**Gestion et partage de données scientifiques**

**Contexte :** charte pour le libre accès aux publications et aux données 🡪 libre accès aux publis permis par la mise à disposition des données et leur réutilisation

Contexte national et eur des infrastruct de recherche : pour être inscrit sur les feuilles de route natio et eur il y a obligation d’avoir une politique de gestion des données et que les données répondent aux objectifs fixés : ouverture des données notamment

Chartes des infrastruct de l’INRA : Traçabilité, accessibilité à des données qualifiées

* + - 5 obj openscience :
* Des infrastruct de recherche connectée
* Une organisation des données pour le partage et la réutilisation
* Des approches prédictives en biologie
* De nvx modes de diffusion de la connaissance
* Le métier et l’environnement du chercheur adapté au numérique

La loi "Pour une République numérique" du 7 octobre 2016 inclue dans son périmètre la problématique des données de la recherche notamment dans son article 30.

Elle dispose que : "Lorsqu'un écrit scientifique issu d'une activité de recherche financée au moins pour moitié par des dotations de l'Etat, des collectivités territoriales ou des établissements publics, par des subventions d'agences de financement nationales ou par des fonds de l'Union européenne est publié dans un périodique [...] son auteur dispose, même après avoir accordé des droits exclusifs à un éditeur, du droit de mettre à disposition gratuitement dans un format ouvert, par voie numérique [...] la version finale de son manuscrit acceptée pour publication, dès lors que l'éditeur met lui-même celle-ci gratuitement à disposition par voie numérique ou, à défaut, à l'expiration d'un délai courant à compter de la date de la première publication. Ce délai est au maximum de six mois pour une publication dans le domaine des sciences, de la technique et de la médecine et de douze mois dans celui des sciences humaines et sociales."

Elle précise ensuite que : "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."

Directive européenne PSI (🡪 transposition en loi nationale) : en eur l’enseignement sup et de la recherche était pas concernée par cette directive alors qu’au nvx national ils ont choisis de les mettre ds la loi

Pas de def légales des données de la recherche. Pas de distonction juridique entre données brute, élaborée, métadonnée.

En général pas de droits d’auteur sur les données : on considère qu’elle appartient à l’établissement

On ne peut plus restreindre la réutilisation, ça peut être commercial ou non

46% des verrous à l’Open data sont des verrous juridiques (cf vidéo datapartage)

Interdiction de diffuser des données : données réalisées en exécution d’un contrat de prestation de service pour des personnes non publiques sauf accord explicite de cette personne

Toutefois, l’accessibilité à des données acquises en partenariats avec des entreprises privées à des fins industrielles ou commerciales peut-être limitée.

<https://guides-formadoct.u-bretagneloire.fr/c.php?g=491597&p=3362372>

Financement H2020 :   
Obligations : dpposer les articles au +tard lors de la publi  
déposer les données générées par le projet   
Etablir un DMP  
Associer des métadonnées aux articles et aux données

Les coûts induits sont éligibles au remboursement

<https://www.inist.fr/wp-content/uploads/formations/Principes-generaux-concernant-le-libre-acces-2017/story_html5.html>

2025.inra.fr

# Les enjeux de l’accessibilité aux données scientifiques

## Enjeux patrimoniaux

Obliger les chercheurs à rendre leurs données accessibles permet tout d’abord d’éviter de perdre des données. En effet, de nombreuses données sont encore dans des formats non pérennes voir peut-être non lisibles car laissées dans des formats de fichier trop vieux pour être remis à jour. De plus, de nombreux chercheurs ne savent pas où ont été rangées leurs données acquises tout au long de leur carrière. Il est estimé qu’environ 80% des données sont perdues 20 ans après publication (Roberge, 2015). D’autre part, les chercheurs ayant changé d’institut au cours de leur carrière n’ont pas toujours pris le temps de stocker correctement leurs données, de même pour les chercheurs partant à la retraite. De plus, dans le cas où les supports de stockage sont encore disponibles et lisibles par d’autres scientifiques, les données sont parfois trop peu décrites à l’aide de métadonnées pour être comprises et réutilisées par d’autres. Les techniques d’acquisition et de traitement des données changent d’années en années ce qui complique encore le travail de restitution d’anciennes données si celles-ci ont été mal décrites ainsi que leur réutilisation. Enfin les données apparaissant dans les publications sont généralement des données transformées et il est souvent difficile de faire le travail inverse pour retrouver les données brutes à l’origine.

Dans certains domaines (météorologie, évolution des sols), les chercheurs ont besoin de données assez anciennes afin de pouvoir établir leurs conclusions. L’accessibilité des données est une solution car elle permet un stockage généralement plus pérenne ainsi que leur réutilisation par d’autres membres de la communauté pour une étude différente que celle pour laquelle elles ont été acquises initialement.

## Enjeux économiques

Tout d’abord, il est inutile que les fonds de l’Etat financent des expérimentations supplémentaires sur un projet alors que les données produites par d’autres instituts pourraient être en partie réutilisées car complémentaires. L’Open data (données ouvertes) permet donc « d’accélérer l’innovation et le retour sur Investissement dans la R&D » (Pôle Données de la Recherche IST, 2018f). A titre d’exemple, on peut citer l’institut européen de bio-informatique (EMBL-EBI), organisation intergouvernementale fournissant gratuitement des données dans le monde entier. Les avantages sont estimés à 1 milliards de livres sterling par an pour les utilisateurs et leurs bailleurs de fonds ce qui correspond à 20 fois le coût opérationnel de l’institut (Beagrie, Houghton, 2016).

## Enjeux scientifiques

D’une part, l’accessibilité aux données garantit une certaine qualité de celles-ci. En effet, de nombreuses questions se posent actuellement sur la qualité des productions scientifiques et notamment sur le niveau de signification statistique des conclusions tirées. Il semblerait que seulement 20% seraient statistiquement fiables (Aumont, 2017). En publiant les données brutes, la recherche peut s’affranchir de ce risque car les données sont la preuve de l’exactitude de la publication ou non. Plus les données seront vues et réutilisées par des collaborateurs et plus le risque d’erreurs ou de conclusions hâtives pas toujours justes sera évité.   
De plus, en amont, le chercheur qui publiera ses données sera plus attentif sur leur qualité. Toutefois, publier ses données entraîne de nombreux changements techniques pour la communauté scientifique et de nouvelles responsabilités.   
Pour l’institut de recherche qu’est l’INRA, l’accessibilité aux données permet d’assurer l’intégrité de ses recherches et produits.

D’autre part, l’accessibilité des données peut être une passerelle entre les instituts de recherche, producteurs de données, et les instituts techniques, utilisateurs de données. En effet, chaque année, 7% des adresses mail des chercheurs ne sont plus fonctionnelles ce qui rend plus difficile l’accès aux informa tions. De plus, l’Open data permettrait également de faciliter l’usage des données notamment par ces communautés techniques.

Enfin, l’ouverture des données les rend citables et améliore la vitesse d’accessibilité à ces dernières. Découle de cela un taux de citation pouvant être plus important ce qui est dans l’intérêt des producteurs de données qui voient leur travail reconnu.

## Enjeux sociétaux

Comme il a été écrit précédemment, l’Open data permet une traçabilité des données et garantit une certaine qualité de ces dernières. L’accessibilité aux données contribue donc à améliorer l’image de la recherche auprès des citoyens ainsi qu’à augmenter leur confiance et donc leur participation à la science (sciences participatives).  
En termes d’éducation, les jeux de données publiés pourraient permettre à des étudiants de les intégrer à leurs productions pour l’argumenter ou encore de s’entraîner sur de vraies données.  
L’Open data permettrait de nouveaux liens entre citoyens et instituts de recherche.

## Enjeux éthiques

L’ouverture des données nécessite de bien réfléchir en amont à la question « Quelles données publier ? ». En effet, d’un point de vue éthique et juridique les chercheurs se doivent de respecter les droits d’auteurs, la vie privée et selon le domaine on peut parfois rencontrer des obligations de secret ou de sécurité.  
Selon le type de données, l’ouverture ne sera donc peut-être pas toujours possible notamment pour des données personnelles ou encore des données concernant le domaine de la santé par exemple.

<https://www6.inra.fr/datapartage/Partager-Publier/Les-enjeux-pour-l-Inra>

<https://www6.inra.fr/datapartage/Partager-Publier/Publier-un-Data-Paper/Seminaire-Data-Papers>

# Produire des données FAIR

FAIR est un acronyme signifiant : « Findable Accessible Interoperable et Reusable » soit en français : Trouvables Accessible Interopérable et Réutilisable. Ces quatre principes s’appliquent aux données produites par la recherche dans le cadre de la commission européenne et plus précisément de l’objectif Horizon 2020 (H2020). Ceux-ci sont censés permettre l’ouverture des données de la recherche. Chacun de ces quatre principes a été décliné en différents niveaux selon le collectif de travail FORCE 11.

## Findable

La première étape pour avoir des données ouvertes est de les rendre trouvables, que ce soit par des humains ou par des machines. 4 niveaux de caractéristiques des données « findable » ont été établies :

F1 : Les données et métadonnées sont associées à un identifiant unique et pérenne.

F2 : Les données sont décrites par des métadonnées riches.

F3 : Les données et métadonnées sont enregistrées ou indexées dans un dispositif permettant de les rechercher

F4 : Les métadonnées spécifient l’identifiant des données

## Accessible

Selon ce principe, les données et les métadonnées doivent être stockées durablement et leur accès facilité. Il est possible de spécifier des conditions d’accès (ouvert ou restreint) et d’utilisation (licence).

A1 : Les données et métadonnées doivent être accessibles par leur identifiant via des protocoles de communication standardisés

A1.1 : Le protocole utilisé est ouvert, libre, et universellement implémentable

A1.2 : Le protocole permet des procédures d’authentification et d’autorisation lorsque cela est nécessaire.

A2 : Les métadonnées sont accessibles même lorsque les données ne le sont plus

## Interoperable

Les données doivent être combinables avec d’autres données, téléchargeables, utilisables et intelligibles par des humains comme par des machines. Les systèmes informatiques doivent pouvoir dialoguer entre eux et échanger des données sans ambigüité (INIST, CNRS, 2018). Le web des données est la continuité du concept d’interopérabilité (DoRANum, 2017).

I1 : Les données et métadonnées utilisent un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances

I2 : Les données et métadonnées utilisent du vocabulaire qui suit les principes FAIR

I3 : Les données et métadonnées incluent liens vers d’autres données ou métadonnées.

