Open science: comment l'inscrire dans ses pratiques de recherche ?

Mariannig Le Béchec
MCF en SIC – Université Lyon 1 - ELICO
co-responsable Urfist de Lyon

Bonjour!

Je suis Mariannig Le Béchec

Maitre de conférences en SIC

Co-responsable Urfist de Lyon

@marilebechec

Source templates : slidescarnival.com

Source inspiration création de cette communication : Millerand, 2012 ; Plantin, 2018

Autres formations : les plans de gestion de données – S. Cocaud et D. L'Hostis, INRA. Urfist Paris – 05 avril 2019 et Initiation aux données de la recherche _ Y. Lafosse, Inist. Urfist Paris – 28 mars 2019

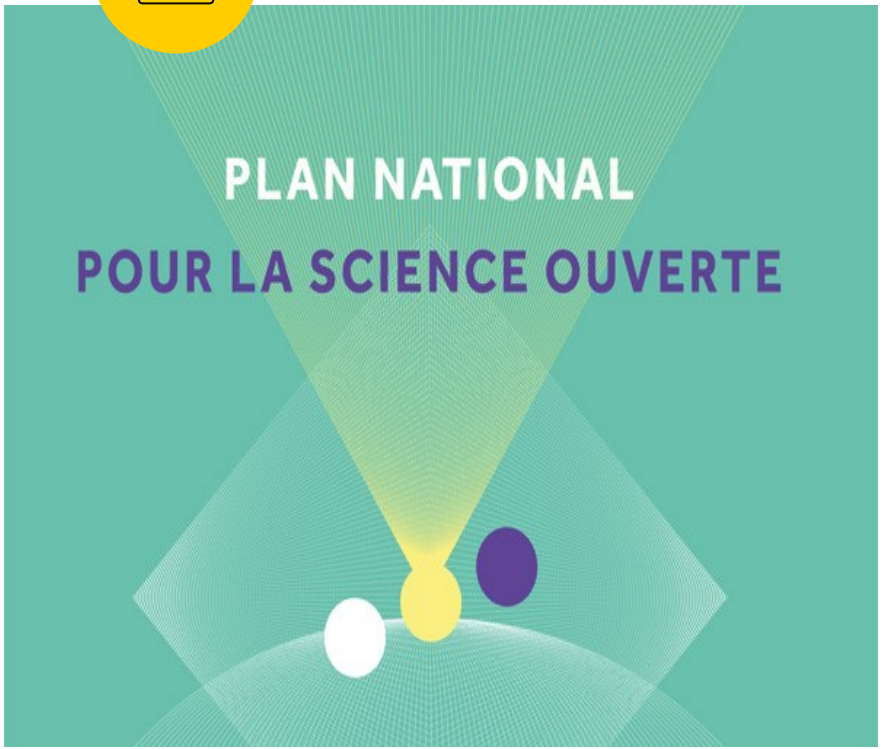
Publics : Personnels des établissements de l’enseignement supérieur et de la recherche

