**Gestion et partage de données scientifiques**

##### Introduction :

Le contexte national et européen des infrastructures de recherche favorise de plus en plus des politiques de gestions des données. En effet, pour être inscrit sur les feuilles de route nationale et européenne, les instituts sont dans l’obligation d’avoir une telle politique. Les données produites doivent répondre à des objectifs fixés dont notamment celui de l’ouverture des données.   
De telles politiques sont donc mises en place au niveau institutionnel, au sein de l’Inra. La charte de l’institut repose sur les six valeurs suivantes : « ouverture, partage, transparence, accès aux données, soutenabilité financière et conformité. », (INRA, 2016a). En effet, en signant la déclaration de Berlin en 2004 qui incite à rendre l’information scientifique librement accessible, l’Inra est entré dans une démarche Open Science. Cette démarche permet d’ouvrir les données de la recherche et de les rendre accessibles à tous. L’objectif est de capitaliser les efforts expérimentaux de chacun et d’éviter les duplications. Cela permettrait d’augmenter l’efficacité de la recherche. La loi Lemaire « Pour une République Numérique » du 7 octobre 2016 incluant la problématique des données de la recherche, renforce la démarche Open Science déjà mise en place. Découle de cela une charte publiée par l’Inra pour le libre accès à ses publications et données scientifiques en février 2017. Cette charte comprend sept points de recommandations destinés aux chercheurs afin de les inciter à entrer dans la démarche Open Science (INRA, 2016b).  
D’autre part, depuis quelques années, avec la révolution du numérique et l’acquisition de jeux de données toujours plus conséquents, l’Inra s’est fixée des objectifs Open Sciences. Ces objectifs comprennent :

* L’organisation des infrastructures de recherche afin qu’elles soient connectés
* L’organisation des données pour faciliter leur partage et leur réutilisation notamment grâce à la création d’un portail de données et de collaborations avec le CEA, le CNRS ainsi que l’Inria et l’Irstea.
* Favoriser les approches prédictives en biologie et en écologie
* Proposer de nouveaux modes de diffusion de la connaissance notamment en adaptant les revues propriétaires à la science ouverte mais également en encourageant des modèles alternatifs de publications.
* Adapter le métier et l’environnement du chercheur au numérique en demandant des compétences en analyse de jeu de données massifs par exemple ou encore en améliorant le processus de dématérialisation.

En outre, Horizon 2020, le programme de financement de la recherche et de l’innovation de l’Union européenne pour la période 2014-2020 s’inscrit lui aussi dans cette démarche de science ouverte. En effet, toutes les publications issues de projets financés par ce programme doivent être en libre accès. De plus, il existe un pilote ORD : Open Research Data qui tend à rendre accessibles le plus de données possibles acquises dans le cadre d’un projet H2020 (Ministère de l’enseignement supérieur, de la recherche et de l’innovation, 2014).  
Enfin, en 2018, apparaît le plan national pour la science ouverte qui « rend obligatoire l’accès ouvert aux publications ainsi qu’aux données issues de recherches financées sur projet ». Ce plan est composé de 3 axes décrivant des mesures afin de « généraliser l’accès ouvert aux publications », « structurer et ouvrir les données de la recherche » ainsi qu’afin de « s’inscrire dans une dynamique durable, européenne et internationale » (Ministère de l’enseignement supérieur, de la recherche et de l’innovation, 2018).  
Cette synthèse présentera donc les modalités liées à l’Open Science, comment s’inscrire dans cette démarche. Pour cela, une première partie abordera les enjeux de l’accessibilité aux données tandis que les parties suivantes se verront plus techniques et permettront de donner un cadre de gestion des données au sens large, toujours dans un objectif d’open science. En effet, ces parties traiteront de la production de données FAIR, de la gestion des données, de leur partage et de leur archivage et enfin des différentes formes de valorisation envisageables. La sixième et dernière partie comprends trois études de cas de data paper afin d’avoir des exemples concrets du cycle de vie des données à travers la structure d’un document de valorisation.

