Ecole technique InnObs - Montpellier 18-22 novembre 2019

Restitution de la séquence introductive sur la gestion des données dans le cadre de l'OPEN DATA

Mardi 19 novembre 2019 10h30-11h15	Données [45'] Mise à niveau	En plénière

« Séquence introduction et mise à niveau sur la gestion des données dans le cadre de l'OPEN DATA »

Durée: 45 mn

Objectif général de la séquence : « Ouverture des données de la recherche, comment l'intégrer dans vos pratiques au cours de vos projets innovants? »

Une séquence de mise à niveau des participants sur les enjeux institutionnels de l'ouverture des données de la recherche dans le cadre de la science ouverte. A cette occasion, il est fait un focus sur les démarches à entreprendre, au sein des projets innovants dans l'observation des évènements biologiques saisonniers, pour atteindre les standards de données permettant leur ouverture.

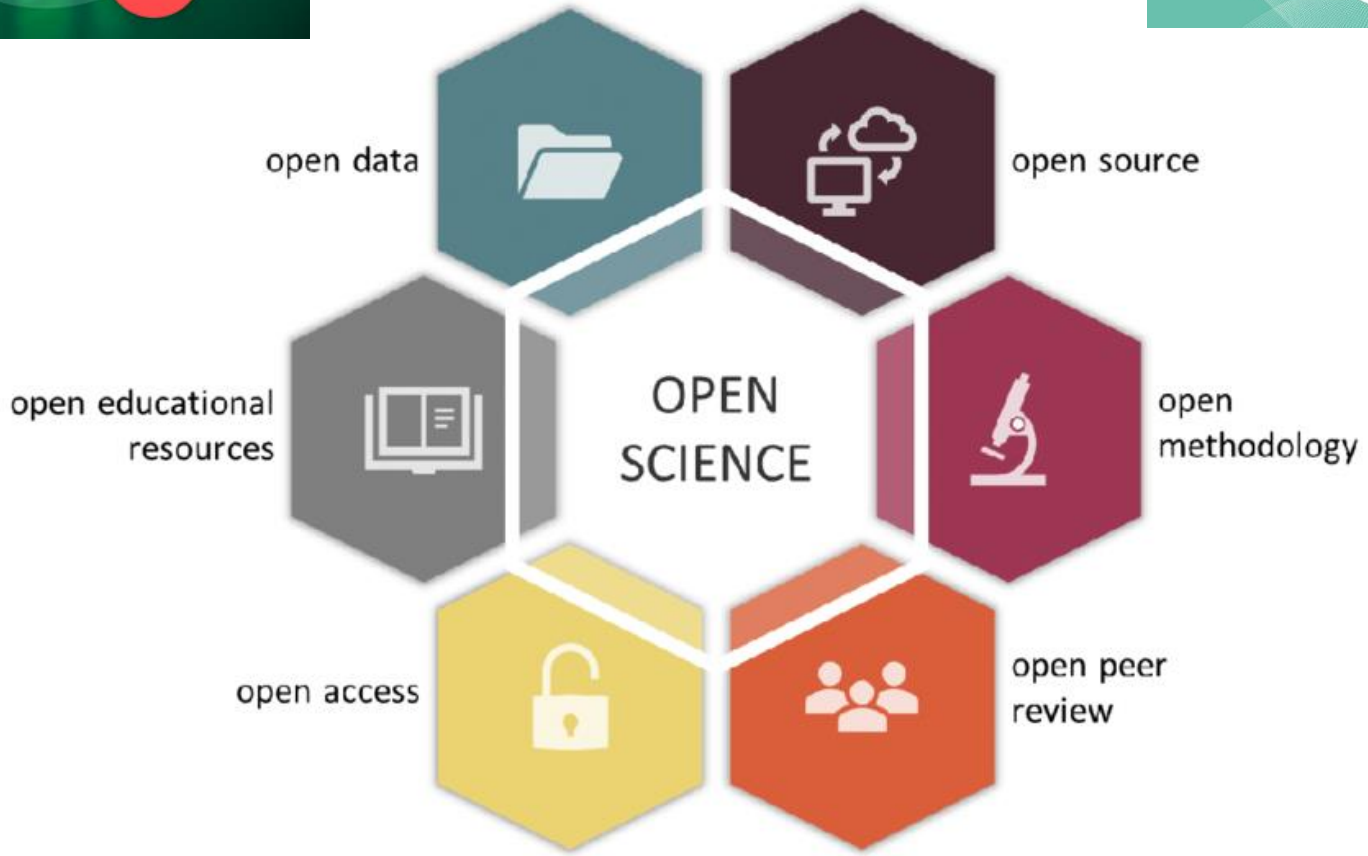
La forme: ludique sous forme d'une interview entre un «spécialiste» (Christian Pichot) qui a une vision globale sur les enjeux des données de la recherche, et un «béotien» (Laurent Burnel) :