Activités : Formations

Programme et inscription
https://sygefor.reseau-urfist.fr/#/program/lyon

Pour trouver des informations sur le sujet

DORANum
Données de la Recherche : Apprentissage Numérique



En France, depuis 2018

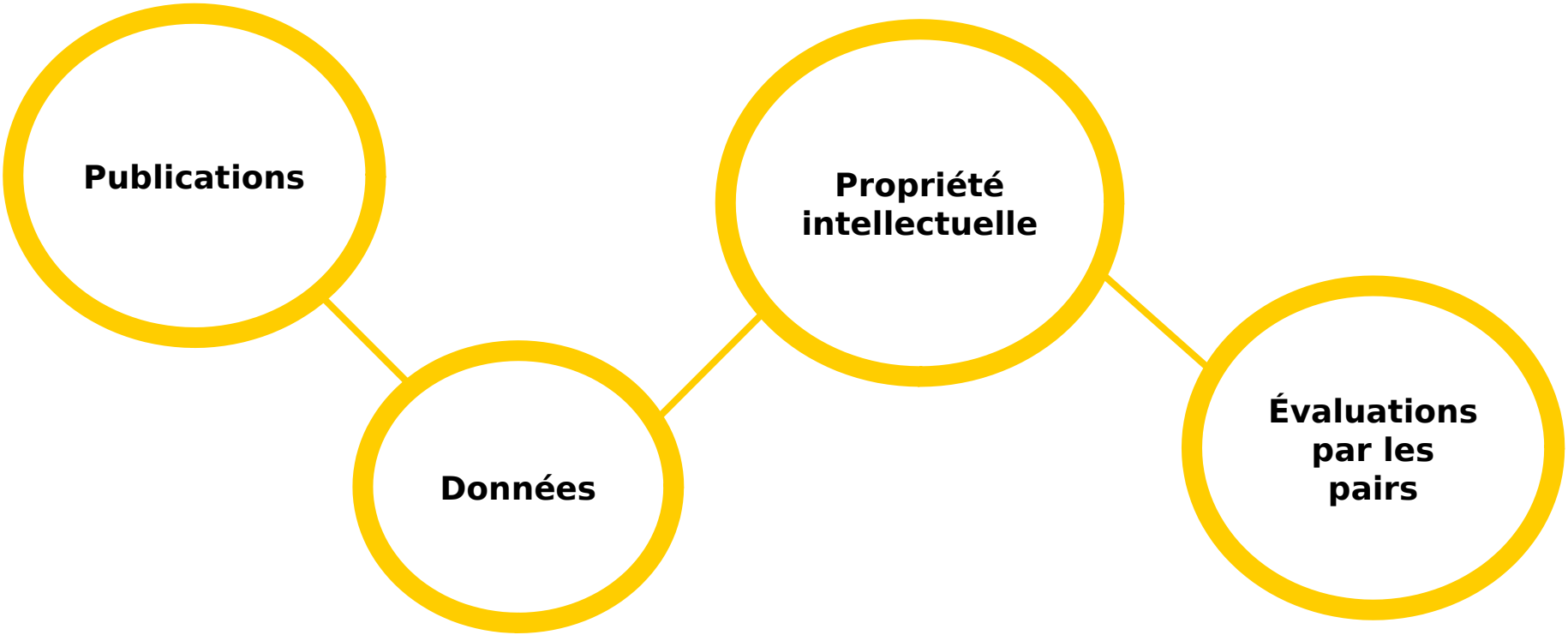
Axes

1. généraliser l’accès ouvert aux publications

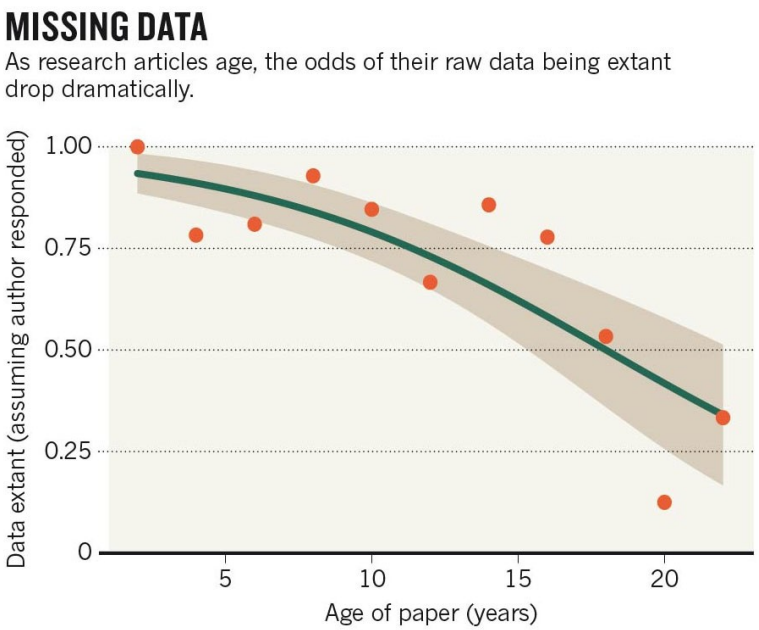
2. structurer et ouvrir les données de la recherche

3. s’inscrire dans une dynamique durable, européenne et internationale

Science ouverte



Scientists losing data at a rapid rate



Elizabeth Gibney & Richard Van Noorden (Nature, 19/12/2013)

Decline can mean 80% of data are unavailable after 20 years.

Vines, T. H. et al., Current Biology, http://dx.doi.org/10.1016/j.cub.2013.11.014 (2013).

1

Data, mais lesquelles ?

Différentes définitions

What is data ?

Data, OCDE, 2007

des 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

Data

are any information or observations that are associated with a particular project, including experimental specimens, technologies, and products related to the inquiry.

Comptes-rendus, carnets de terrains

Cahiers de laboratoire

Bases de données

Script, algorithmes

Méthodes

(protocoles, plan d’expérimentation…)

Échantillons

Questionnaires, enquêtes

Fichiers (textes, audio, vidéo, images)

« Raw data »
&
Invisible work

2

Data, mais qui les produit ?

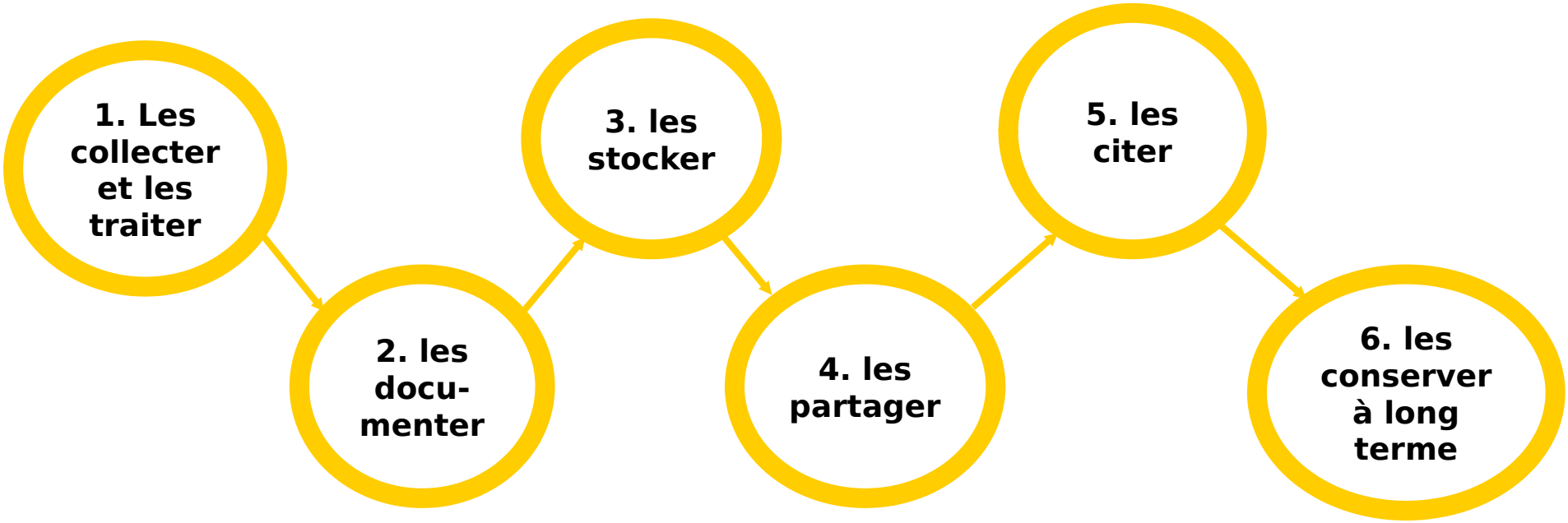
Différents acteurs

Petite histoire

Acteurs

Marie Chercheuse	Valentin Chargé de projet de recherche	Inès Informaticienne	Lana Professionnelle de l'Information
			

6 étapes
pour gérer les données



1/6 comment les collecter Et les traiter ?

- 2 situations
  - Les données sont disponibles : veille
  - Les données n’existent pas: co-création de formulaire, transfert dans la base de données

2/6 Comment les documenter ?

- Expliquer comment et pourquoi le chercheur a créé les données
- Document : txt, pdf ou vidéo
- Contenu : hypothèse, méthodes, échantillonnage, traitement éventuels (logiciels, algorithmes utilisés) et instruments



3/6. Comment les stocker ?

- Déterminer les lieux et supports de stockage selon le volume et la fréquence de consultation (coût)
- Qualifier les données sensibles et leur niveau de protection
- Déterminer la durée de stockage pour éliminer les fichiers temporaires et alimenter la conservation
- Anticiper le partage en hiérarchisant le contenu, les états, les types, etc.
- Programmer les sauvegardes



4/6. comment les partager ?