## Reusable

Le but ultime des principes FAIR est d’optimiser la réutilisation des données. Pour cela les données et métadonnées doivent être doivent être correctement décrites.

R1 : Les données et métadonnées ont des attributs multiples et pertinents

R1.1 : Les données et métadonnées sont mises à disposition selon une licence claire et accessible.

R1.2 : Les données et métadonnées sont associées à leur provenance.  
R1.3 : Les données et métadonnées correspondent aux standards des communautés indiquées.

<https://www6.inra.fr/datapartage/Produire-des-donnees-FAIR/Reusable>

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

# Gérer des données

Les principaux risques dans la gestion des données reposent sur l’obsolescence matérielle, logicielle ou encore du format de fichier ainsi que la perte de contenu. On peut également citer d’autres risques plus secondaires tels qu’un mauvais lieu de rangement ou encore un mauvais nommage des fichiers. Nous allons donc voir dans cette partie comment éviter ces risques afin que les données respectent les principes FAIR. La première étape dans la gestion des données est celle d’établir un plan de gestion des données (Marie Claude Quidoz, 2018).

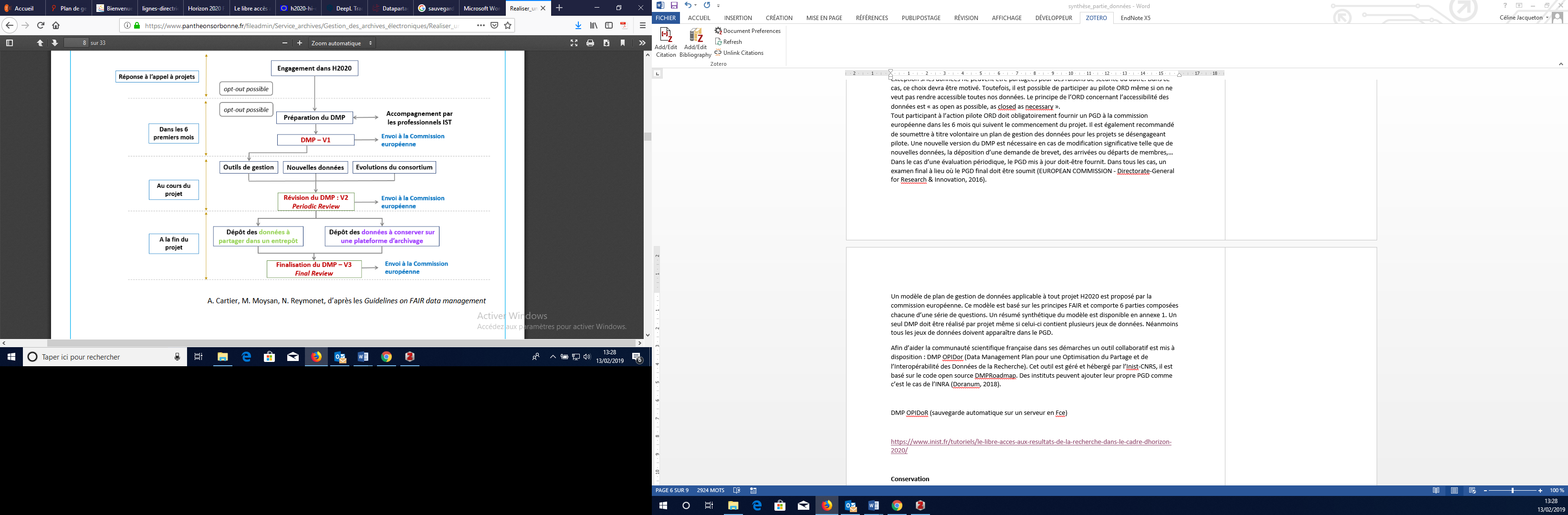
## Etablir un plan de gestion des données (PGD)

Un plan de gestion des données est un document formel et évolutif qui décrit de façon synthétique toutes les étapes du cycle de vie de la donnée : de la collecte à l’archivage et au partage, basé sur les principes FAIR. Ce document est considéré comme une bonne pratique de gestion et permet de gagner en temps, en efficacité et de mieux prévoir le budget à allouer aux données. Il n’est pas figé mais évolutif au cours du projet et se remplit pendant et après le projet.

Dans le cadre d’un projet financé par H2020 tout scientifique doit participer à l’action pilote ORD : « Open Research Data » soit en français « libre accès aux données issues de la recherche » sauf exception si les données ne peuvent être partagées pour des raisons de sécurité ou autre. Dans ce cas, ce choix devra être motivé. Toutefois, il est possible de participer au pilote ORD même si on ne veut pas rendre accessible toutes nos données. Le principe de l’ORD concernant l’accessibilité des données est « as open as possible, as closed as necessary ».   
Tout participant à l’action pilote ORD doit obligatoirement fournir un PGD à la commission européenne dans les 6 mois qui suivent le commencement du projet. Il est également recommandé de soumettre à titre volontaire un plan de gestion des données pour les projets se désengageant pilote. Une nouvelle version du DMP est nécessaire en cas de modification significative telle que de nouvelles données, la déposition d’une demande de brevet, des arrivées ou départs de membres,… Dans le cas d’une évaluation périodique, le PGD mis à jour doit-être fournit. Dans tous les cas, un examen final à lieu où le PGD final doit être soumit (EUROPEAN COMMISSION - Directorate-General for Research & Innovation, 2016).

Un modèle de plan de gestion de données applicable à tout projet H2020 est proposé par la commission européenne. Ce modèle est basé sur les principes FAIR et comporte 6 parties composées chacune d’une série de questions. Un résumé synthétique du modèle est disponible en annexe 1. Un seul DMP doit être réalisé par projet même si celui-ci contient plusieurs jeux de données. Néanmoins tous les jeux de données doivent apparaître dans le PGD.

Afin d’aider la communauté scientifique française dans ses démarches un outil collaboratif est mis à disposition : DMP OPIDor (Data Management Plan pour une Optimisation du Partage et de l’Interopérabilité des Données de la Recherche). Cet outil est géré et hébergé par l’Inist-CNRS, il est basé sur le code open source DMPRoadmap. Des instituts peuvent ajouter leur propre PGD comme c’est le cas de l’INRA (Doranum, 2018). Cet outil permet de partager le PGD avec ses collaborateurs et offre des sauvegardes automatiques sur un serveur en France. Toutefois, pour avoir toutes les versions du PGD il est fortement recommandé de l’exporter (Filatre, Hensens, 2017).



## Documenter et organiser les données

Cette partie est très importante, que ce soit pour l’auteur ou l’utilisateur des données. En effet une bonne gestion des règles de nommage, d’organisation et de renseignements des métadonnées permet de gagner en intelligibilité et d’accroître largement le potentiel de réutilisation des données.

### Nommage et organisation des fichiers et dossiers

Le nommage et l’organisation des fichiers (ou dossiers) de données permettent un gain de temps considérable, que ce soit pour l’auteur pendant le projet qui peut plus facilement localiser et distinguer les fichiers ou pour l’utilisateur lorsqu’il veut réutiliser les données. Une organisation structurée et un nommage clair permettent donc de faciliter le partage des jeux de données.  
En outre, cela permet également de réduire les risques de suppression ou de déplacements malencontreux et donc de préserver l’intégrité des données (Pôle Données de la Recherche IST, 2016).

Concernant le nommage, il est tout d’abord conseillé de choisir une convention de nommage adapté aux fichiers et aux contenus. Cette convention doit être appliquée pour tous les fichiers et toujours de la même façon. Il est préférable d’écrire cette convention quelque part afin de pouvoir s’y référer lors du nommage de nouveaux fichiers ou pour mieux (re)comprendre les titres des fichiers nommés ultérieurement.  
Quelques règles de bases :

* Le nom de fichier doit être court : l’INRA recommande au maximum 25 caractères (Pôle Données de la Recherche IST, 2018d). Les noms de fichiers peuvent inclure des abréviations, dans ce cas il est préférable de choisir des abréviations usuelles telles que par exemple AP pour Avant-Projet, CR pour Compte Rendu, VP pour Version Provisoire, VF pour Version Finale,…). Il est aussi possible et conseillé d’y faire figurer les initiales du chercheur ainsi que le numéro de version ou encore des dates. Attention cependant à ce dernier point, les dates doivent avoir un format consistant de type AAAA-MM-JJ selon la norme ISO 8601 (Unil, 2016). Les fichiers doivent être identifiables par leurs noms même en cas de changement du lieu de stockage
* Il est impératif d’éviter les caractères spéciaux ainsi que les accents, les cédilles ou encore les espaces. Ces derniers peuvent être remplacés par des underscores ou encore des tirets (Unil, 2016). Attentions certaines applications ne font pas la différence entre majuscules et minuscules donc supprimer les espaces en les remplaçant par des majuscules en début de mot n’est pas toujours pertinent (Pôle Données de la Recherche IST, 2018d).
* Dans le cas d’une numérotation il est préférable d’utiliser des zéros devant le numéro. Par exemple il vaut mieux écrire 001, 010, 100 plutôt que 1, 10, 100. Cette règle permettra d’assurer le bon ordre leur dans leur affichage (Unil, 2016).
* Il est également conseillé de ne pas garder toutes les versions intermédiaires d’un même fichier afin d’éviter les confusions.

Concernant l’organisation de l’arborescence un plan de classement est recommandé et tous les contributeurs au projet doivent idéalement participer à son élaboration (Laperdrix, 2016). Il est conseillé d’aller du plus général au plus spécifique et de mettre en place une organisation durable, qui ne serait pas impactée par des changements (Arseneau, 2015).

### Documenter les données

Afin de documenter les données, des métadonnées sont utilisées. Etymologiquement métadonnées signifie « données sur les données », c’est en fait un ensemble de données structurées permettant de décrire, expliquer, localiser les données afin que celles-ci soient plus compréhensibles. Elles permettent d’appliquer les principes FAIR et donc de faciliter (DoRANum, 2017), (INIST, CNRS, 2018) :

* L’accessibilité : elles favorisent la recherche des données notamment grâce à la recherche multicritères.
* L’interopérabilité : elles doivent permettre aux données d’être interprétables par des humains comme par des machines
* La pérennité : les métadonnées doivent permettre de documenter les éventuelles migrations d’un format de document à un autre
* La réutilisation : elles permettent une meilleure gestion et organisation des productions scientifiques notamment grâce à l’identification de l’auteur et la définition des usages.