# Les enjeux de l’accessibilité aux données scientifiques

L’open data est un concept récent parfois encore mal connu du personnel scientifique. Les enjeux de l’ouverture des donnés sont multiples, que ce soit pour les personnes ou les instituts producteurs de données ou pour les utilisateurs des données (Pôle Données de la Recherche IST, 2018f), (Aumont, 2017).

## Enjeux patrimoniaux

Inciter 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 trop vieux pour être remis à jour. De plus, de nombreux chercheurs n’ont plus connaissance des lieus de stockage des données acquises au long de leur carrière ou n’ont pas toujours pris le temps nécessaire pour stocker leur données correctement. En effet, il est estimé qu’environ 80% des données sont perdues 20 ans après publication (Roberge, 2015). D’autre part, le problème du stockage des données est d’autant plus important pour des chercheurs ayant travaillé dans plusieurs instituts différents au cours de leur carrière. 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. Enfin les données apparaissant dans les publications sont généralement des données traitées et il est souvent difficile de retrouver les données brutes initiales à partir de celles-ci. Or, l’accessibilité à des données plus anciennes est impérative dans certains domaines comme la météorologie par exemple ou tous les domaines concernant l’évolution des sols.

## 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. 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, les données étant la preuve de l’exactitude de la publication. Plus les données seront vues et réutilisées par des collaborateurs et plus le risque d’erreurs ou de conclusions hâtives sera évité.   
De plus, en amont, le chercheur qui publiera ses données sera plus attentif. 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 informations. 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. 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 donc 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 il est parfois possible de rencontrer des obligations de secret ou de sécurité.  
Selon le type de données, l’ouverture ne sera donc pas toujours possible notamment pour des données personnelles ou encore des données concernant le domaine de la santé par exemple.

# 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 des 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.

# 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 son contenu. On peut également citer d’autres risques plus secondaires tels qu’un mauvais lieu de rangement ou encore un mauvais nommage des fichiers. Cette partie comprend donc des explications sur la manière dont il est possible de s’affranchir de ces risques afin que les données respectent les principes FAIR.   
La première étape dans la gestion durable des données consiste à 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 et en efficacité ainsi que de prévoir au mieux le budget à allouer aux données. Le plan de gestion 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 (voir même vivement conseillé) bien que la totalité des données ne soit pas rendue accessible. 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 du pilote. Une nouvelle version du DMP est nécessaire en cas de modifications significatives du projet tel que l’acquisition 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 la version finale du PGD doit être soumise (EUROPEAN COMMISSION - Directorate-General for Research & Innovation, 2016) (Figure 1).

Afin d’aider les participants au projet pilote ORD, 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 par l’Inist-CNRS : DMP OPIDoR (Data Management Plan pour une Optimisation du Partage et de l’Interopérabilité des Données de la Recherche). Cet outil 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). DMP OPIDoR permet de partager le PGD avec des 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).