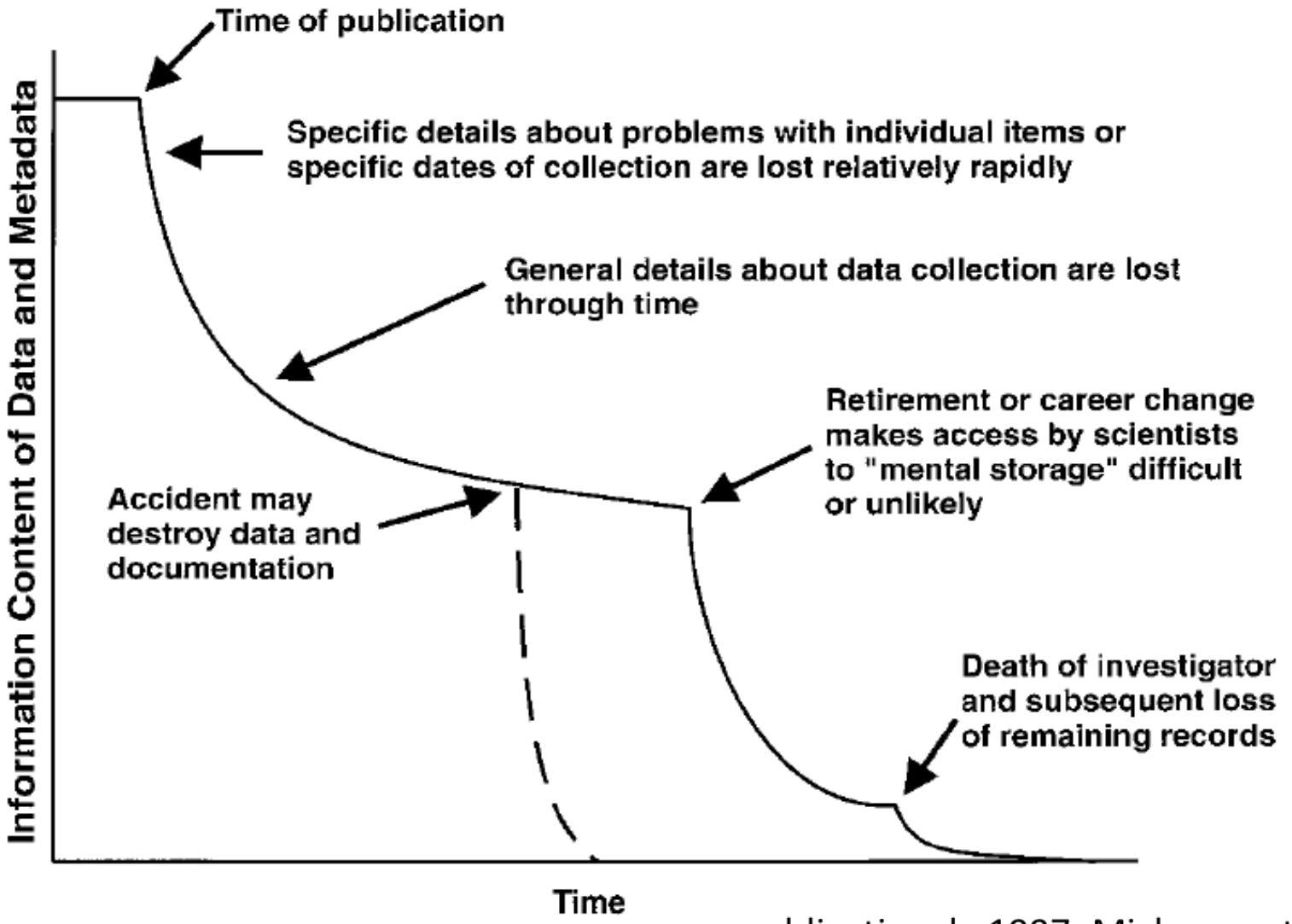
Il se présente sous forme d'une série de questions posées par le béotien, le spécialiste répondant de suite à chaque question, éventuellement à l'aide d'un support si nécessaire (support utile au moins pour la présentation des principes FAIR et pour le cycle de vie de la données)

Ce support est une tentative de restitution des échanges réalisés entre Christian Pichot et Laurent Burnel.
Les questions et commentaires sont à la suite des diapositives concernées pas ces questions et commentaires

Vous avez dit ouverture des données?

Contexte





publication de 1997: Michener et al.

Commentaires diapo n°3 et 4

Question 1: on va parler pendant cette semaine de gestion de données dans ce nouveau contexte du partage des données mais Open science et données de la recherché quèsaco? tu pourrais d'abord nous faire un petit historique de tout cela ?
= Un bref historique? Enjeux scientifiques, éthiques et économiques?

Réponse 1: Oui, assez brièvement et en se focalisant sur les données et non pas sur les publications.
La description des étapes 'marquantes' depuis les années 90, fait l'objet de quelques synthèses dont
https://www.ouvrirlascience.fr/un-historique-du-libre-acces-aux-publications-scientifiques-et-aux-donnees/

Le processus d'acquisition des connaissances….et de leur exploitation

Jusqu'à présent, le partage des résultats de la recherche se faisait très majoritairement au travers des publications scientifiques et techniques, et très peu via les données. Les situations sont néanmoins assez différentes selon les domaines de recherche en fonction d'une part des contraintes d'acquisition des données (ex : astronomie) et d'autre part du caractère ± récent des recherches (ex génomique, données de séquences…) et de leur concomitance avec le développement du numérique.