- As open as possible, as closed as necessary
- Optimiser l'impact de la diffusion des données : DataPaper ou article + jeu(x) de données
- Tenir compte du cadre légal et contractuel (PI, copyright, licences de réutilisation)
- Via des infrastructures de données thématiques (stockage, accès aux données, interface de consultation, de création de communautés de chercheurs)
- Attention aux standards

OpenAIRE

BioMedBridges

4/6. comment les partager ?

- ex. projet RELOC (2016-2017)
- Accès pendant le projet
- Accès après le projet → quelle diffusion en respectant les clauses contractuelles ?
- Comment signifier les enrichissements ?
- Quelle anonymisation ?

ex. choix des formats

Type	Recommended	Avoid for data sharing
Tabular data	CSV, TSV, SPSS portable	Excel
Text	Plain text, HTML, RTF PDF/A only if layout matters	Word
Media	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	Quicktime H264
Images	TIFF, JPEG2000, PNG	GIF, JPG
Structured data	XML, RDF	RDBMS

Service de validation des formats : https://facile.cines.fr/
Further examples: http://www.data-archive.ac.uk/create-manage/format/formats-table

Comment les faire citer ?

Identifiant unique

Interopérable

Persistant

= DOI

digital object Identifier

DC : identifiants Dublin Core (15 éléments)

http://datadryad.org/resource/doi:10.5061/dryad.77r3p?show=full

DRYAD About For researchers For organizations

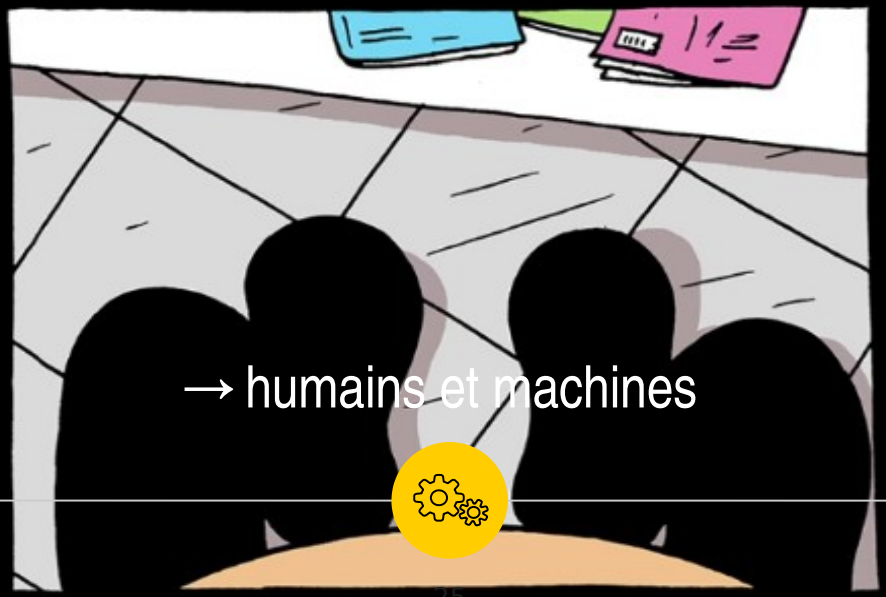
Show simple item record

dc.contributor.author	Dubuc Messier, Gabrielle
dc.contributor.author	Garant, Dany
dc.contributor.author	Bergeron, Patrick
dc.contributor.author	Réale, Denis
dc.date.accessioned	2012-08-17T19:33:27Z
dc.date.available	2012-08-17T19:33:27Z
dc.date.issued	2012-09-27
dc.identifier	doi:10.5061/dryad.77r3p
dc.identifier.citation	Dubuc Messier G, Garant D, Bergeron P, Réale D (2012) Environmental conditions affect spatial genetic structures and dispersal patterns in a solitary rodent. Molecular Ecology 21(21): 5363–5373.
dc.identifier.uri	http://hdl.handle.net/10255/dryad.41672
dc.description	The study of the spatial distribution of relatives in a population under contrasted environmental conditions provides critical insights into the flexibility of dispersal behaviour and the role of environmental conditions in shaping population relatedness and social structure. Yet

EXIF

Format

Heure



→ humains et machines

Global Positioning System	
GPS Altitude	31.9 m
GPS Latitude	6deg 14' 7.620"
GPS Longitude	106deg 49' 30.210"
Image Information	
Date and Time	2018:08:24 15:47:27
Manufacturer	Apple
Model	iPhone 6s
Photograph Information	
Aperture	F2.2
Exposure Bias	0 EV
Exposure Mode	Auto
Exposure Program	Auto
Exposure Time	1/874 s
Flash	No, auto
FNumber	F2.2
Focal Length	4.2 mm
ISO Speed Ratings	25
Metering Mode	Multi-segment
Shutter speed	1/874 s
White Balance	Auto

Le Monde

Martin Vindberg, Lemonde

6/6 comment les conserver sur le long terme ?

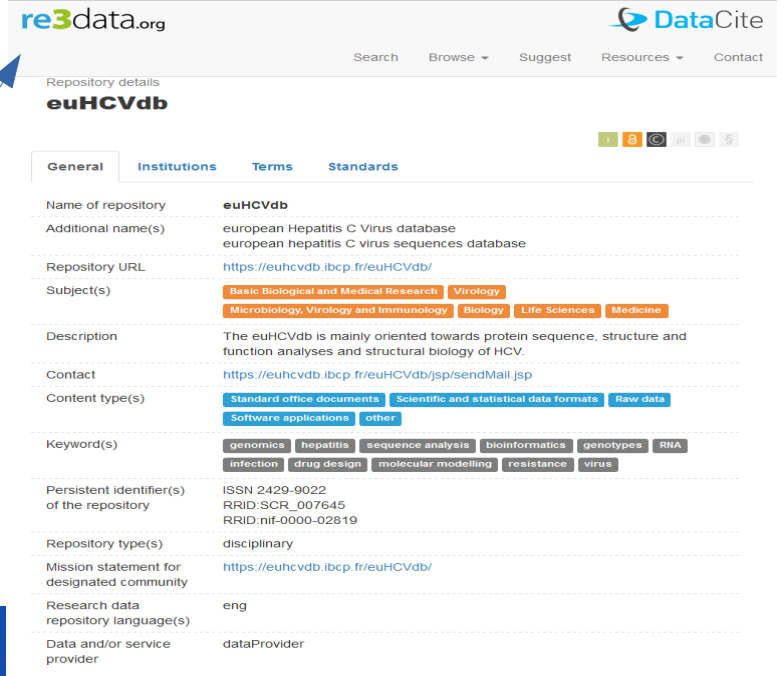
- Entrepôts de données
- 1 répertoire d'entrepôts
- Ex d'entrepôts :

