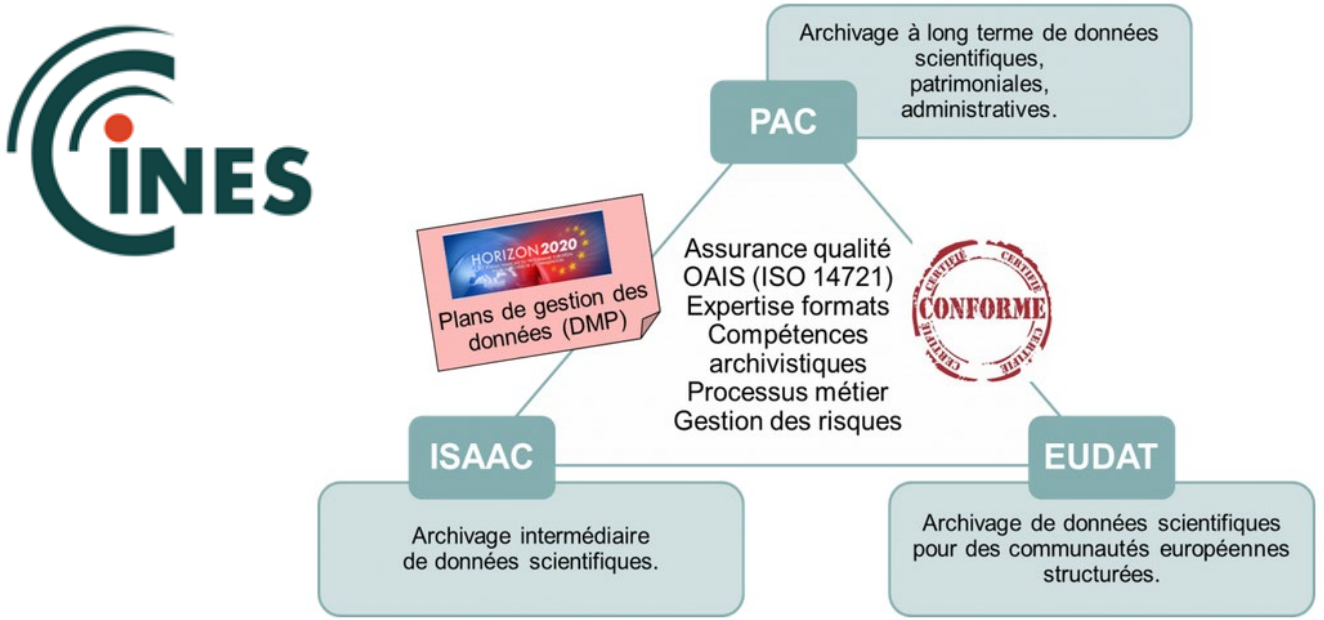
creative commons
BY



QUEL ARCHIVAGE POUR LES DONNÉES ?

Qu'est-ce que l'archivage et la conservation à long terme ?

CENTRE INFORMATIQUE NATIONAL
DE L'ENSEIGNEMENT SUPÉRIEUR



Est-ce bien raisonnable ?

UN SECOND BILAN
COMMENT NOUS SITUONS NOUS ?

SECOND AUTODIAGNOSTIC :

Sur une échelle de 1 à 5 étoiles, quel score (reflétant sa qualité) donneriez vous à votre stratégie actuelle de gestion des données de recherche ?

BON D'ACCORD ET MAINTENANT ?