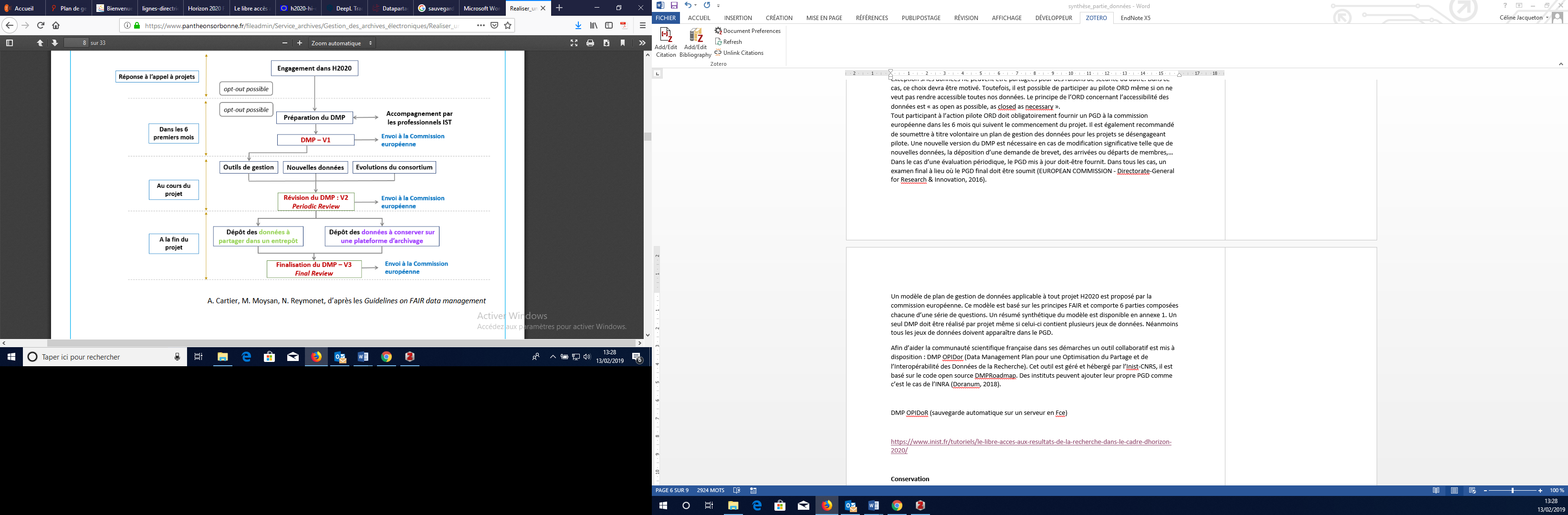


Figure 1 : Marche à suivre en terme de plan de gestion des données dans le cadre d'une participation au pilote ORD. A. Cartier, M. Moysan, N. Reymonet, d’après les Guidelines on FAIR data management.

## 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 que 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 adaptée 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 la date. 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, il est recommandé d’établir un plan de classement pour les projets impliquant plusieurs personnes. Idéalement, tous les collaborateurs doivent participer à son élaboration (Laperdrix, 2016). Pour l’élaboration d’un tel plan, 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 quelconques (Arseneau, 2015).  
Enfin, afin d’éviter tout problèmes de stockage et de perte de fichiers, il est conseillé d’utiliser la règle du 3-2-1 à savoir : produire 3 copies, sur 2 supports différents dans au moins 1 lieu distant (Marie Claude Quidoz, 2018).

### 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 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 la prise en compte de 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. Le site FAIRsharing propose également des standards de métadonnées selon la discipline étudiée. 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 également 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 identifiants 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 aux fonctionnalités de base. De ce fait, c’est le DOI qui est le plus souvent utilisé (INIST, CNRS, 2018), (DoRANum, 2018b). C’est également ce dernier qui est recommandé par l’INRA.