  INRA Dataverse,

  Irsteadata

  Figshare

  Nakala



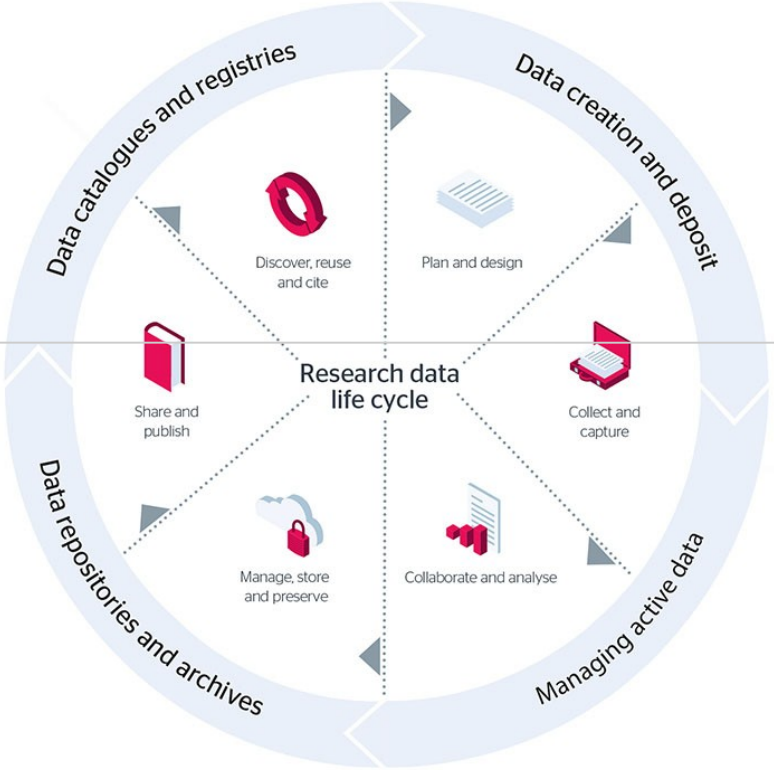
Acteurs

- Chercheurs
- Responsable du projet
- Informaticien
- Correspondant administrateur des données du Ministère
- Professionnels de l’information (conservateur, bibliothécaire, archiviste…)
- Juriste,
- Data curator, etc.

3

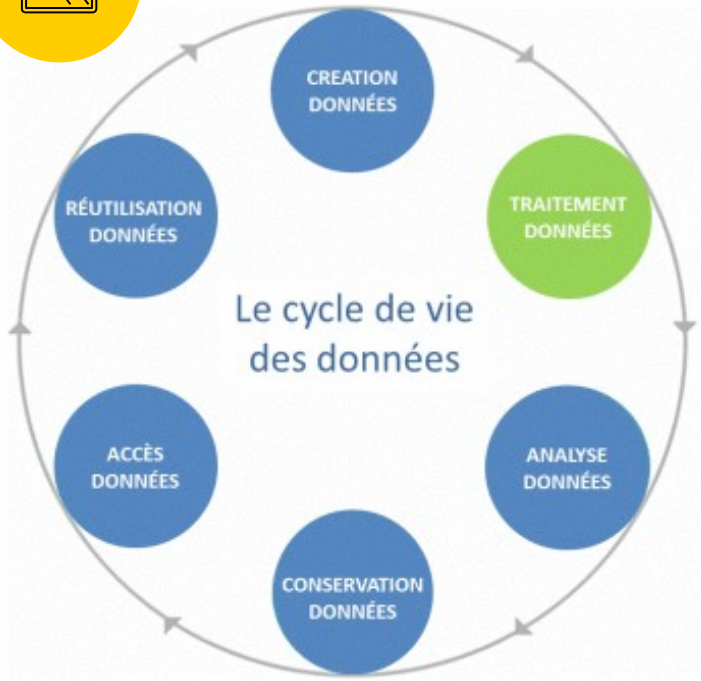
Data, mais quel management ?

Un ou des cycles



le management de données ?

Image from JISC under CC-BY-NC-ND, Caroline Ingram



Cycle de vie des données de la recherche (adapté de UK Data Archive)