Les métadonnées répondent généralement aux questions : « Quoi ? », « Où ? », « Quand ? », « Comment ? », « Pourquoi ? » et définissent (INIST, CNRS, 2018) :

* Le contexte de production : protocoles expérimentaux, procédés de traitement et d’analyse, auteur, date, localisation, objectifs de la collecte des données, financeur
* Le contenu intellectuel : définitions, titre, résumé, domaine de recherche, mots-clés, type de ressource
* Les caractéristiques techniques des données : format de fichiers, taille, organisation, liens
* Les propriétés et droits d’usage : détenteur des droits, licence, droits d’exploitation, conditions d’accès, outils spécifiques.

Ces métadonnées peuvent être dans un fichier : embarquées ou bien externes, c’est-à-dire dans un catalogue d’accompagnement par exemple ou encore un annuaire d’entrepôt. Elles peuvent être gérées par plusieurs approches différentes :

* Un tableau de type Excel par exemple : cette approche, bien que largement utilisée, est déconseillée car elle ne prend pas en compte la sémantique et ne permet pas de faire des requêtes facilement. Un tableau construit dans un format csv est préférable car il permet l’interopérabilité (lisible par humains et machines).
* Une structure en arbre de type xml : cette approche permet une meilleure organisation des métadonnées ainsi que la possibilité de lier des concepts avec d’autres.
* Un graphique sémantique avec des triplets RDF : les métadonnées sont organisées de façon structurées et cette approche permet de prendre en compte la sémantique.

Il existe des standards de métadonnées permettant de porter la recherche sur des critères définis et donc de la faciliter. Ces standards donnent une trame à suivre pour la description des données. On les choisit en fonction de la destination des données (certains entrepôts ont leurs propres standards) ou encore par domaines spécifiques. Pour aider les chercheurs dans leur démarche, le site internet du DCC (Digital Curation Center) et celui du RDA (Research Data Alliance) proposent des listes de standards de métadonnées et d’outils d’aide à la description. Il est nécessaire d’indiquer les standards utilisés pour les jeux de données dans le DMP.  
Ces standards peuvent être sous forme de schémas ayant pour avantages d’indiquer les relations existants entre les différentes métadonnées. Dans le cas d’un schéma de métadonnées il est souvent nécessaire d’utiliser le format xml permettant une structuration en arborescence.

Les schémas sont composés de 4 parties décrivant chaque élément :

* la signification (auteur, date, titre,…)
* le contenu (nombre, texte)
* la formulation (« Nom, Prénom », texte libre, norme ISO 8601)
* la valeur

Un schéma définit également le caractère obligatoire des métadonnées (obligatoire, recommandé et facultatif), les règles d’ajout d’éléments ainsi qu’éventuellement des règles plus spécifiques (obligation de renseigner tel champ si tel autre champ est rempli par exemple). Il est possible d’utiliser le schéma DataCite par exemple ou encore Dublin Core qui et générique et qui a fait l’objet d’une norme ISO (certifié ISO 9001)

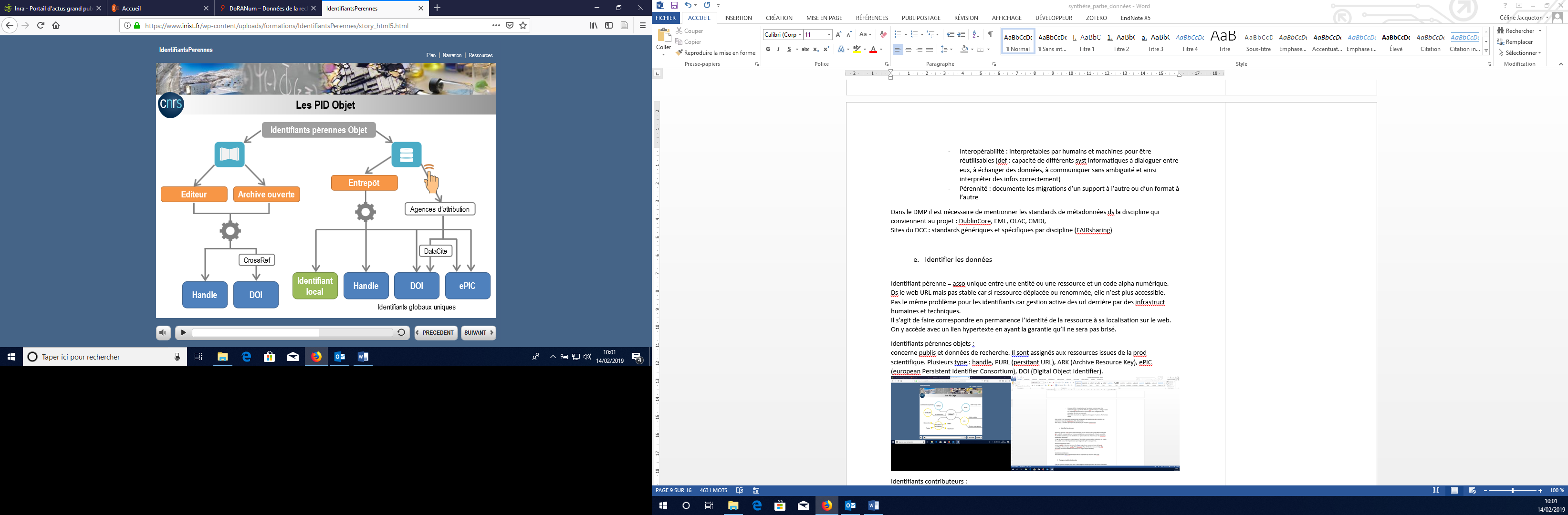
Au sein de ces schémas les métadonnées peuvent être décrites à l’aide de vocabulaire contrôlé, de thésaurus ou encore d’ontologies qui sont respectivement des listes de termes prédéfinis, des répertoires structurés de termes et « un ensemble structuré de termes et concepts représentants le sens d’un champ d’informations » (Wikipédia). Ces outils améliorent la visibilité des données ainsi que leur réutilisation. En agronomie, il est possible d’accéder à ces outils de description déjà existants en passant par exemple par l’Agroportal.

### Identifier les données

Dans le cadre du projet H2020, il est obligatoire de « faire référence à un système standard d’identification » (COMMISSION EUROPÉENNE - Direction Générale de la Recherche et de l’Innovation, 2016). On utilise donc un identifiant pérenne (PID) afin d’identifier ses données ou même plus largement les contributeurs et auteurs du jeu de données. Le PID sont des identifiants uniques et pérennes permettant d’associer chaque jeu de données à une ressource ou entité. Ce sont en quelques sortes les équivalents des URL dans le web. Toutefois, les URL ne sont pas des identifiants stables contrairement aux PID car la ressource n’est plus accessible dès lors qu’elle est déplacée ou renommée. La gestion active des PID par des infrastructures humaines et techniques permet d’éviter ce problème. Il existe deux types d’identifiants pérennes : objets et contributeurs.

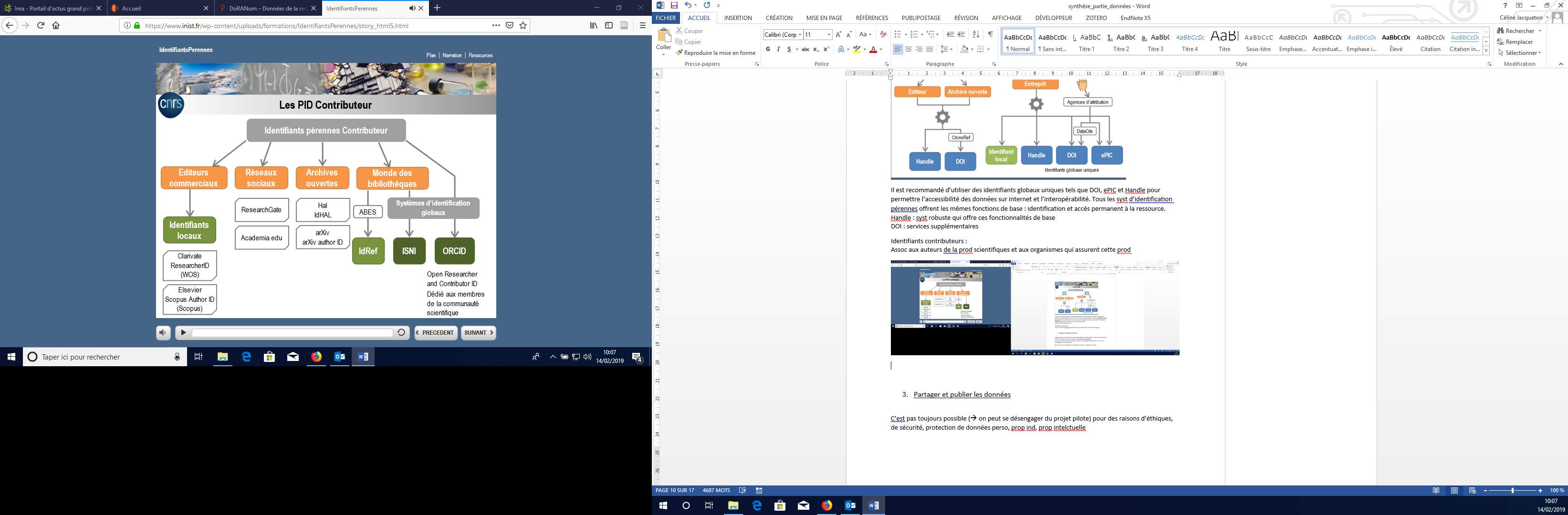
#### Identifiant pérenne objet :

Ces identifiants concerne des objets c’est-à-dire des publications et des jeux de données. Ils permettent la découverte, le partage, la citation et la réutilisation des données. De plus, ils font également le lien entre les données ayant permis la publication et la publication elle-même. Il existe une grande variété d’identifiants objets : handle, PURL (Persitstant URL), ARK (Archive Resource Key), ePIC (european Persistant Identifier Consortium), DOI (Digital Object Identifier). Il existe aussi des identifiants appelés URI qui étendent le principe des URL et permette de prendre en compte la sémantique. Toutefois ces identifiants ne seraient pas tout à fait persistants.   
Il est recommandé d’utiliser plutôt des identifia nts de type Handle, ePIC et DOI car ces derniers sont globaux et acceptés dans un grand nombre d’entrepôts. De tels identifiants garantissent l’interopérabilité et l’accessibilité des données.   
Par ailleurs, tous les identifiants pérennes garantissent des fonctionnalités de base qui sont l’identification et l’accès permanent au jeu de donnée. Toutefois, le système Handle serait plus robuste et le DOI offre des services supplémentaires. De ce fait, c’est le DOI qui est le plus souvent utilisé (INIST, CNRS, 2018), (DoRANum, 2018b). D’autre part, c’est également le DOI qui est recommandé par l’INRA.