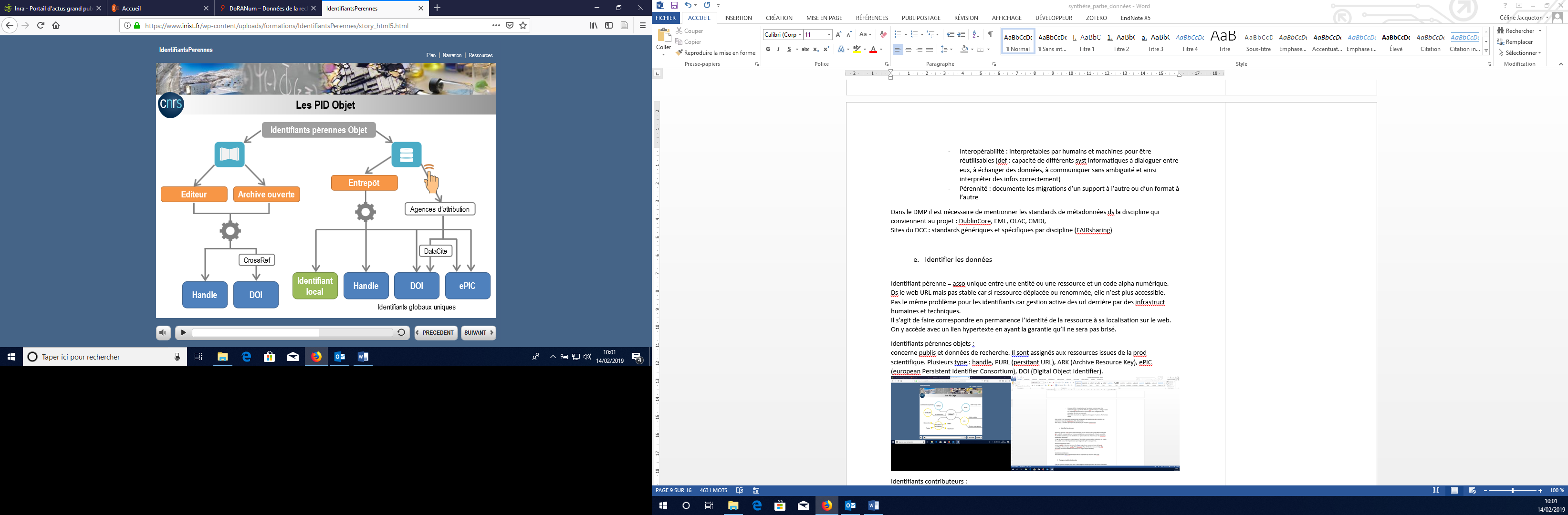


Figure 2 : Schéma des principaux identifiants objets pérennes, (INIST, CNRS, 2018).

#### 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 les mêmes avantages que les identifiants objets en termes de citation et de partage. Comme pour les identifiants objets, il en existe plusieurs types (*Figure 3*). L’obtention de ces identifiants se fait souvent de manière automatique sauf pour le researcher ID du WOS (Web Of Science) ainsi que 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.

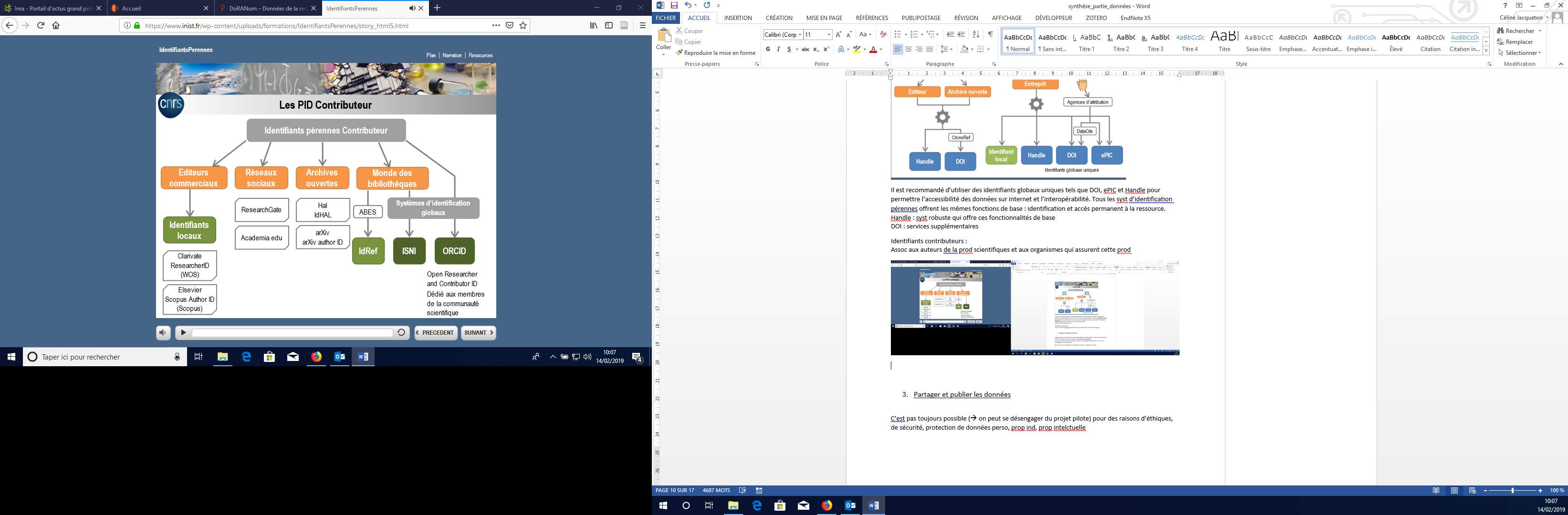


Figure 3 : Schéma des principaux identifiants pérennes contributeurs, (INIST, CNRS, 2018).