https://www.ukdataservice.ac.uk/manage-data/lifecycle

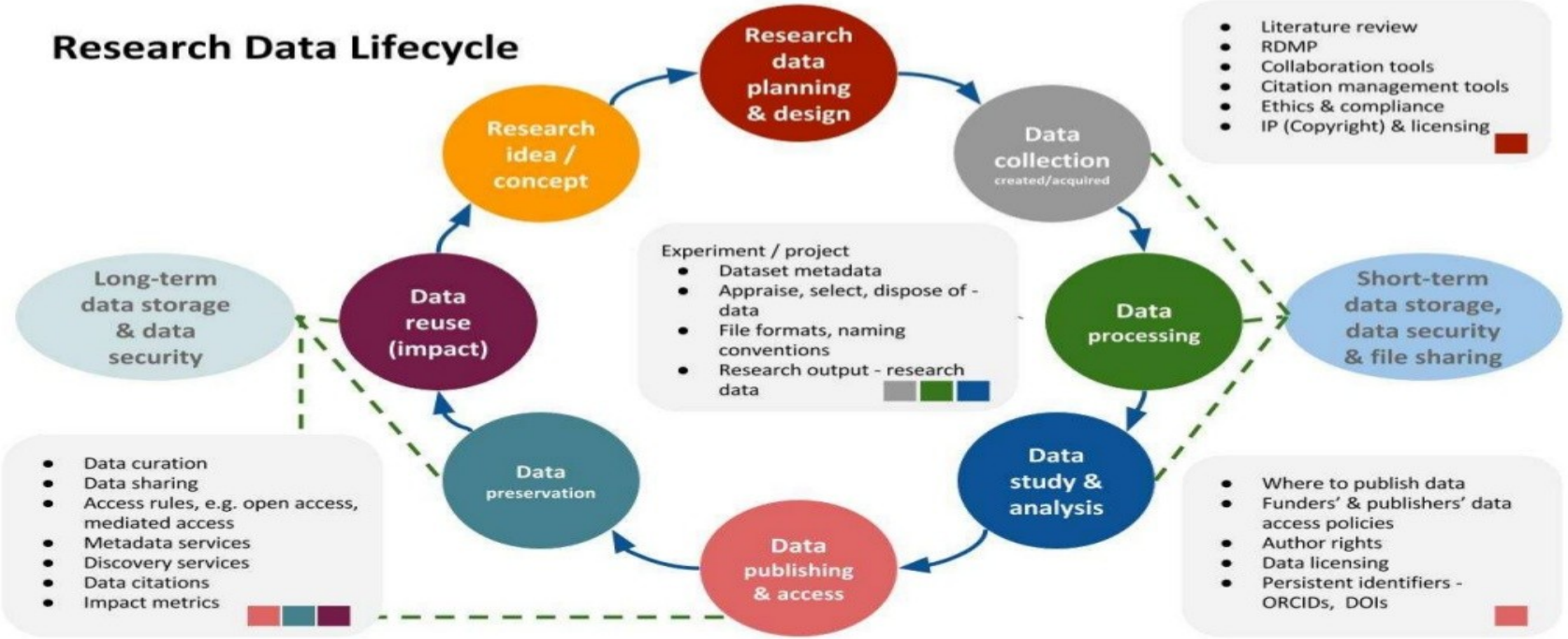
DMP

Rédigé au début d'un projet de recherche et qui définit ce que les chercheurs feront de leurs données pendant et après le projet en explicitant la mise à disposition des données

Contient :

- Cycle de vies des données
- Descriptif des données
- Politique des données
- Coûts

Autre définition



Rscd 2017 bo f data lifecycle data skills for libs, 21 mars 2017

Mise en place de bonnes pratiques de gestion des données

Photo by Quino Al on Unsplash

étape 7. et après

- formaliser la gestion de son travail sur les données (gain de temps)
- Augmenter l’impact de sa publication (Piwowar & Vision en 2013, même si corpus limité de 26 articles)
- Assurer la réutilisation par des tiers des données
- Exiger par les financeurs

Did you find the process of writing / supporting a H2020 DMP positive or negative?
[189 respondents, 119 comments. Answers: "Positive" 60%, "Negative" 16%, "Not applicable" 24%]

Overall experience with writing/supporting a H2020 DMP

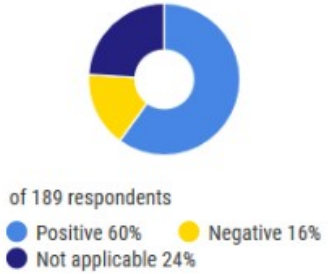


Figure 003 - Q3: reaction to H2020 DMP process

OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans
January 9, 2018 Dataset Open Access

Marjan Grootveld; Ellen Leenarts; Sarah Jones; Emilie Hermans; Eliane Fankhauser
https://zenodo.org/record/1120245

"I cannot say it was negative or positive. It stimulated interesting discussions among the leading scientists, who had very different view what a data management plan was about in the first place. Some intended to write legacy, but could be convinced that this was not about developing new data policy, because the project only re-uses data from many different sources. This specialty was also the reason that the due date of a first data management plan after 6 months was felt far too early for this particular project. At this stage, it was not sufficiently settled what data sources would be used and how the data should be handled, stored, and made accessible to external users. Therefore, to some extent, there was unnecessary effort with the risk of writing or agreeing things that would become obsolete very quickly. Because of "theoretical" base in the beginning, some scientists were tempted to write more a proposal than data management plan."

Management

Coordonner

Différents métiers pour une même finalité et compétences

Concilier des points de vue (ex. durée de conservation des données, embargo)

Réguler les accès

Maîtriser les coûts

Humains

Matériels

Documenter les connaissances

Produire des contenus réutilisables

Et des procédures pérennes

Éthique et légal

Consentement par rapport à ce partage

Confidentialité et anonymisation

Choix des licences

4

Data, quels enjeux ?

Photo by jesse orrico on Unsplash

Un enjeu éthique et économique

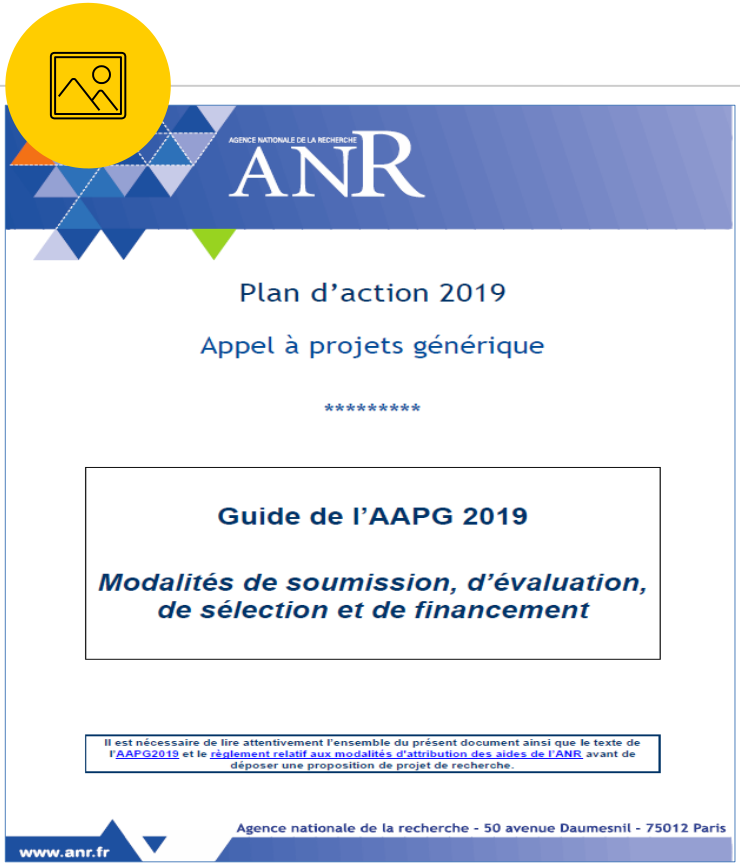
Science ouverte, five open schools of thought (Fecher, Friesike, 2014)