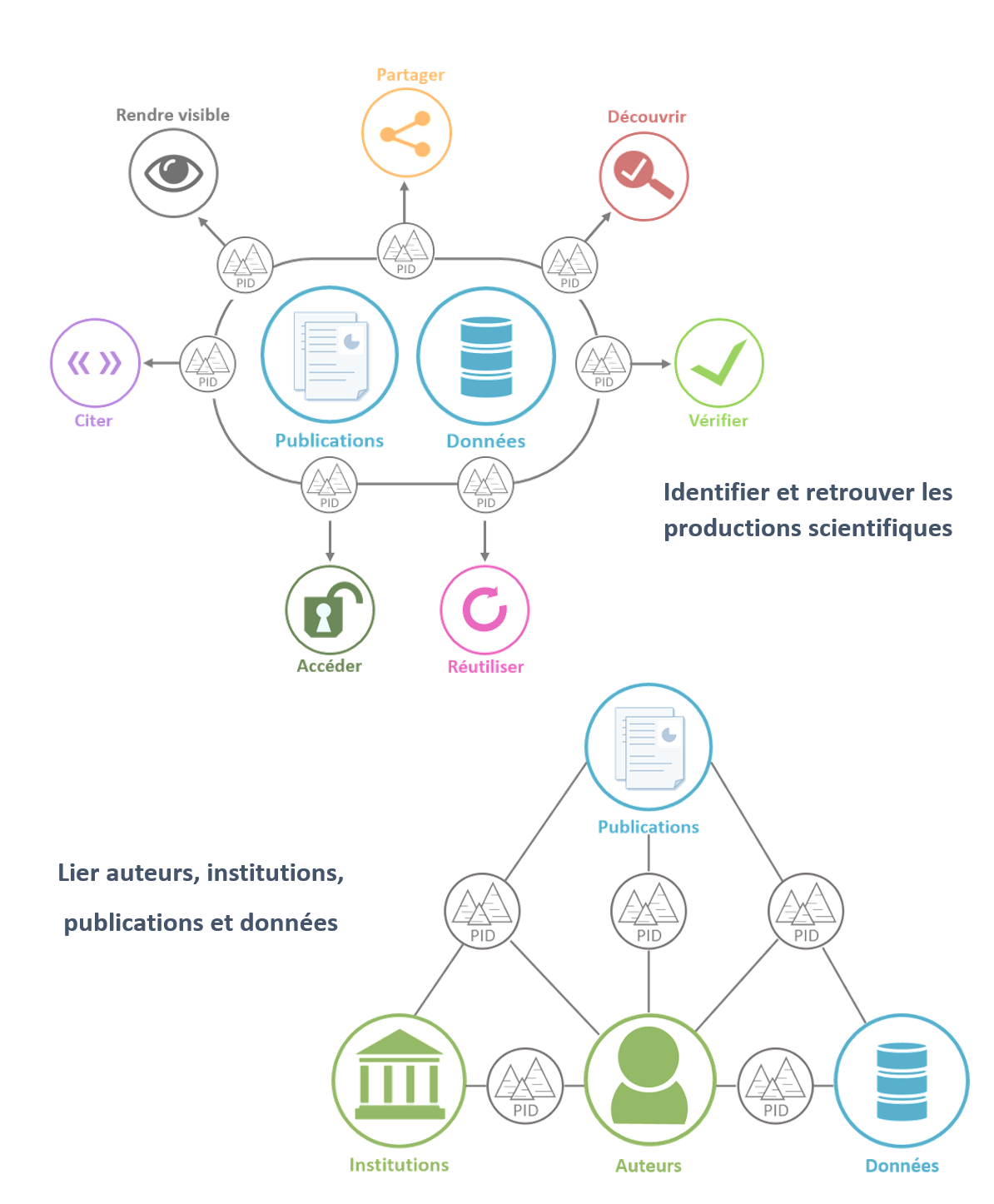


#### Identifiant pérenne contributeur :

Ces identifiants sont associés aux auteurs d’une production scientifique ou encore à un organisme de recherche assurant cette production. Cela permet notamment d’être plus visible et offre également les mêmes avantages que les identifiants objets en termes de citation et de partage. Comme pour les identifiants objets, il en existe plusieurs types, visible ci-dessous. L’obtention de ces identifiants se fait souvent de manière automatique sauf pour le researcher ID du WOS (Web Of Science) et les identifiants liées aux archives ouvertes (idHAL et ArXiv Author ID). Il est également possible d’avoir un identifiant global grâce aux plateformes ISNI (International Standard Name Identifier) et ORCID.



ABES : Agence Bibliographique de l’Enseignement Supérieur



# Partager et archiver des données

La gestion de données se fait bien souvent dans un but de partage, que ce soit avec d’autres contributeurs, la communauté scientifique ou alors la communauté citoyenne. Dans le cadre de l’Open Science et du projet H2020, les données doivent être autant ouvertes que possible. Toutefois, comme il a été écrit plus haut, le partage des données n’est pas toujours possibles pour des raisons éthiques, de sécurité ou encore de propriétés individuelle ou intellectuelle. En cas de doutes, l’INRA a développé un outil comprenant huit questions afin de conseillé les chercheurs sur la diffusion de leurs données (Andro et al., 2018) (http://www.bibliotheque-numerique.fr/DonneesDiffusables.php).  
Si le partage de données est partiellement ou entièrement impossible, les raisons doivent être explicitées dans le DMP. Le désengagement du projet pilote n’est pas nécessaire mais s’il y a lieu, les raisons du désengagement doivent elles aussi être explicitées dans le DMP.  
L’objectif du partage des données est l’accès direct et immédiat par un tiers aux données, facilitant la réutilisation (INIST, CNRS, 2018).  
Se pose également à cette étape la question de quelles données partager ? Sous entendu : données brutes ou données traitées ? Cette question doit être étudiée en amont dans le DMP en fonction du public à qui seraient destinées les données et de la réutilisation qu’il pourrait en faire.

L’archivage des données, également réfléchit en amont dans le PGD, a aussi pour objectif de permettre l’accessibilité des données. Attention toutefois à ne pas confondre stockage et archivage : l’archivage se réalise à la fin du projet, avant qu’il ne soit totalement terminé et ne porte que sur la dernière version des fichiers de données contrairement au stockage qui a lieu au cours du projet et qui permet de conserver à court terme les différentes versions.  
L’objectif principal de l’archivage de données est la conservation de l’information sur le long terme. A cette étape, les risques principaux sont donc liés à la lisibilité, l’intégrité et l’intelligibilité des données. En effet, les formats de fichiers peuvent devenir obsolètes empêchant leur lisibilité et le manque éventuel de documentation sur les données ne permettra pas leur compréhension à long terme lorsque les techniques d’analyses auront changées par exemple. L’archivage pérenne nécessite donc de mettre en place une veille sur les supports et formats de fichiers, les systèmes d’exploitation et les technologies matérielles.

### Format des fichiers

Il existe une grande diversité de formats de fichiers liés généralement à la nature de celui. Les principaux risques dans le choix du format sont ceux associés à la lisibilité du fichier. En effet, les mises à jour logicielles peuvent ne plus supporter le format, le format a peut évoluer et les versions précédentes peuvent ne plus être disponibles.

Pour avoir des données FAIR il est donc conseillé de choisir des formats :

* Ouverts donc sans restriction d’accès ou de mise en œuvre
* Non propriétaires. En effet, bien que le risque d’obsolescence soit également réduit avec des formats ouverts propriétaires ceux-ci comporte un autre risque : l’éditeur peut décider de fermer le format (Pôle Données de la Recherche IST, 2015).
* Largement utilisés par la communauté scientifique (M.-C Jacquemot-Perbal, Ciolek-Figiel, 2016).
* Normalisés (ISO, W3C) (CINES, 2017).

Tim Berners-Lee propose un programme Open data 5 étoiles concernant les formats de fichiers destinés à la publication de données ouvertes dans le cadre de la normalisation W3C (Wolrd Wide Web Consortium) (Michael Hausenblas, 2012) :

  
OL : Open Licence  
RE : Real Estructure ?  
OF : Open Format  
URI : Uniform Resource Identifier  
LD : Linked Data

La première étoile correspond simplement au fait de publier ses données sur le web avec une licence ouverte. C’est donc une manière très simple de publier pour le chercheur toutefois les données sont à l’intérieur du document et donc non moissonables ou extractables par des machines.

La deuxième étoile correspond à des formats où les données sont structurées c’est-à-dire lisible par des machines. Toutefois les données sont toujours à l’intérieur du document et il est possible de les extraire seulement avec un logiciel propriétaire. Néanmoins il est facile de publier des données de ce type.

La troisième étoile correspond à des fichiers de formats ouverts permettant une plus large utilisation par autrui. Toutefois pour publier ce type de format il peut être nécessaire de passer par des convertisseurs si les données sont dans un format propriétaire à l’origine. De plus, un format type csv entraîne une perte d’information par rapport au xls ou xslx telles que les formules ou les commentaires. Pour palier cela, les informations perdues peuvent être listées au préalable dans un fichier de type « lisez-moi » (Filatre, Hensens, 2017).

La quatrième étoile correspond à des formats où les fichiers sont « dans » le web, c’est-à-dire qu’elles peuvent avoir des liens avec d’autres ressources qui pointent vers elles (woddiscovery, 2010). Cela est rendu possible grâce à un URI (Uniform Resource Identifier) qui est un identifiant universel étendant le principe de l’URL sans qu’il y ait forcément besoin d’une page web derrière. Ces identifiants ne sont pas tout à fait pérennes mais des mécanismes permettent de faire les conversions nécessaires. L’avantage de ce format est de pouvoir lier ces données à d’autres. Toutefois, les graphes RDF sont plus difficiles à comprendre pour un utilisateur lambda qu’un fichier xls par exemple. De plus un tel format demande plus de réflexion au chercheur et un peu plus de travail.

Enfin la cinquième étoile correspond à « Linked data » c’est-à-dire que les données sont reliées à d’autres données. Ainsi les données sont d’autant plus visibles et leur valeur s’en voit augmentée.

Les standards du W3C sont ouverts et extensibles (INES et al., 2011) mais ne sont pas toujours largement utilisés et sont réservés, à partir de la quatrième étoile à des personnes ayant des bases en informatique.