Le constat fait est celui d'une perte assez rapide des données. Dans le domaine de l'écologie où l'hétérogénéité des objets d'étude et des pratiques est grande, ce processus a notamment été décrit dans une publication de 1997 (Michener et al.). Dans un premier temps, les informations les plus spécifiques sur le contexte dans lequel ont été réalisés les travaux sont oubliées puis, au départ des protagonistes (chercheurs, techniciens) soit vers de nouvelles activités soit en retraite seul 50 % de l'information (données/métadonnées) est disponible et au décès de ceux-ci il n'en resterait qu'environ 10 %.

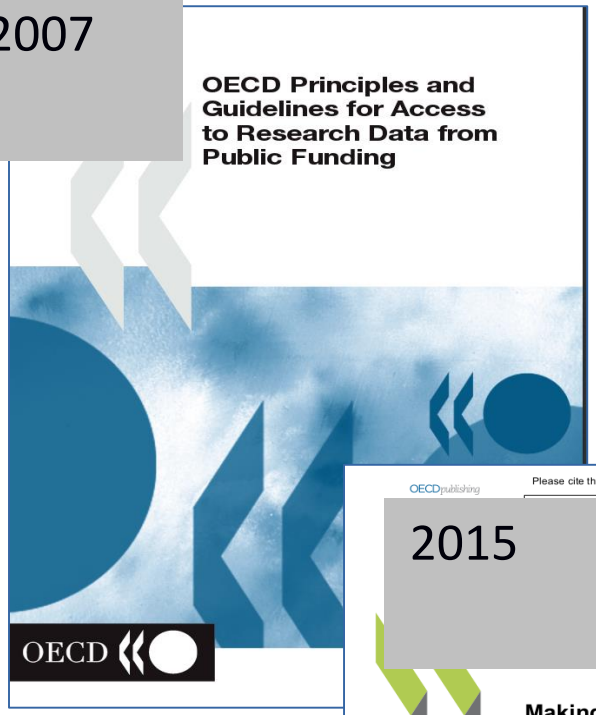
En 2011 Reichman et al. estimaient que moins de 1% des données acquises en écologie étaient accessibles après publication des résultats de la recherche

C'est un constat généralisé, fait notamment par les agences de financements qui ont souhaité disposer de plus de lisibilité sur les données acquises, pouvoir vérifier l'effectivité des travaux, et rendre possible la réutilisation des données pour la recherche ET l'économie.

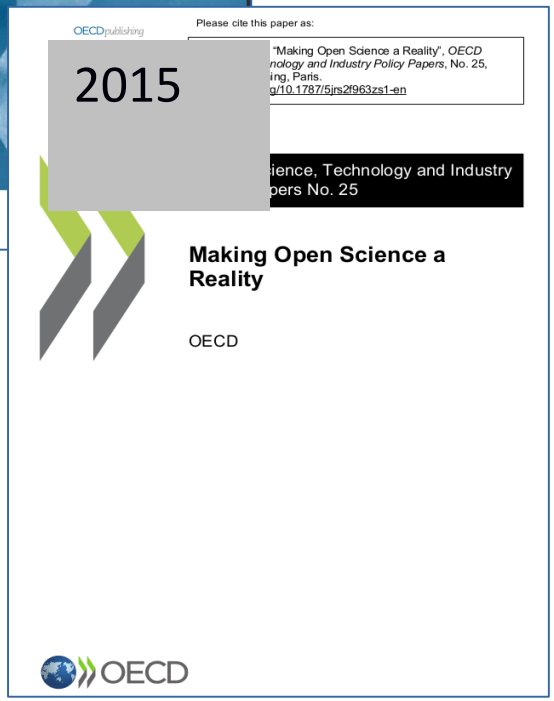
Open access to research data in an Open Science Context

SYSLOG Training
'Open Access Policies and Requirements to Publications and Research Data in Horizon 2020'
September 2015

2007



2015



Open Research Data (ORD)

- ORD refers to making research data freely available for reuse beyond the purpose for which they were originally collected
- Making Research data freely available aid further discovery, make scientific process more cost efficient and reliable
- ORD is part of a broader change: data driven science underpinning Open Science