[Article 29.3 Modèle de convention de subvention]

Déposer dans un entrepôt pour l'accès, l'exploration, l'exploitation, la reproduction et la diffusion par un tiers (gratuitement)
-données (et métadonnées) nécessaires à la validation des résultats présentés dans les publications scientifiques ;
- données (et métadonnées) selon délais stipulés dans le PGD

Fournir des infos sur outils et instruments à disposition des bénéficiaires et nécessaires à la validation des résultats

En France, Agence Nationale pour la recherche (ANR)

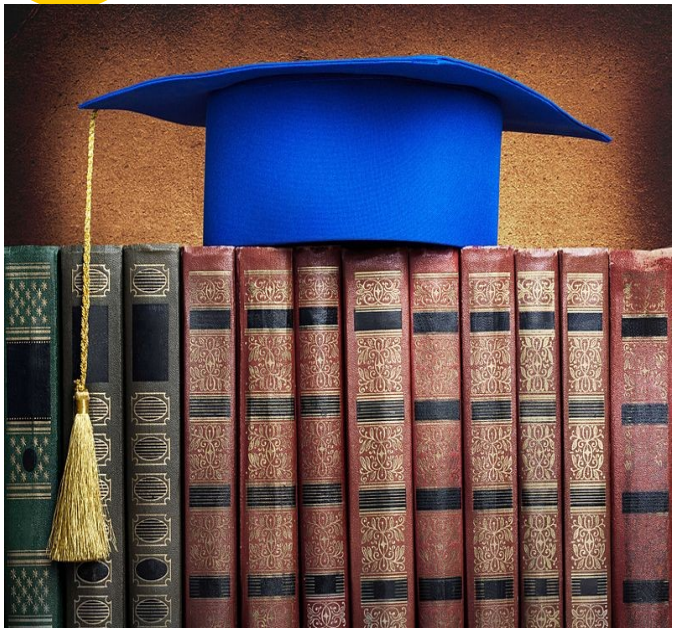


PGD (DMP) obligatoire au démarrage du projet financé en 2019

Respect des obligations loi République numérique et plan national en faveur des archives ouvertes

→ dépôt des publications scientifiques (texte intégral) soit dans HAL soit dans une archive institutionnelle locale

→ fournir un PGD (DMP)



Source : A.lakrafi

Arrêté du 22 février 2019

définissant les **compétences des diplômés du doctorat** et inscrivant le doctorat au répertoire national de la certification professionnelle

Bloc 3 Valorisation et transfert des résultats d'une démarche R&D, d'études et prospective

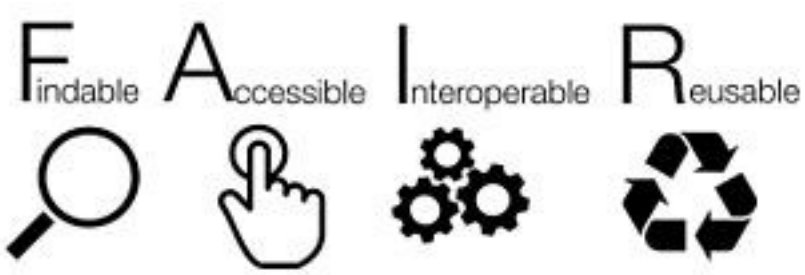
« … mobiliser les techniques de communication de données en « open data » pour valoriser des démarches et résultats. »

Des données FAIR

→ humains et machines

Photo by Quino Al on Unsplash

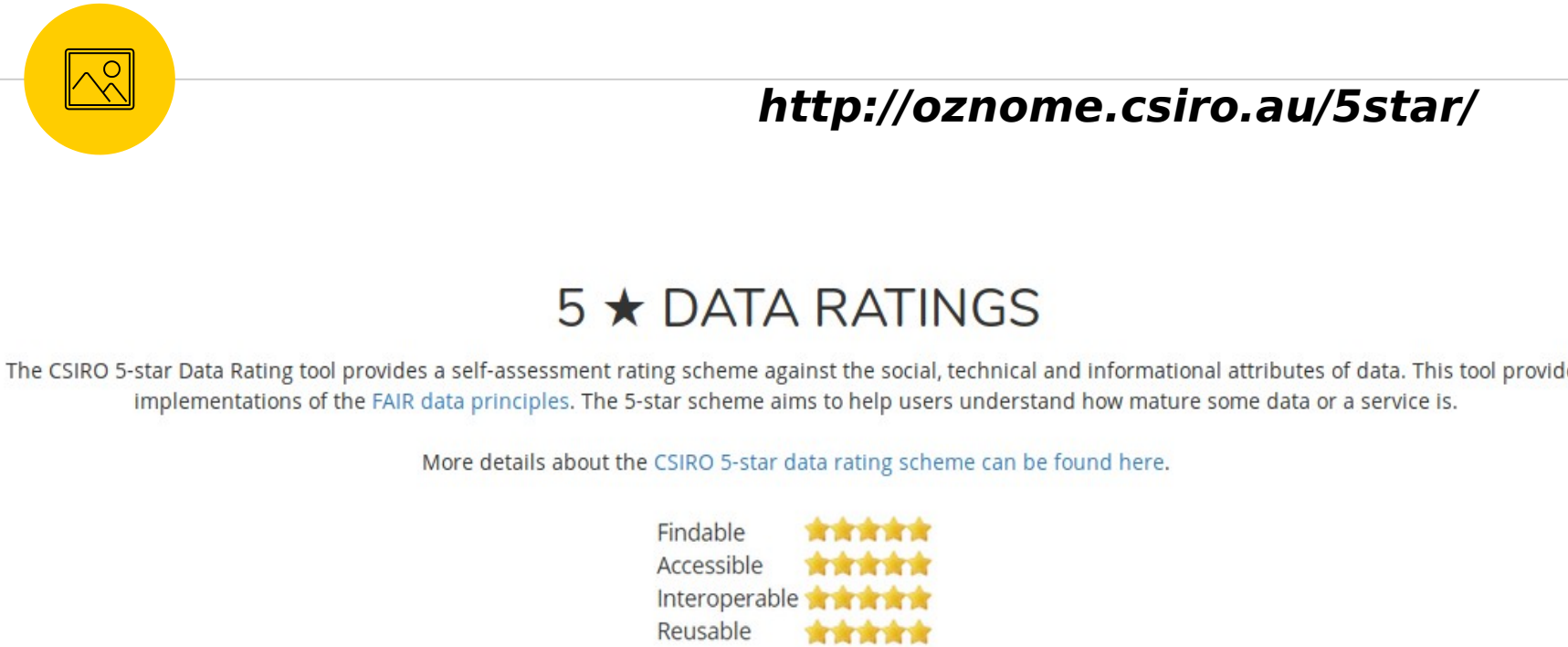
Plan national pour la science ouverte MESRI (juillet 2018)