Les formats généralement les plus conseillés sont :

* Le PDF/A : Le PDF Archivage est un format normalisé ISO adapté à l’archivage, largement utilisé et ouvert. Bien que ce soit un format propriétaire, il est documenté. De plus, les métadonnées sont intégrées dans le document sous un format XMP. Ce format est considéré comme pérenne, en effet, ISO ne peut pas annuler ce standard (PDF Tools AG, 2016). Toutefois ces types de fichiers pouvant contenir aussi bien du texte que des images, la migration vers un autre format est plus difficile (Gautier Poupeau, 2010).
* Le csv semble également être un format considéré comme pérenne (Archives d’Etat, 2011) (Arms et al., 2013) toutefois ce format peut entraîner des pertes d’informations notamment au niveau des commentaires et des formules. Il peut donc être nécessaire de le compléter avec un autre fichier où figureraient ces informations (Marie Claude Quidoz, 2018).
* Le xml, recommandé en 1998 par une publication du W3C (INES et al., 2011) est largement utilisé par la communauté scientifique et est lui aussi considéré comme pérenne (Rouchon, 2010) notamment s’il est associé avec un fichier xsd (Marie Claude Quidoz, 2018).

### Choix de la licence

La licence est un élément juridique définissant les modalités de partage (droits de copies et de diffusion) et de réutilisation des données (INIST, CNRS, 2018). Elle assure la préservation des droits d’auteur en imposant la citation de la source pour toute réutilisation.  
La communauté européenne recommande une licence creative commons notamment les licences CC 0 et CC BY. Ces licences sont gratuites et permettent aux utilisateurs des données de pouvoir les reproduire, les distribuer et les communiquer gratuitement (| Creative Commons France, [sans date]).  
Toutefois, l’attribution de ces licences est définitive. Il est possible de stopper la diffusion de l’œuvre sous licence Creative Commons ou encore de changer de licence mais dans ce cas toutes utilisations précédent ce changement resteront soumises aux conditions de la licence. Il faut également faire attention aux DRM (Data Rights Management) pouvant être mis en place par certains éditeurs afin de restreindre l’accès et l’utilisation de l’œuvre. En effet, ces DRM sont des protections techniques, incompatibles avec les licences creative commons (Fily, 2015).

Les licences CC BY :

Avec une licence de ce type, l’utilisateur doit citer l’auteur de l’œuvre et indiquer les éventuelles modifications apportées.   
Il existe différentes clauses possibles d’ajouter aux licences CC BY, créant ainsi 6 licences CC-BY différentes (| Creative Commons France, [sans date]) :

* BY NC : Ce type de licence empêche que l’œuvre soit réutilisée à des fins commerciales. Pour toutes les licences creative commons, le partage de fichiers n’est pas considéré comme une réutilisation commerciale (Fily, 2015).
* BY SA : Ce type de clause oblige le partage de l’œuvre, même modifiée, dans les mêmes conditions c’est-à-dire sous une licence identique à celle choisie initialement par l’auteur. Avec cette clause il est donc impossible pour l’utilisateur de combiner ensuite deux œuvres ayant des licences différentes (Fily, 2015).
* BY ND : Cette licence autorise le partage et la diffusion de l’œuvre initiale mais n’autorise pas sa modification sauf autorisation préalable. De ce fait, une telle clause interdit la traduction de l’œuvre (Fily, 2015).

Ces trois clauses peuvent ensuite être combinés : il existe des licences BY NC ND qui n’autorisent ni diffusion commerciale ni modification de l’œuvre initiale et des licences BY NC SA qui n’autorisent pas de d’utilisation commerciale non plus et qui oblige à partager d’éventuelles modifications sous les mêmes conditions de licence (| Creative Commons France, [sans date]).  
Dans le cadre d’une démarche OpenScience, l’INRA, comme le CIRAD, déconseillent l’utilisation des clauses ND et NC (Pôle Données de la Recherche IST, 2018c), (Fily, 2015). Le CIRAD va même plus loin en déconseillant également la clause SA à cause des possibilités réduites de compilation de différents jeux de données.

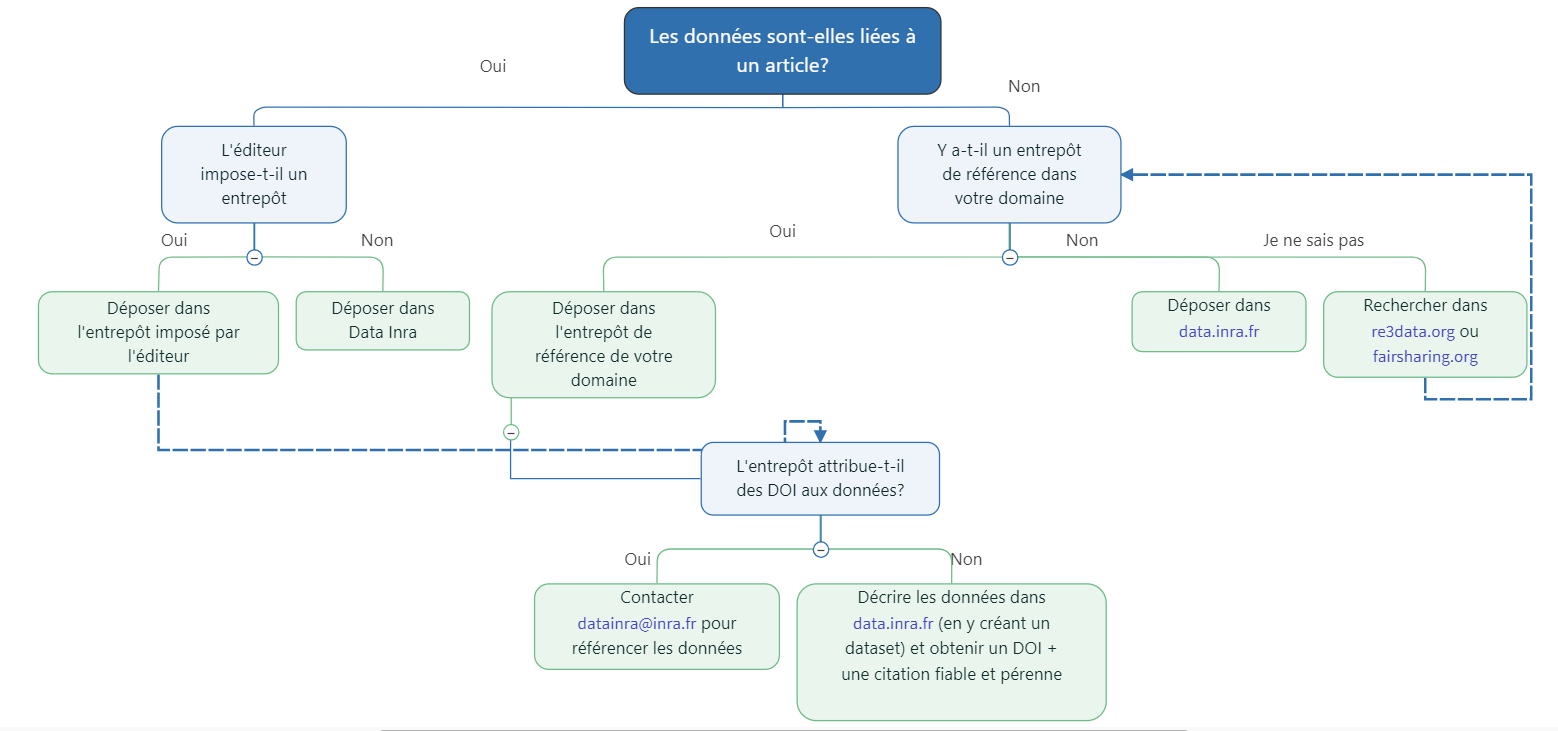
La licence CC 0 :

Cette licence permet de placer les données dans le domaine public, encourageant une large diffusion et réutilisation, même à des fins commerciales. Le titulaire des droits renonce alors à tous ses droits sur l’œuvre (INIST, CNRS, 2018). De ce fait, aucune restriction d’usage ne s’applique sur les données et l’auteur n’est pas obligatoirement cité (Fily, 2015).  
  
Plusieurs outils sont disponibles afin d’aider les auteurs dans le choix de la licence adéquate à leur jeux de données. Le site de creative commons propose par exemple une interface comportant plusieurs questions pour guider les auteurs dans leur choix (| Creative Commons France, [sans date]). Un assistant de licence EUDAT B2SHARE a également été développé (COMMISSION EUROPÉENNE - Direction Générale de la Recherche et de l’Innovation, 2016).

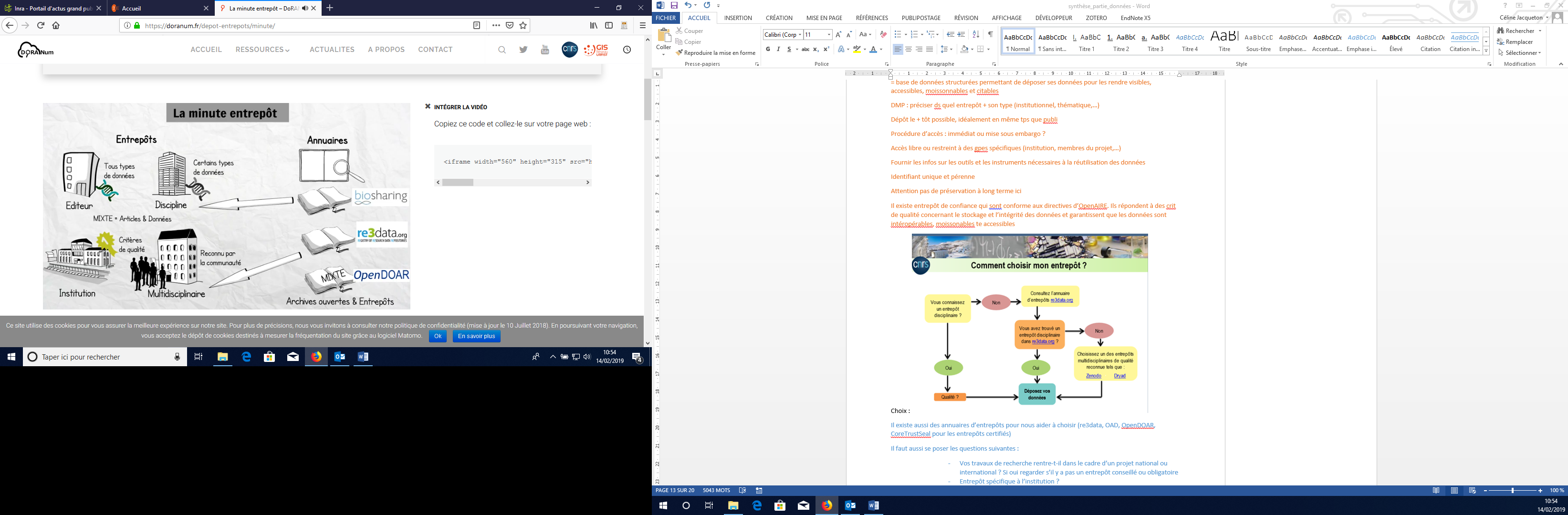
### Entrepôt de données et plateforme d’archivage