PERSPECTIVES

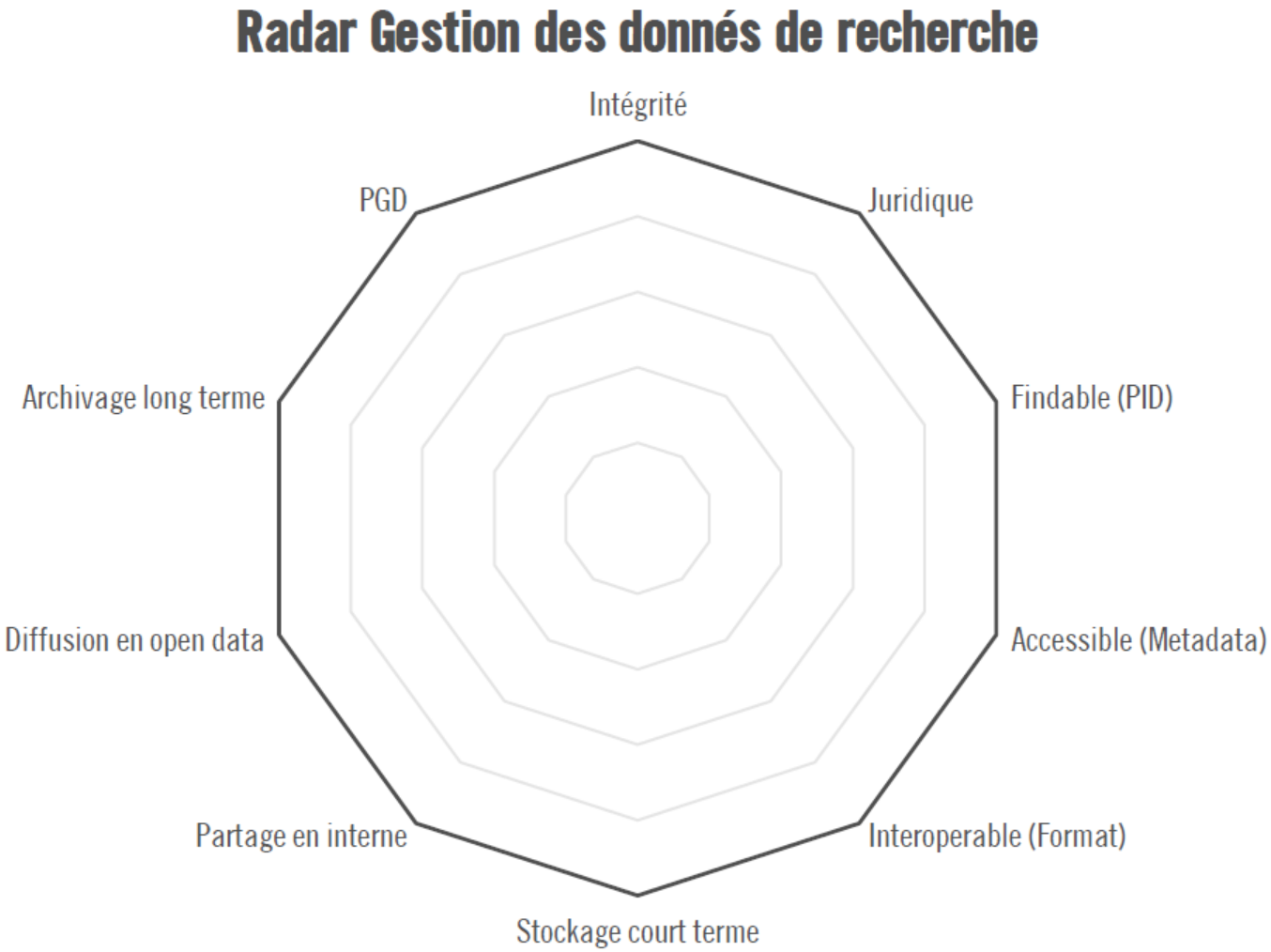
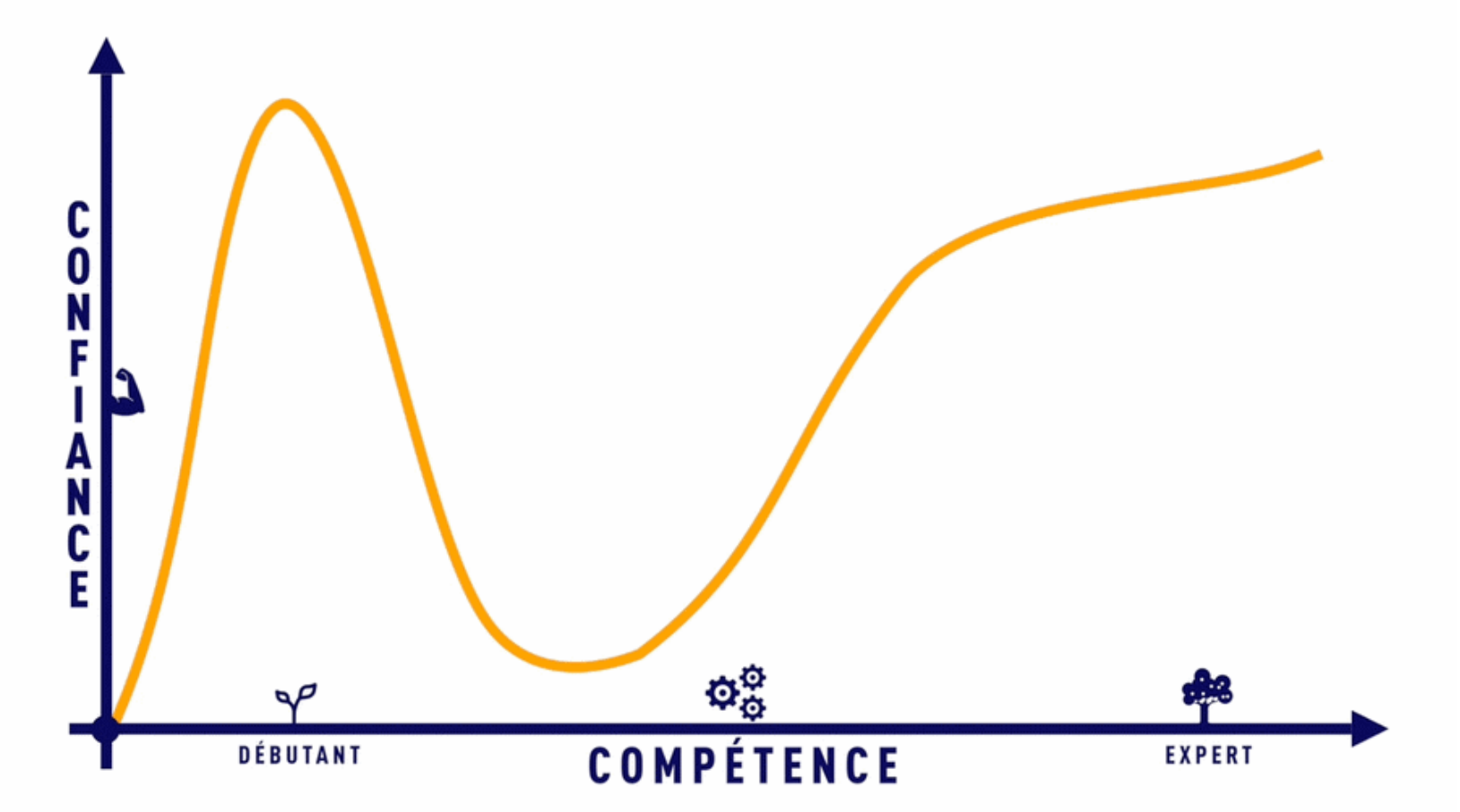


Figure de Germain Faity (2020)

PERSPECTIVES

BON D'ACCORD ET MAINTENANT ?

Représentation en graphique de dispersion XY des données illustrant l'auto-évaluation d'un sujet au cours de l'acquisition d'une compétence. Effet Dunning-Kruger (Wikipedia)



PERSPECTIVES

BON D'ACCORD ET MAINTENANT ?

- Pour le cas particulier du CBGP, une formation de trois jours qui **implique simultanément** porteur/porteuse de projet, personnel technique et doctorant/doctorante
- De nombreuses autres formations moins sur-mesure diffusées par les référents Données locaux et les directions de nos tutelles (DipSO + data officers des départements pour INRAE)
- De très nombreuses journées évènementielles dans le Monde entier et plus proche de nous à venir dès le retour d'une situation sanitaire compatible
- Des ressources d'autoformation très riches, dont celles du Comité Ouvrir la Science