SangyaPundir

« [...] les données produites par la recherche publique française soient progressivement structurées en conformité avec les principes FAIR (Findable -Facile à trouver-, Accessible – Accessible-, Interoperable – Interopérable-, Reusable - Réutilisable-), préservées et, quand cela est possible, ouvertes. »

http://oznome.csiro.au/5star/

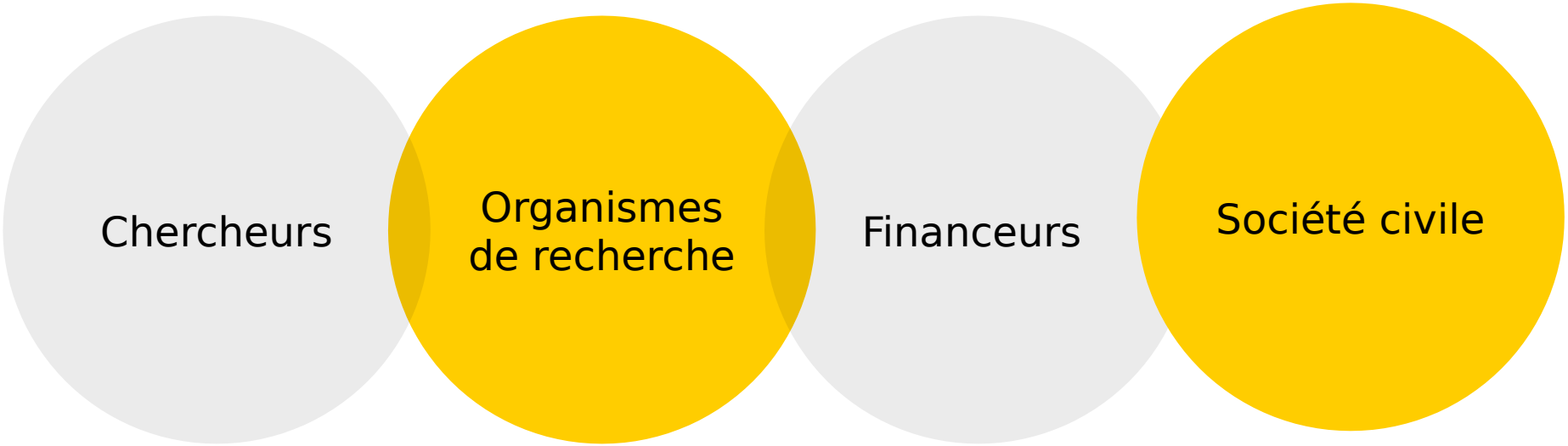


RDA

Pour une automatisation de la lecture et de l’écriture d’information de ou dans un DMP

Interopérabilité pour toutes disciplines

Intérêts


Chercheurs
Organismes de recherche
Financeurs
Société civile

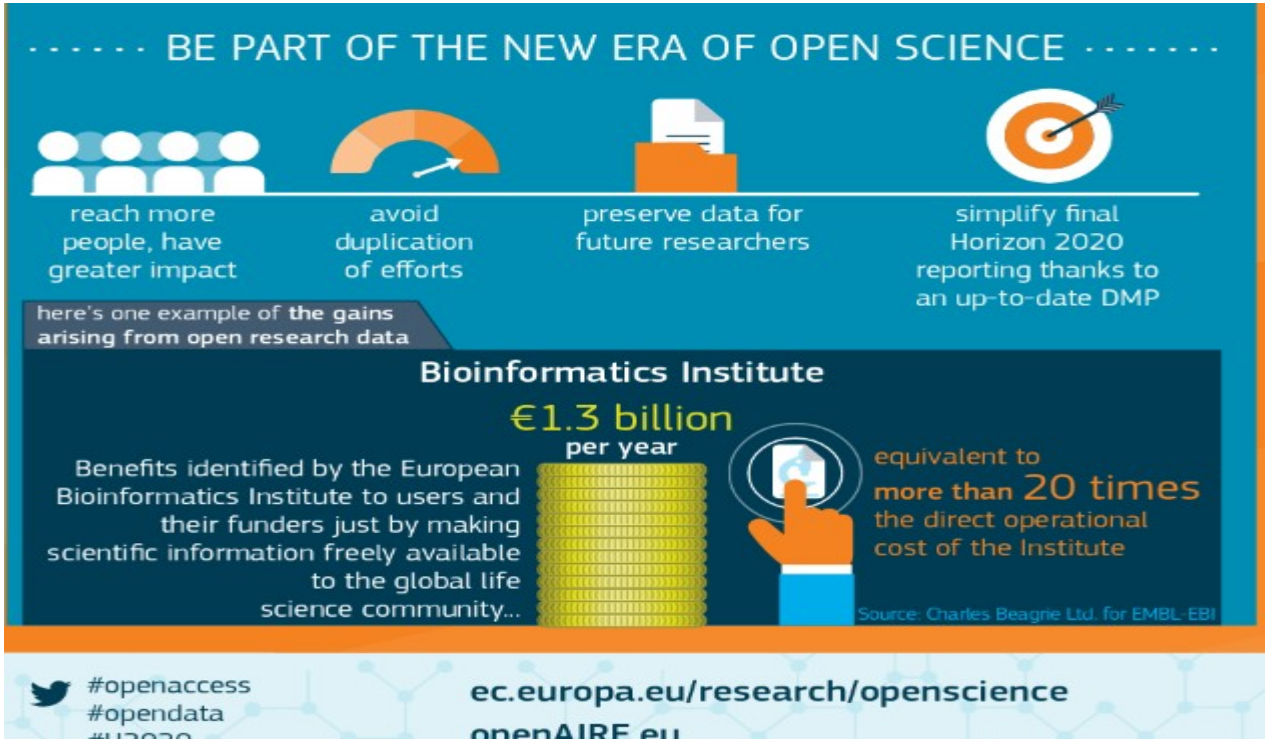
Les missions des personnels de la recherche publique

Elles comprennent:

- le développement des connaissances,
- leur transfert et leur application dans les entreprises et dans tous les domaines contribuant au progrès de la société,
- la diffusion de l'information et de la culture scientifique et technique dans toute la population et notamment parmi les jeunes,
- la participation à la formation initiale et à la formation continue,
- l'administration de la recherche,
- l'expertise scientifique.