Un cadre juridique et politique

Commentaires diapo n°6 et 7

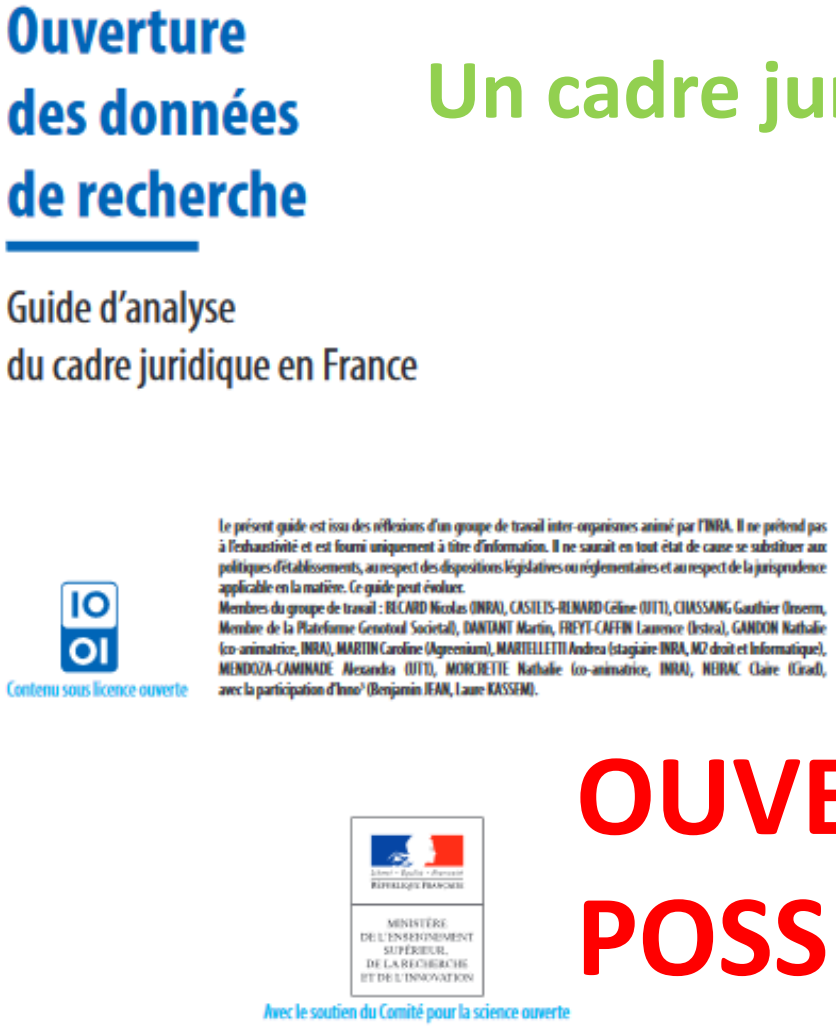
Question 2 : Les données sont donc précieuses, ont une valeur économique, elles sont stratégiques au cœur des dispositifs de recherche ?
Réponse 2: OUI les données en tant que production le sont!
Cette analyse, partagée au niveau international, est à l'origine des recommandations voire contraintes formulées par les financeurs en matière de soumission des projets.
Parmi ceux-ci l'UE a exprimé depuis une petite dizaine d'années une volonté politique forte se traduisant par les actions ou exigences progressives en matière de partage des données : « Open Reserch Data Pilot » ensuite généralisé à l'ensemble des données relevant du programme H2020.
In the 2014-16 work programmes, the ORD pilot included only selected areas of Horizon 2020. Under the revised version of the 2017 work programme, the Open Research Data pilot has been extended to cover all the thematic areas of Horizon 2020:
Les injonctions sont les suivantes:
=> Disposer d'un document de planification (PGD, on y reviendra)
qui s'appuie sur une liste de principes (les FAIR)
=> Choisir un entrepôt pour rendre accessible les données
Tout ceci avec une philosophie très pragmatique 'As open as possible, as closed as necessary'.

L'objectif ne se limite pas à dynamiser la Recherche mais aussi et très explicitement à stimuler l'innovation et l'économie. Ce point a notamment fait l'objet d'un rapport (en 2007) de l'OCDE établissant les principes de l'accès aux données obtenues sur fonds publics puis d'un second document, en 2015, sur la mise en œuvre en 2015 (Making Open Science a reality).

Tu veux dire que nos données sont précieuses, couteuses ?
=> C'est un capital à conserver, à protéger, à partager, à réutiliser!

Question 3 : ouverture oui mais dans un cadre bien défini au niveau européen et français non ?
Réponse 3: Cadre juridique – politique française
Le cadre juridique européen définit le partage des données du secteur public et invite au partage des données de la recherche.
Ce cadre est moins est moins contraignant que les exigences de l'UE pour le financement des projets
Le cadre juridique français est plus exigeant que l'européen. Il est aussi beaucoup plus simple depuis la loi Lemaire (ou république numérique, oct 2016) : financements majoritairement publics => ouverture des données (et des logiciels….considérés en fait comme tout autre 'document administratif'.
L'engagement français s'est traduit notamment par le lancement, en juillet 2018, du Plan Science Ouverte.
Pour autant l'approche se veut pragmatique et identifie beaucoup d'exceptions : « Ouvert autant que possible, fermés autant que nécessaire »

Un cadre juridique et politique


Ouverture des données de recherche

Guide d'analyse du cadre juridique en France

le présent guide est issu des réflexions d'un groupe de travail inter-organismes animé par l'INRA. Il ne prétend pas à l'exhaustivité et est fourni uniquement à titre d'information. Il ne saurait en tout état de cause se substituer aux politiques d'établissements, au respect des dispositions législatives ou réglementaires et au respect de la jurisprudence applicable en la matière. Ce guide peut évoluer.

Membres du groupe de travail : BECARD Nicolas (INRA), CASILES-RENARD Céline (UTT), CHASSANG Gauthier (Inserm, Membre de la Plateforme Genotoul Societal), DANTANT Martin, FREYT-CAFFIN Laurence (Irstea), GANDON Nathalie (co-animatrice, INRA), MARTIN Caroline (Agreenium), MARTELLETTI Andrea (stagiaire INRA, M2 droit et Informatique), MENDOZA-CAMINADE Alexandra (UTT), MORCHETTE Nathalie (co-animatrice, INRA), NEIRAC Claire (Cirad), avec la participation d'Inno³ (Benjamin JEAN, Laure KASSEM).