Un entrepôt de données est une base de données structurée où les auteurs peuvent déposer leurs données afin de les rendre visibles, accessibles, moissonables et citables. Le dépôt des données doit être réalisé le plus tôt possible, en même temps que la publication s’il y en a une. De plus, le choix de l’entrepôt et son type (institutionnel, thématique,…) doit être précisé dans le DMP. Avant de déposer ses données dans un entrepôt il faut s’assurer que toutes les étapes précédentes aient bien été respectées, à savoir : le nommage et l’organisation des fichiers, la documentation des données, le choix d’une licence, le choix d’un format de fichier pérenne et ouvert et l’attribution d’un identifiant pérenne et unique. De plus il faut également vérifier que le volume des fichiers est conforme à la taille maximale prise en charge par l’entrepôt choisi (DoRANum, 2018a).  
Si tel est le cas, il faut ensuite déterminer les procédures d’accès. L’accès aux données peut être immédiat ou non. En effet, il existe des procédures de mise sous embargo permettant de bloquer l’accès aux données pendant un certain temps afin de soumettre la publication associée aux données par exemple. D’autre part il est également possible de restreindre l’accès à des groupes spécifiques comme par exemple les membres du projet ou encore les membres de l’institution (INIST, CNRS, 2018).

Afin de choisir l’entrepôt le plus adéquat au dépôt des données, l’INRA propose le schéma ci-dessous (Pôle Données de la Recherche IST, 2018b).



En effet, si les données sont liées à une publication, la première chose à savoir est si l’éditeur impose un entrepôt. Si ce n’est pas le cas alors les données peuvent être déposé dans data inra qui est un entrepôt institutionnel. Si les données ne sont pas liées à une publication il est conseillé de les déposer dans un entrepôt spécifique au domaine scientifique. Le cas échéant il est possible de consulter des annuaires d’entrepôts tels que re3data, fairsharingOAD, OpenDOAR ou encore CoreTrustSeal pour les entrepôts certifiés. Il existe en effet des entrepôts certifiés conformes aux directives d’OpenAIRE et répondant à des critères de qualité en ce qui concerne le stockage et l’intégrité des données. Ces entrepôts garantissent l’interopérabilité, l’accessibilité et le fait que les données soient moissonnables (INIST, CNRS, 2018).  
Toutefois, pour le choix de l’entrepôt, il faut également prendre en compte les recommandations du financeur s’il y en a. De plus il semble important de vérifier que l’entrepôt soit connu de la communauté scientifique et que le moteur de recherche soit performant afin de garantir la visibilité de ses données (DoRANum, 2018a).



Certains entrepôts permettent également l’archivage pérenne des données. Cependant, si ce n’est pas le cas de l’entrepôt choisit il faut soumettre les fichiers de données à une plateforme d’archivage. Il est principalement attendu d’une plateforme d’archivage qu’elle garantisse la sécurité et l’intégrité physique des données en mettant en place des veilles technologiques. Elle doit également pouvoir traiter de gros volumes de données.   
Il est recommandé de choisir une plateforme suivant le modèle OAIS (Open Archival Information System) qui est un modèle conceptuel de référence, aujourd’hui devenu une norme ISO (ISO 14721). Ce modèle est abstrait et définit « l’architecture logique et les fonctionnalités d’un système d’archivage » (CINES, 2016). Ce modèle recense six fonctions d’archivage (INIST, CNRS, 2018), (CINES, 2016) :

* Entrée : réception et contrôle des objets à archiver. Les données sont ensuite transmises à l’entité « stockage » et leur documentation à l’entité « gestion de données ».
* Stockage
* Gestion de données
* Administration : pilote le système et établit les règles internes.
* Planification de la pérennisation : assure la veille technologique et prépare et planifie les éventuelles évolutions.
* Accès : interface directe avec les utilisateurs leur permettant de trouver l’objet qu’ils cherchent.

Et quatre acteurs principaux :

* L’archive : acteur interne, opérateur du système d’archivage.
* Le management : décideur politique.
* Les producteurs : fournissent les objets à archiver.
* Les utilisateurs : peuvent accéder aux objets archivés.

Ainsi que différent type d’informations :

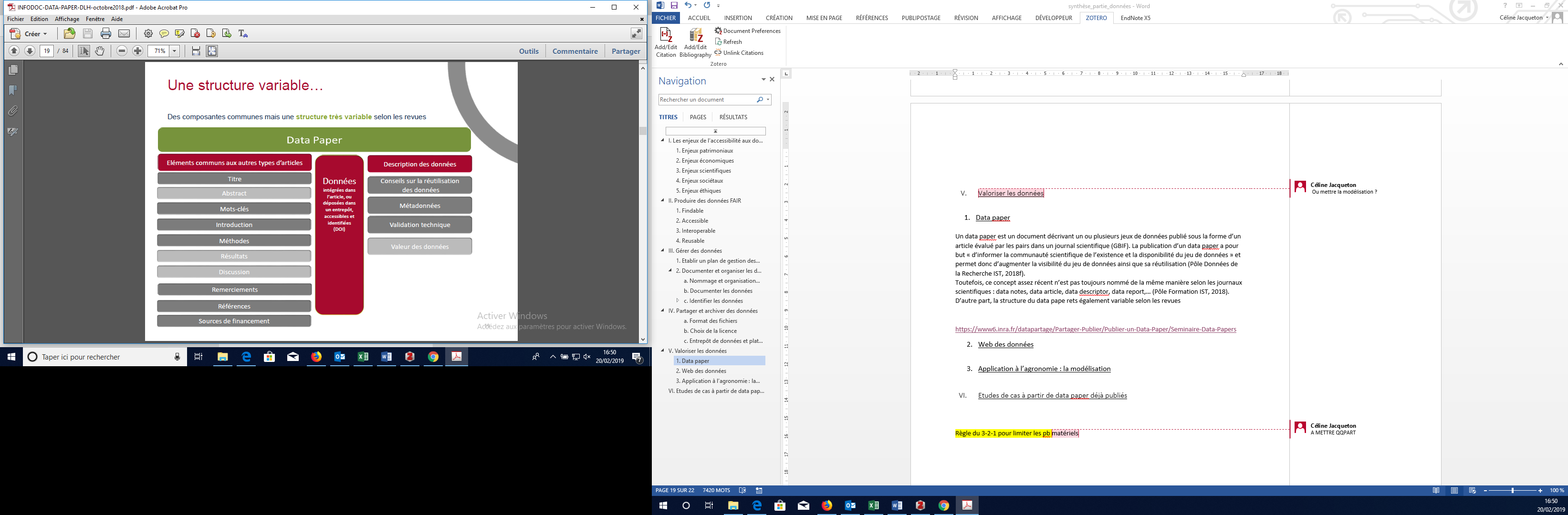
* Information de représentation qui peut être une information de structure ou une information sémantique
* Information de pérennisation : informations de provenance, de contexte, d’identification et d’intégrité
* Information d’empaquetage : relie les informations de pérennisation et de représentation
* Information de description : transmise à l’entité « gestion de données », elle permet de constituer le référentiel de base de données.

Il existe également une norme AFNOR NF Z42-013 décrivant les exigences opérationnelles et techniques à mettre en œuvre au sein d’un système d’archivage électronique ainsi qu’une norme de certification ISO 16-363.  
En complément des différentes certifications existantes, un outil en ligne permettant l’évaluation d’une archive numérique est disponible : Data Seal of Approval. De plus, il existe des normes sur l’expression métadonnées en vue d’un archivage telles que METS et PREMIS.  
Enfin le CINES propose également un outil permettant de tester la validité des formats de fichiers en vue d’un archivage. De plus cet organisme dispose également d’une plateforme d’archivage pérenne, PAC, certifiée et conforme aux normes OAIS et AFNOR Z42-013 (Pôle Données de la Recherche IST, 2018a) (INIST, CNRS, 2018).

# Valoriser les données

## Data paper

Un data paper est un document associé à un identifiant décrivant un ou plusieurs jeux de données publié sous la forme d’un article évalué par les pairs dans un journal scientifique. La publication d’un data paper a pour but « d’informer la communauté scientifique de l’existence et la disponibilité du jeu de données » et permet donc d’augmenter la visibilité du jeu de données ainsi que sa réutilisation (Pôle Données de la Recherche IST, 2018g).   
Toutefois, ce concept assez récent n’est pas toujours nommé de la même manière selon les journaux scientifiques : data notes, data article, data descriptor, data report,… (Pôle Formation IST, 2018).  
D’autre part, la structure du data paper est également variable selon les revues :



En effet, un abstract n’est pas toujours présent, de même pour l’introduction. Cependant les méthodes sont toujours expliquées et l’on y trouve également des remerciements, des références ainsi que les sources de financement. Contrairement à l’article scientifique, la partie la plus importante du data paper comporte sur la description des données et notamment sur leur réutilisation potentielle. Souvent on trouve un tableau récapitulatif des données après l’abstract. Certains data paper contiennent le jeu de données alors que d’autres renvoient aux jeux de données via un lien pérenne. Le jeu de donnée renvoie lui aussi au data paper également via un lien pérenne.