Code de la recherche (Art. L411-1)

Innovation, un leitmotiv ?



RESEARCH DATA - OPEN BY DEFAULT

FAIR DATA!

Accessible

Interoperable

Findable

Re-usable

HORIZON 2020 GRANTEES ARE REQUIRED

take measures to ensure open access to the **data underlying their scientific publications**

provide open access to **any other research data of their choice**

Horizon 2020 grantees are **encouraged to also share datasets** beyond publication

PROJECTS MUST HAVE

DATA MANAGEMENT PLAN (DMP)

Provides information on:

- the data the research will generate
- how to ensure its curation, preservation and sustainability
- what parts of that data will be open (and how)

Data management costs are fully eligible for funding

No repository imposed: deposit data where you want

AS OPEN AS POSSIBLE, AS CLOSED AS NECESSARY

Grantees have the right to **opt-out**, but need to say **why**

Top three reasons for **opt-out**:

- **intellectual property rights**
- **privacy**
- **might jeopardise project's main objective**

The approach has been tested during a Horizon 2020 pilot action

2015

of 431 signed projects

65.4%

opted to share data

from 2017

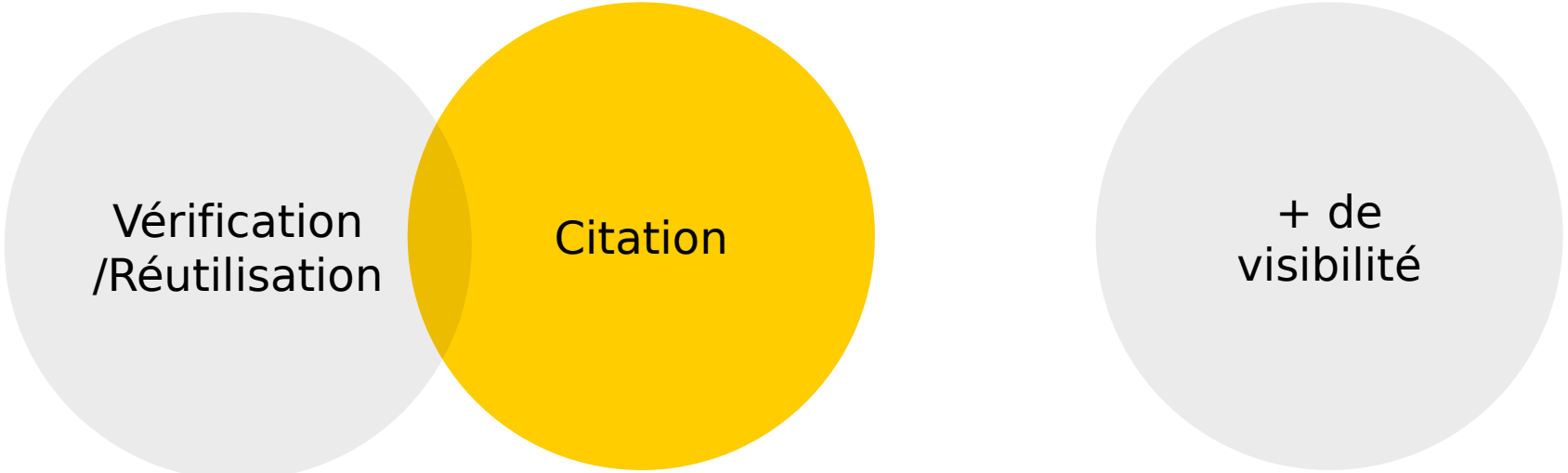
the current Open Research Data Pilot expands to cover all areas of Horizon 2020, with the same rules

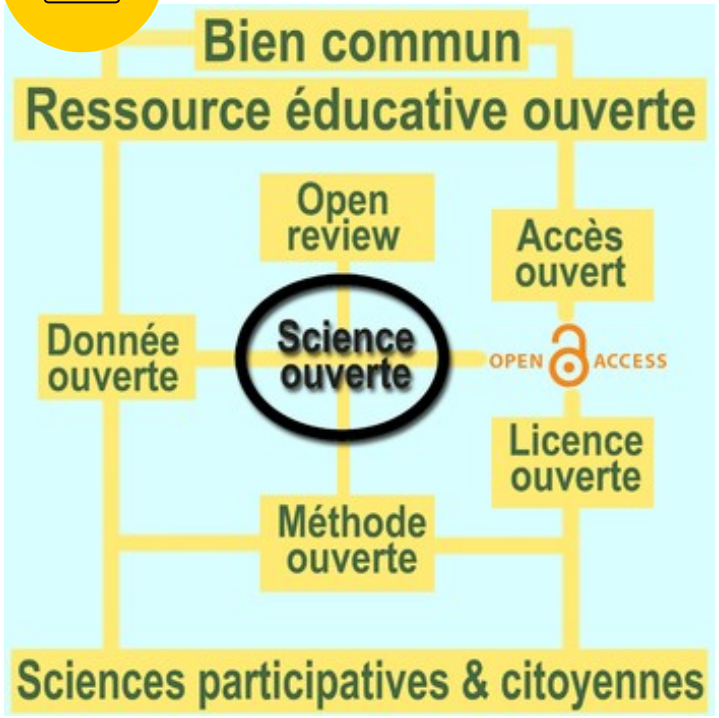
HOW IT WORKS

1. Get funding
2. Write DMP
3. Gather data during research
4. Finish research

- Publish results
- Choose repository and deposit data
- **Inform OpenAIRE** the EU-funded Open Access portal

Objectifs





Open science

is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods.

https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition

4 Conclusion

Rendre visible un travail invisible et collectif

- Les gestionnaires d’information « invisibles » (Millerand, 2012)
- Visibility and invisibility of data processors, data curators (Plantin, 2018) → « pristine »
- Place des chercheurs dans les infrastructures ?