Contenu sous licence ouverte

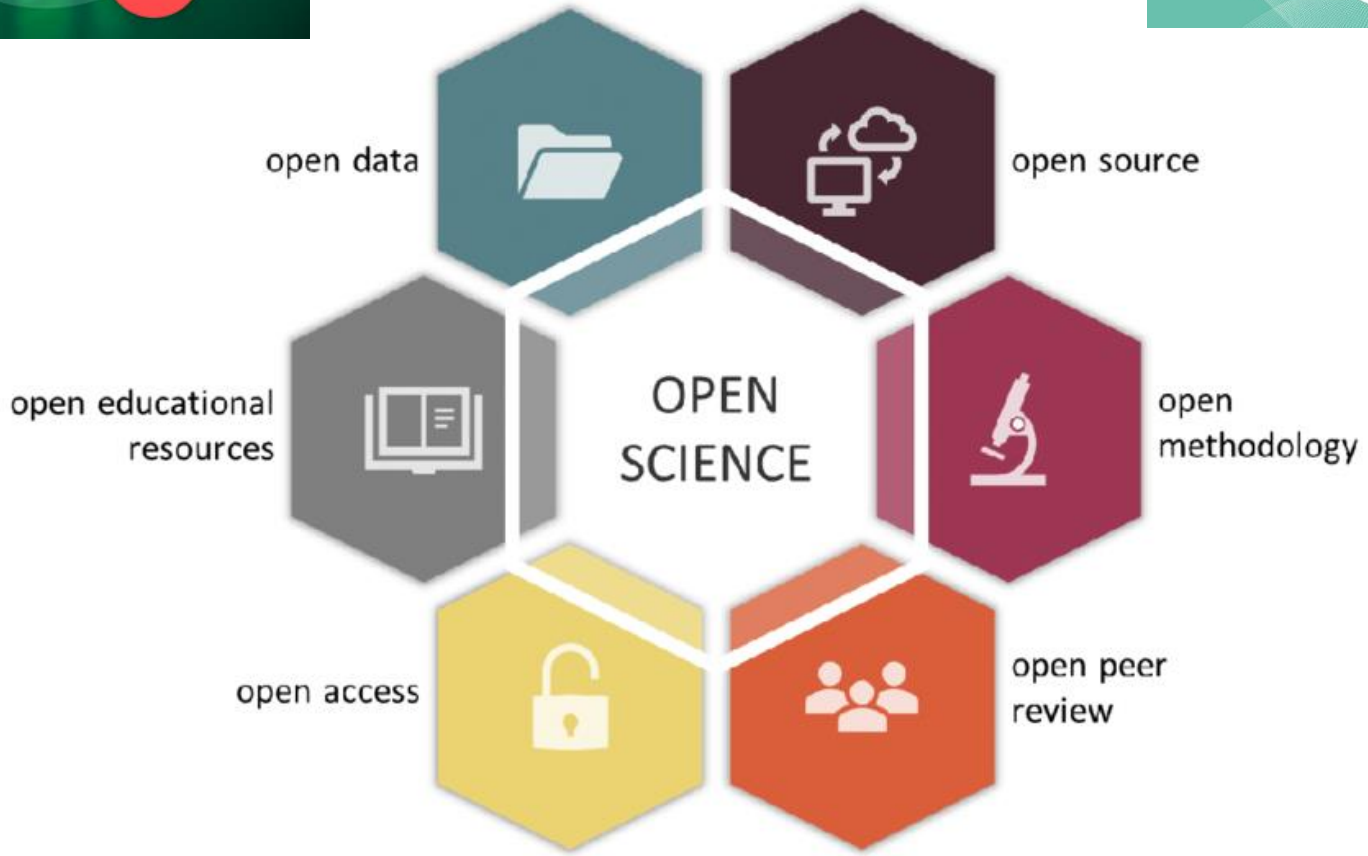
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION

Avec le soutien du Comité pour la science ouverte

V2 - Décembre 2017

OUVERT AUTANT QUE POSSIBLE, FERME AUTANT QUE NECESSAIRE

Contexte



Commentaires diapo n°9 et 10

Question 4 : il y a bien donc des restrictions à l'ouverture des données, le fameux RGPD, tu peux en dire quelques mots ?

Réponse 4: Oui, les données sensibles, personnelles ou relevant de la sécurité n'ont pas obligation a être ouvertes.
D'autre part et assez récemment la protection des données personnelles à fait l'objet d'un règlement européen, le RGPD (mai 2018). Ce règlement a pour objectif de 1) Renforcer les droits des personnes, 2) responsabiliser les acteurs traitant des données et 3) crédibiliser la régulation. Il oblige notamment a mettre en place par défaut ou anticipation les outils/procédures permettant de garantir la protection des données gérées (« privicy by design »). Les natures et objectifs de traitement des données doivent être renseigné dans un registre.
A l'INRA, la Déléguée à la Protection des Données :Nathalie Gandon.

Question 5 : alors pourquoi faut-il ouvrir les données ?