Beaucoup de revues scientifiques classiques publient également des data paper mais il existe également des revues dédiées en libre accès. Certaines de ces revues proposent des modèles ou des outils afin d’aider la rédaction et guident les chercheurs dans leur démarche en proposant des entrepôts. Les revues de publication peuvent être choisies selon :

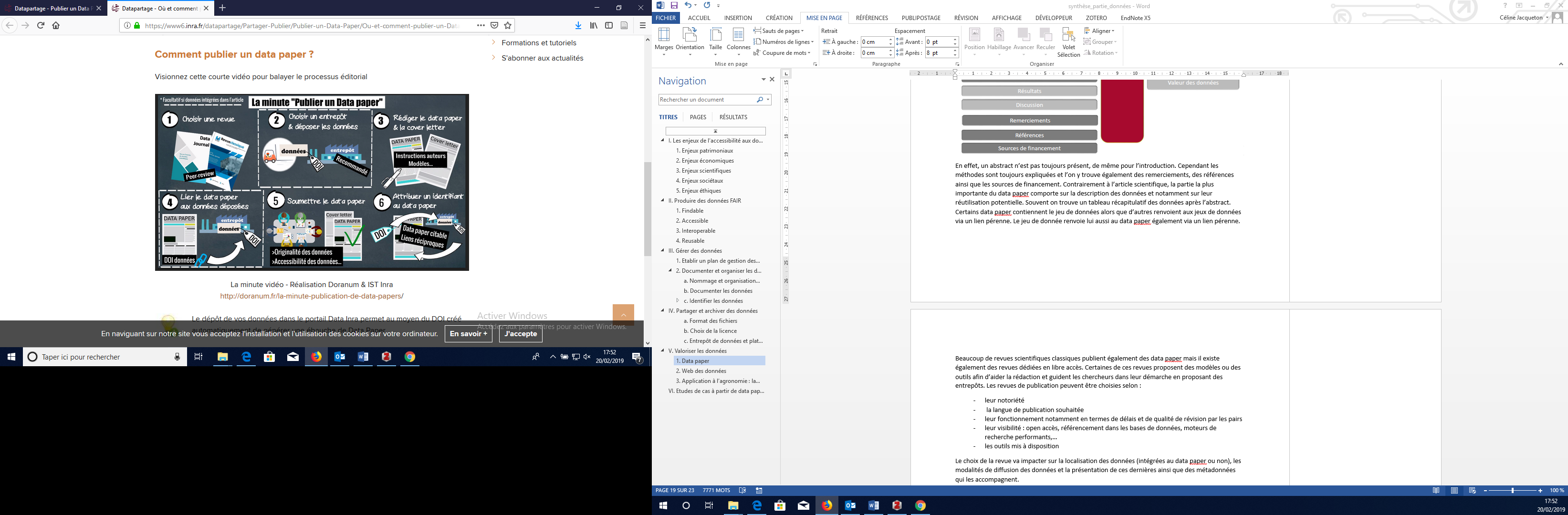
* leur notoriété
* la langue de publication souhaitée
* leur fonctionnement notamment en termes de délais et de qualité de révision par les pairs
* leur visibilité : open accès, référencement dans les bases de données, moteurs de recherche performants,…
* les outils mis à disposition

Le choix de la revue va impacter sur la localisation des données (intégrées au data paper ou non), les modalités de diffusion des données et la présentation de ces dernières ainsi que des métadonnées qui les accompagnent.   
L’INRA met à disposition un fichier excel recensant un grand nombre de revues publiant des data paper avec leurs modalités de fonctionnement et les frais liés à la publication

Pour qu’un data paper soit accepté les critères classiques concernant la qualité générale du manuscrit sont évalués à savoir les citations, le respect des instructions ainsi que plusieurs points concernant les données :

* leur importance et leur originalité
* leur valeur de réutilisation
* leur qualité et leur fiabilité
* leur accès
* la qualité et la rigueur de la méthode de collecte
* le choix des métadonnées et des formats

Le plus gros frein à la publication d’un data paper réside essentiellement dans son coût.



<https://www6.inra.fr/datapartage/Partager-Publier/Publier-un-Data-Paper/Seminaire-Data-Papers>

## Web des données

Le web sémantique ou web des données est une évolution du web proposée par Tim Berners-Lee en 2001 et standardisé par le W3C. Ce concept permet de favoriser la découverte et la réutilisation de données en proposant des standards d’échange (respectant les principes FAIR) en termes de formats et de protocoles s’appuyant sur le modèle RDF (Resource Description Framework). Le web sémantique est en fait un web intelligent qui comprend les informations plutôt que de les stocker, permettant d’ainsi mieux répondre aux attentes des utilisateurs (Dunod, 2008).

Publier dans le web sémantique apporte donc une plus-value aux données qui sont d’autant plus interopérables et accessibles. De plus le web sémantique étant considéré par certains comme l’évolution future du web d’aujourd’hui, il est bien de l’utiliser d’ors et déjà.

Le web sémantique s’appuie sur les principaux standards suivants(France, 2017), (Pôle Données de la Recherche IST, 2018e) :

* Les URI (Uniform Resource Identifier) dont nous avons déjà parlé un peu plus haut. Ces identifiants permettent d’accéder directement à la ressource mais étant une extension du principe des URL ils ne sont pas tout à fait pérennes. Toutefois il existe des mesures permettant une meilleure gestion de leur pérennité, on les appelle alors « URI http ».
* Le modèle RDF : ce modèle est un standard pour l’échange de données et a pour but de faciliter leur diffusion. L’élément de base est le triplet, il se compose de : sujet, prédicat, objet. Le RDF permet de lier les données et fournit une organisation qui peut être requêtée par les machines.
* L’OWL (Web Ontology Language) : ce langage permet la représentation des connaissances sur des classes et/ou sur des propriétés afin de les intégrer dans des environnements informatiques et pouvoir exercer des raisonnements dessus.
* SKOS (Simple Knowledge Organization System) : c’est un « modèle de données standards pour les systèmes d’organisation de connaissances » (Pôle Données de la Recherche IST, 2018e) tels que les vocabulaires contrôlés ou les thésaurus.
* SPARQL (Simple Protocole and RDF Query Language) : c’est un langage de requête pour un graphe RDF. Il permet de rechercher, modifier, ajouter ou supprimer des données RDF sans pour autant les télécharger.

Il existe un guide des bonnes pratiques pour publier des données sur le web à l’adresse suivante : <https://www.w3.org/TR/dwbp/#ProvideMetadata> dont les grandes lignes reprennent ce qui a été déjà dit dans ce document en termes de gestion et de partage des données

Site intéressant à consulter plus tard si leur outil est développé : https://www.dtls.nl/fair-data/find-fair-data-tools/

## Application à l’agronomie : la modélisation

# Etudes de cas à partir de data paper déjà publiés

## A global experimental dataset for assessing grain legume production, Cernay, Pelzer and Makowski, 2016

### Généralités sur le journal

Ce data paper est issu du journal scientific data qui est un journal en open access, proposé par le groupe de publication Nature. La publication n’a d’ailleurs pas le nom de data paper mais celui de « Data Descriptor ». Selon le journal, ce type de publication doit aider à la réutilisation des données et permettre la reconnaissance de ceux qui publient. Les publications de ce journal sont automatiquement indexée dans PubMed, MEDLINE, Google Scholar et Clarivate’s Web et automatiquement déposés dans PubMed Central qui est une archive ouverte.

Les data descriptor combinent du contenu narratif traditionnel et une description structurée des données, permettant de fournir un nouveau cadre de partage et réutilisation des données. Les principes du data descriptor s’alignent sur les principes FAIR.

La structure d’une telle publication s’articule de la manière suivante :

* Titre (110 caractères maximum)
* Résumé (170 mots maximum, sans références)
* Eléments de contexte (700 mots maximum)
* Méthodes
* Données
* Validation technique
* Notes d’utilisation
* Références

En général, pas plus de 3 tableaux et 3 figures par manuscrit.

* 2-year impact factor: 5.305
* 5-year impact factor: 5.862
* Immediacy index: 0.843
* Eigenfactor® score: 0.00855
* Article influence score: 2.597

### Structure

Concernant la structure, la partie la plus développé dans ce data paper est la partie sur les méthodes. Les parties sur le contexte et la réutilisation des données sont également plus développées que les autres parties. Le data paper comprend 7 figures au total et 2 tableaux.

ABSTRACT : Il se compose d’une partie rédigée décrivant le contenu et la structure du jeu de données et un tableau récapitulatif comprenant les colonnes suivantes : = comprenant le type de conception, de mesures, de technologies, de facteurs et les caractéristiques de l’échantillon

* Design Type(s)
* Measurement Type(s)
* Technology Type(s)
* Factor Type(s)
* Sample Characteristic(s)

BACCKGROUND AND SUMMARY  
Contexte de l’Assemblée Générale des Nations Unies + definition et intérets des légumineuses  
Pourcentages sur la production de légumineuse : %age de surfaces légumineuses allouées par sp  
Intérêt d’un tel jeu de données.  
Description du jeu données (1 carte comprenant les climats des différentes régions d’études et leur localisation (of course !) + 1 tableau récap du jeu de données +1 tableau pztant la struct de la base de données (nbe d’attributs classés par types et par tables)  
Utilisation possibles d’un tel jeu de données  
suggestion des données qu’on pourrait y ajouter

METHODE  
Biblio : ils expliquent où ils ont cherché et comment : sur quels critères ils se sont basés (les critères sont parfois justifiés mais ce n’est pas le cas pour tous)  
Structure de la base de données : Cette partie donne des informations sur l’organisation même de la base de données (tables mères et filles, liens, langage de requête), explique les valeurs (d’où vient la valeur « NULL » ainsi que la source et le pourcentage d’erreur estimé par rapport aux valeurs) et décrit les fichiers disponibles en plus de la base de données (leur format et l’endroit où les télécharger).

DONNEES  
Décrit une partie des attributs (à mon avis tous ceux pour lesquelles il pourrait avoir des doutes sur la signification, leur calcul, la façon de procéder en cas de données manquantes, données exclues…) + règles de décision : du style dans tel cas, on a procédé de telle façon  
Enonce également les valeurs extrêmes pour plusieurs attributs

VALIDATION TECHNIQUE  
Tous les articles ont été lus 3 fois par la même pers pour en extraire les données  
toutes les données reportées dans les tables ont également été checkées 3 fois par la même pers pour éviter les erreurs  
requêtes pour vérif que tt colle (relations   
tt le contenu des tables a été checké sur R  
17 auteurs contactés qd doutes

NOTES D’UTILISATION  
1 phrases de rappel sur ce sur quoi est basé l’article  
Utilisations possibles de la base de données : c’est hyper détaillé, ils donnent non seulement plusieurs issues mais également la marche à suivre pour faire telle ou telle chose  
exemples de ce qu’on peut ajouter et de si c’est compliqué ou non (du style : compliqué de rajouter une table mais fastoche de rajouter des données)  
Avantages du format csv pour la réutilisation  
Exemple de requête SQL pour extraire des données

REF   
188 ref (car basé sur la littérature)

INFO SUPP  
licence choisie

## Data on spatio-temporal representation of mineral N fertilization and manure N application as well as ammonia volatilization in French regions for the crop year 2005/06, Génermont et al., 2018

### Généralités sur le journal

Ce journal est proposé par le groupe ELSEVIER et permet de faciliter le partage et la réutilisation des jeux de données, quelque soit le domaine scientifique auquel ils appartiennent. Pour ce faire, ce type d’article permet une description détaillée des données et une recherche facilitée qui contribue à augmenter le nombre de citations et ouvre des portes à de nouvelles collaborations.  
Ce journal est en open access (l’article peut être librement lu, téléchargé, copié et distribué) et est examiné par les pairs. Le coût d’une telle publication s’élève à 500 USD hors taxes. Toutefois, ELSEVIER a mis en place des collaborations avec des organismes de financement ce qui permet aux auteurs de se voir rembourser les frais de publication à condition de se conformer aux politiques de ces organismes.  
Data in brief utilise un contrat de licence exclusif ce qui permet aux auteurs de conserver leurs droits d’auteurs et elsevier se verra accorder des droits d’édition et de distribution.  
Le taux de citation est de 0.7, le SNIP (Source Normalized Impact per Paper) de 0.287 et le SJR (SCImago Journal Rank) de 0.341. Le taux d’acceptation est de 77%.