Réponse 5: Au delà des exigences des financeurs et de la valorisation que peut en tirer le monde économique, l'ouverture des données constitue une opportunité pour le monde de la recherche :

- les producteurs se voient reconnaître leur activité (doi, citation..) avec licences type CC-BY
- les « utilisateurs » dispose(ro)nt d'une matière première considérable pour consolider leur travaux et/ou développer des nouveaux axes, utiliser des nouvelles méthodos (ex data driven, IA)

….et cette ouverture permet la vérification des résultats publiés

- Pour le citoyen, il s'agit aussi de la bonne utilisation de ses impôts. L'économie, pour l'Europe, que représenterait le partage les données est évaluée à plus de 10 milliards par an.

Question 6 : Quelle est la place de l'open data dans l'open science ?

Réponse 6: L'Open Science représente une « nouvelle » façon de pratiquer la recherche en rendant accessible, en partageant autant que possible l'ensemble des éléments qui y contribuent : publications (ce n'est pas nouveau), données, code logiciel, échantillons.
C'est un partage pour une réutilisation libre par n'importe quel acteur, pas uniquement le monde de la recherche.
On peut considérer que l'OpenScience est la poursuite des pratiques de partage des résultats scientifique via leur publication dans des revues spécialisées initiée au XVIIème. On distingue généralemt 2 composantes « Open Access » et « Open Data »
L'Open Access rend libre l'accès aux publications.
L'Open Data est son équivalent pour les données. Il s'agit d'une évolution beaucoup plus importante car historiquement les données étaient rarement partagées. Elles prennent aujourd'hui une valeur en tant que telles et c'est vraiment nouveau. Leur publication peut d'ailleurs s'effectuer dans des revues de données...en Open Access !
Tout ceci est techniquement rendu possible par le numérique

Question 7 : un peu de vocabulaire dans ce contexte, c'est quoi une donnée? Une métadonnée?

Réponse 7: Bonne question ! Les précédentes aussi d'ailleurs
Une donnée (de la recherche) est une mesure prise ou une observation faite sur un objet d'intérêt, par exemple un arbre pour lequel on détermine son espèce, son âge et sa hauteur.
Une donnée (de la recherche) est une mesure prise ou une observation faite sur le contexte (un des élément du contexte) dans lequel a été prise la donnée, par exemple la date de la prise de mesure ou le nom de la personne qui l'a réalisée.
= « describe the content, context, quality, structure, and accessibility of a data set »

On dit généralement que la métadonnée est de la donnée sur la donnée.
La différence n'est pas toujours évidente et dépend du niveau auquel on s'intéresse. Par exemple si l'on s'intéresse à la biomasse de chacune des branches de cet arbre, les caractéristiques d'espèce, d'âge et de hauteur peuvent être considérées comme des métadonnées.
« MetsTaDonnée » a parfois été mal compris. Il ne s'agit pas d'une injonction à « mettre sa donnée » d'où les réticences parfois rencontrées. Ce doit être une démarche volontaire car bénéfique, qui porte sur les deux composantes, données et métadonnées, car il est évident qu'une donnée sans ses métadonnées ne peut-être réutilisée.

Question 8 : venons-en aux fameux principes F.A.I.R, c'est quoi ces 4 lettres ?

Commentaires: ci-joint 4 vidéos qui illustrent ces principes, issues de la plateforme de services ouverte aux partenariats « doranum » (https://doranum.fr/), projet engagé par le réseau des Urfist et l'Inist-CNRS.

Les principes FAIR

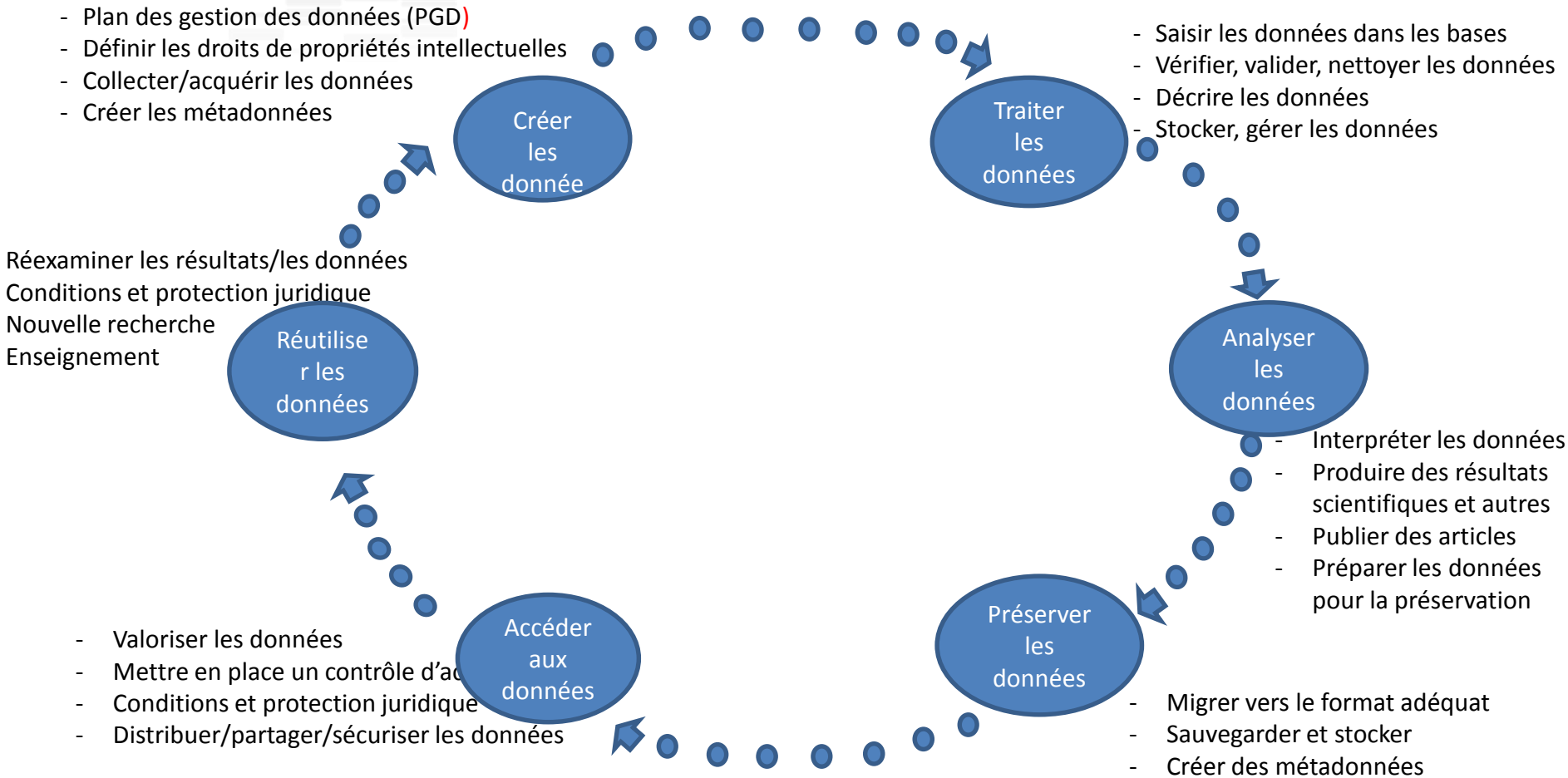
Réutilisation : https://doranum.fr/enjeux-benefices/minute-validation-reutilisation-donnees/

Accessibilité : https://doranum.fr/acces-visualisation/minute/

Findable : https://doranum.fr/identifiants-perennes-pid/minute/

Interoparabilité : https://doranum.fr/metadonnees-standards-formats/minute-interconnexion-donnees-recherche/

le cycle de vie des données



Commentaires diapo n°14

Question 9 : qui est concerné par l'ouverture des données ?

Réponse 9:
C'est l'affaire de tous, de tous les acteurs au cours du cycle de vie des données !
Les différents intervenants: collecteurs, gestionnaires, scientifiques etc. (= affaire de tous!)
-Au travers du cycle de vie des données au travers de 6 étapes. La mise en place de bonnes pratiques de gestion à toutes les étapes du cycle de vie des données se fait avec un outil: le PGD (plan de gestion des données) qui est un document qui décrit la façon dont les données seront obtenues, traitées, organisées, stockées, sécurisées, préservées, partagées,… au cours et à l'issue d'un projet (voir aussi diapo n°16)

Question 10 : « Qu'entend on par cycle de vie des données ? »

*Comment collecter, traiter et analyser? = données existantes/disponibles ou à créer/récolter

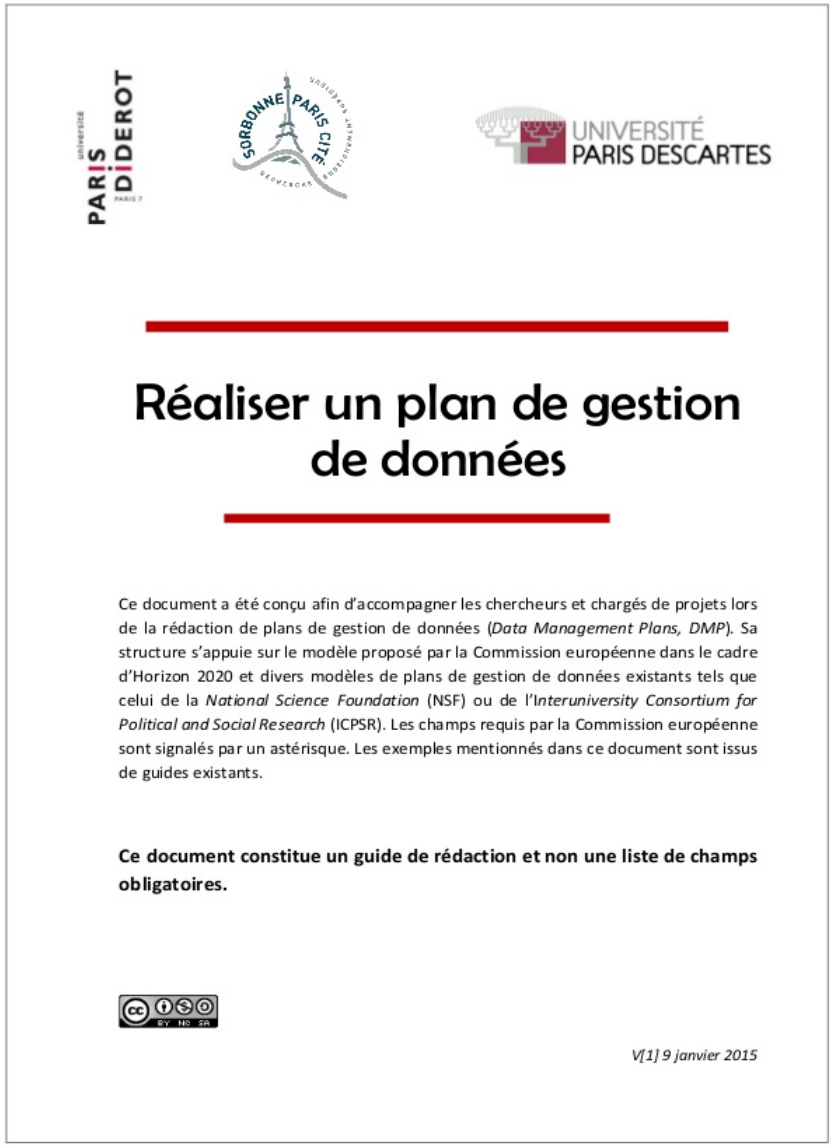
*Comment documenter? = documents descriptifs des hypothèses, des méthodes (protocoles, mode opératoires), plan d'échantillonnage, traitement éventuels (logiciels, algorithmes utilisés) et instruments/matériels