Donne un DOI aux publis  
Les données de l’article doivent avoir été produites et appartenir à l’auteur ou à son institution

### Structure

* Titre, auteurs et labo
* Abstract : purement descriptif, donne la localisation des données et se doit de citer l’article scientifique lié aux données le cas échéant. Non obligatoire mais fait dans l’exemple : cite la licence utilisée.
* Tableau des spécifications :

|  |  |
| --- | --- |
| Subject area | *E.g., physics, chemistry, biology, economics, psychology* |
| More specific subject area | *Describe narrower subject area.* |
| Type of data | *Table, image (x-ray, microscopy, etc.), text file, graph, figure* |
| How data was acquired | *Microscope, survey, SEM, NMR, mass spectrometry, etc.; if an instrument was used, please give the model and make.* |
| Data format | *Raw, filtered, analyzed, etc.* |
| Experimental factors | *Brief description of any pretreatment of samples/data* |
| Experimental features | *Very brief experimental description* |
| Data source location | *City, country,institution and/or latitude and longitude (and GPS coordinates) for collected samples/data.* |
| Data accessibility | *State if data is with this article or in public repository; if public repository, please explicitly name repository and data identification number, and provide a direct URL to data. We recommend* [***Mendeley Data***](https://data.mendeley.com/)*if you do not have a trusted repository.* |
| Related research article | *If your data article is submitted as a companion paper to a research article, please cite your associated research article here; you may reference this as “in press” [1].*  *If this is a direct submission to* Data in Brief*, please cite the most relevant research article here from your reference list.* |

* Valeur des données : 3 à 5 points expliquant en quoi les données ont de la valeur pour la communauté scientifique, comment elles peuvent être réutilisées, avec quoi elles peuvent être comparées.
* Données : description brève du jeu de données inclus dans l’article (env. 250 mots). Dans l’exemple : 1 phrase d’introduction/contexte et description également des données supplémentaires.
* Plan expérimental, matériels et méthodes : décrit brièvement la manière dont ont été obtenues les données (env. 1p dans l’exemple)
* (Remerciements)
* Références (10 dans l’exemple)

Pas d’introduction, d’interprétation et de conclusion.

## Fifteen-year record of soil temperature at the Bear Brook Watershed in Maine, Patel et al., 2018

Scientif data également (Nature) mais ici les données ne sont pas sous forme de base de données ni issues de recherche dans la littérature. Ce sont des données expérimentales

Tableau comparatif des 2 revues

<https://www.nature.com/sdata/about#aims-scope>

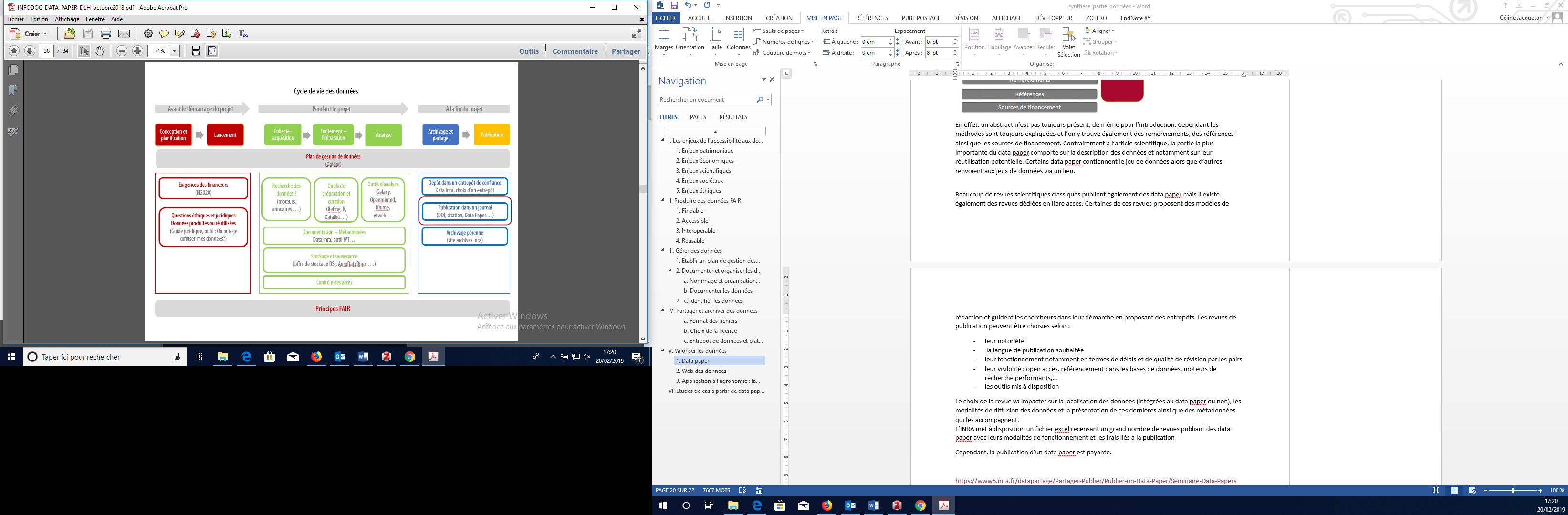
Genermont 2018

(Li 2018)

https://creativecommons.org/licenses/by/4.0/

🡪 rajouter cette licence dans la partie correspondante (elle est généralement utilisée pour data in brief et scientific data)

Règle du 3-2-1 pour limiter les pb matériels



Annexe

|  |  |
| --- | --- |
| Elément du PGD | Questions à examiner |
| 1. **Résumé des données** | • Indiquer l’objectif de la collecte/génération de données.  • Expliquer la relation par rapport aux objectifs du projet.  • Préciser les types et formats des données générées/collectées.  • Préciser si des données existantes sont réutilisées (le cas échéant).  • Préciser l’origine des données.  • Indiquer le volume prévu des données (s’il est connu).  • Mettre en évidence le potentiel de réutilisation des données : à qui seront-elles utiles ? |
| 1. **Données FAIR**   2.1 Rendre les données trouvables, y compris la fourniture des métadonnées | • Souligner la trouvabilité des données (fourniture des métadonnées).  •Décrire la procédure d’identification des données et faire référence à un système standard d’identification. Utilisez-vous des identifiants pérennes et uniques tels que les DOI ?  • Définir les conventions de nommage utilisées.  • Présenter la démarche de recherche par mot-clé.  • Décrire l’approche utilisée pour un versionnage explicite.  • Préciser les standards de création de métadonnées (le cas échéant). Si aucun standard n’existe dans votre discipline, décrire comment les métadonnées seront créées ainsi que leur type. |
| 2.2 Rendre les données librement accessibles | • Préciser quelles données seront librement disponibles. Si certaines données ne sont pas mises à disposition, le justifier.  • Préciser comment les données seront rendues disponibles.  • Préciser quelles techniques ou outils logiciels sont nécessaires pour accéder aux données. La documentation sur le logiciel nécessaire pour accéder aux données est-elle fournie ? Est-il possible d’intégrer le logiciel adéquat (p. ex. en code open source) ?  • Préciser où les données et les métadonnées associées, la documentation et le code sont déposés.  • Préciser comment l’accès sera fourni au cas où des restrictions s’appliquent. |
| 2.3 Rendre les données interopérables | • Évaluer l’interopérabilité de vos données. Préciser quels vocabulaires décrivent les données et métadonnées, quels standards ou méthodologies seront appliqués pour faciliter l’interopérabilité.  • Préciser si vous utiliserez du vocabulaire standard pour tous les types de données présents dans votre jeu de données pour permettre une interopérabilité interdisciplinaire. Dans le cas contraire, fournirez-vous un alignement avec les ontologies les plus fréquemment utilisées ? |
| 2.4 Accroître la réutilisation des données (au moyen de licences explicites) | • Préciser quelle licence sera attribuée à vos données afin de permettre la réutilisation la plus large possible.  • Préciser quand les données seront accessibles pour leur réutilisation. Le cas échéant, préciser pour quelle raison et pendant combien de temps un embargo sur les données est nécessaire.  • Préciser si les données produites et/ou utilisées dans le projet sont exploitables par des parties tierces, en particulier après la fin du projet. Si la réutilisation de certaines données est restreinte, expliquer pourquoi.  • Décrire les processus d’assurance qualité des données.  • Préciser la durée pendant laquelle les données resteront réutilisables. |
| 1. **Allocation des ressources** | • Estimer les coûts pour rendre vos données conformes aux principes FAIR. Décrire comment vous comptez financer ces coûts.  • Identifier clairement les responsabilités de gestion des données dans votre projet.  • Décrire les coûts et la valeur potentielle de la conservation à long terme |
| 1. **Sécurité des données** | • Examiner la récupération des données ainsi que le stockage sécurisé et le transfert de données sensibles. |
| 1. **Aspects éthiques** | • Pour être couvert dans le cadre de l’évaluation éthique, la section éthique de la Description de l’action et les livrables éthiques. Inclure des références et les aspects techniques liés s’ils ne sont pas couverts par la précédente évaluation. |
| 1. **Autres** | • Mentionner les autres procédures nationales/sectorielles/départementales/de financeurs pour la gestion des données que vous utilisez (le cas échéant). |

Annexe 1 : Tableau récapitulatif des questions à examiner dans le PGD – selon (COMMISSION EUROPÉENNE - Direction Générale de la Recherche et de l’Innovation, 2016)