*Comment sauvegarder et stocker? = formats adéquats, référentiels de métadonnées, lieux, supports, durée

*Comment partager? = DataPaper, propriétés intellectuelle, RGPD, standards, entrepôts de données, R_Metadata/GeoFlow

*Comment citer? = DOI Le DOI est un identifiant international qui offre un accès pérenne aux ressources numériques (publications, données, revues, rapports, etc.) grâce à un lien unique et stable. Il est formé d'un préfixe et d'un suffixe séparés par un slash " / ".DIGITAL OBJECT IDENTIFIER- ".Exemple de DOI : 10.23638/LMCS-13(2:15)2017

*Comment conserver à long terme? = Entrepôts de données ex: ZENODO, DATA INRA



Réaliser un plan de gestion de données

Ce document a été conçu afin d'accompagner les chercheurs et chargés de projets lors de la rédaction de plans de gestion de données (*Data Management Plans, DMP*). Sa structure s'appuie sur le modèle proposé par la Commission européenne dans le cadre d'Horizon 2020 et divers modèles de plans de gestion de données existants tels que celui de la *National Science Foundation* (NSF) ou de l'*Interuniversity Consortium for Political and Social Research* (ICPSR). Les champs requis par la Commission européenne sont signalés par un astérisque. Les exemples mentionnés dans ce document sont issus de guides existants.

Ce document constitue un guide de rédaction et non une liste de champs obligatoires.

V[1] 9 janvier 2015

Question 11 : on a vu que FAIR repose sur des principes mais comment s'organiser concrètement, avec quels outils, quelles pratiques pour contribuer au partage des données ?

Commentaires: c'est un document évolutif au cours du projet, le PGD permet de se poser les bonnes questions pour…
* identifier les risques liés à la gestion des données, assurer la sécurité et la préservation des données sur le long terme,
* identifier les responsabilités, les rôles de chacun dans la gestion des données, planifier les ressources et compétences nécessaires à cette gestion,

- garantir des données fiables et bien gérées, compréhensibles, disponibles et préservées sur le long terme pour une réutilisation future (démarche FAIR)

- Ex: d'outil: OPIDOR https://dmp.opidor.fr

Déployé par l'Inist-CNRS

Voir: Cocaud, S., L'hostis, D. (2019). Pourquoi et comment rédiger un plan de gestion de données ? (Cours). 128 slides https://prodinra.inra.fr/record/447192

En résumer pour être concret!

1. ***Je rédige un PGD dès le début de mon projet pour anticiper et organisation tout cela !***
2. ***Je documente mes données (jeux de données) : Métadonnées et standards de métadonnées…***
3. ***Et pour les données à partager:***

- ***J'organise mes fichiers :*** *choix d'un format ouvert, un nom compréhensible…*
- ***Je vérifie que les principes éthiques sont respectés*** *(RGPD)*
- ***J'attribuer des identifiants pérennes :*** *citation des données - DOI*
- ***Je choisis une licence*** *(de diffusion) utilisables pour ouverture des données et droits de diffusion*
- ***Je choisis des Entrepôts***

Et cerise sur le gâteau je fais un DATAPAPPER !

Je consulte le n° spécial du cahier des techniques de l'INRA

Question 10 : Pour finir/ en conclusion ?

Réponse 10: De nouvelles pratiques pour le bénéficie de tous. Les données constituent une partie de notre bien commun.
Plus que de s'en persuader, il faut le comprendre, y adhérer.
Sans aller jusqu'à dire qu'il faut prendre le pli, les principes FAIR nous aider à repasser,
remettre sur la table nos pratiques de gestion de données.

Bien entendu on y croit dur comme FAIR!

Quelques références:

https://www.ouvrirlascience.fr/un-historique-du-libre-acces-aux-publications-scientifiques-et-aux-donnees/

https://ec.europa.eu/research/openscience/index.cfm

https://doranum.fr/

https://www6.inra.fr/datapartage

https://www.go-fair.org/fair-principles

https://gardian.bigdata.cgiar.org/metrics.php#!

https://research.csiro.au/oznome/tools/oznome-5-star-data