\*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 envisageable 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 conseiller 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. 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 des données. 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 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) (*Figure 4*).



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

Figure 4 : Programme Open data 5 étoiles, (Michael Hausenblas, 2012).

La première étoile correspond 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 permettant la structuration des donnés et donc la lisibilité par les 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 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 solides 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). Néanmoins, 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 précédant ce changement resteront soumises aux conditions de la première 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 est possible d’ajouter différentes clauses aux licences CC BY :

* 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 l’utilisation commerciale non plus et qui obligent à partager d’éventuelles modifications sous les mêmes conditions de licence.  
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 la plus utilisée dans le cadre d’une démarche Open Science semble être la licence CC BY 4.0. Cette licence permet le partage et la réadaptation du jeu de données, que ce soit à des fins commerciales ou non.

#### 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 est indispensable de 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 suivant (*Figure 5*).

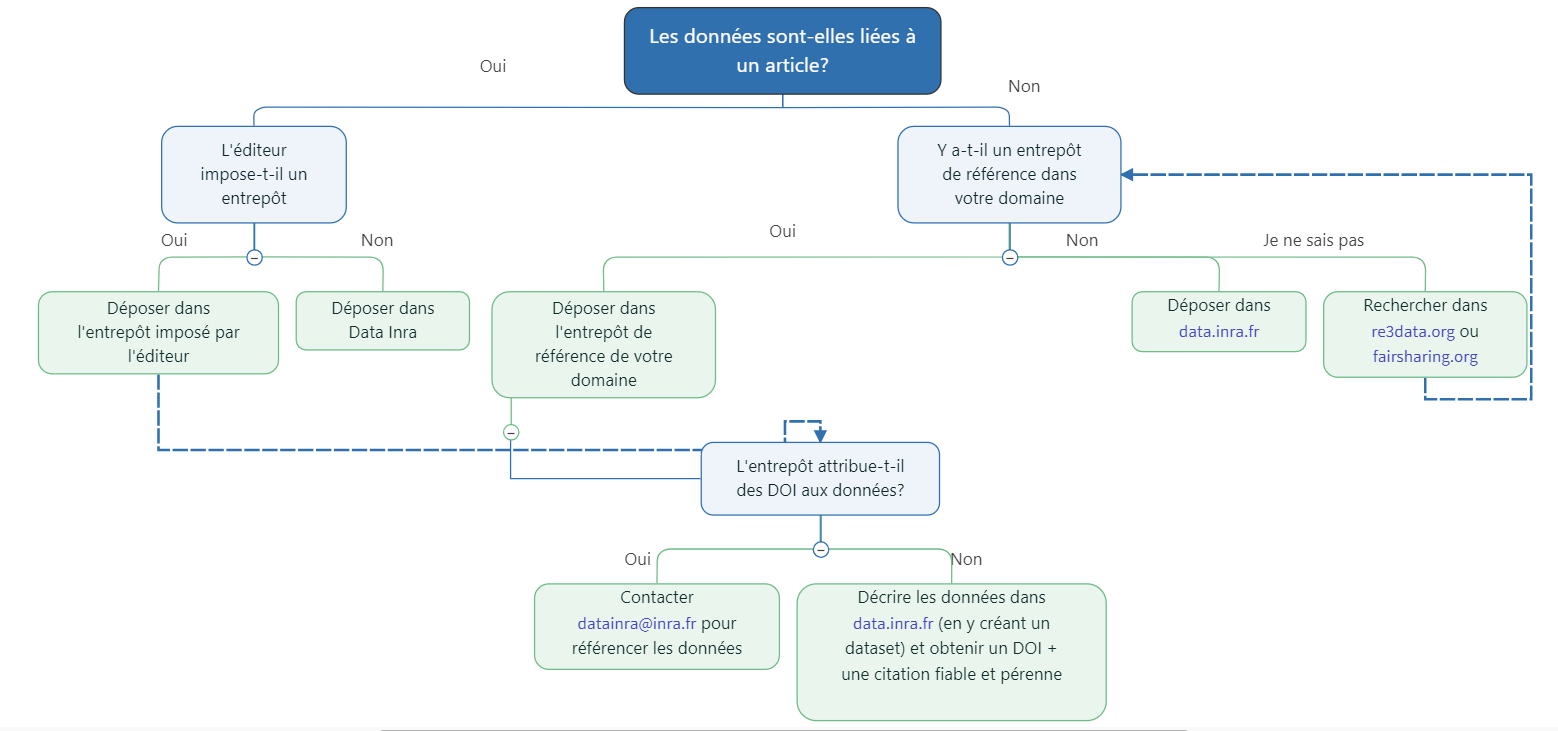


Figure 5 : Schéma d'aide au choix d'un entrepôt, (Pôle Données de la Recherche IST, 2018b).

En effet, si les données sont liées à une publication, il faut savoir si l’éditeur impose un entrepôt. Si ce n’est pas le cas alors les données peuvent être déposées 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) (*Figure 6*).

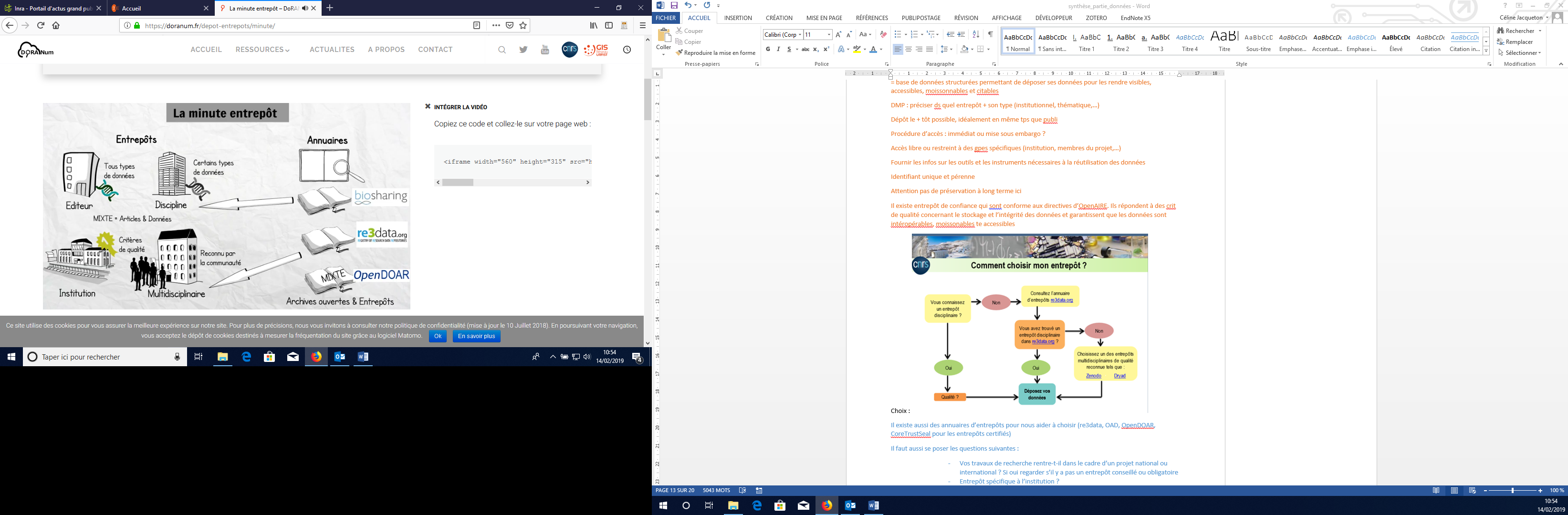


Figure 6 : Schéma présentant la démarche du choix d'un entrepôt, (DoRANum, 2018a).

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). Il 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 des 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 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 d’un jeu de données », (Cirad, 2018) 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 (*Figure 6*).

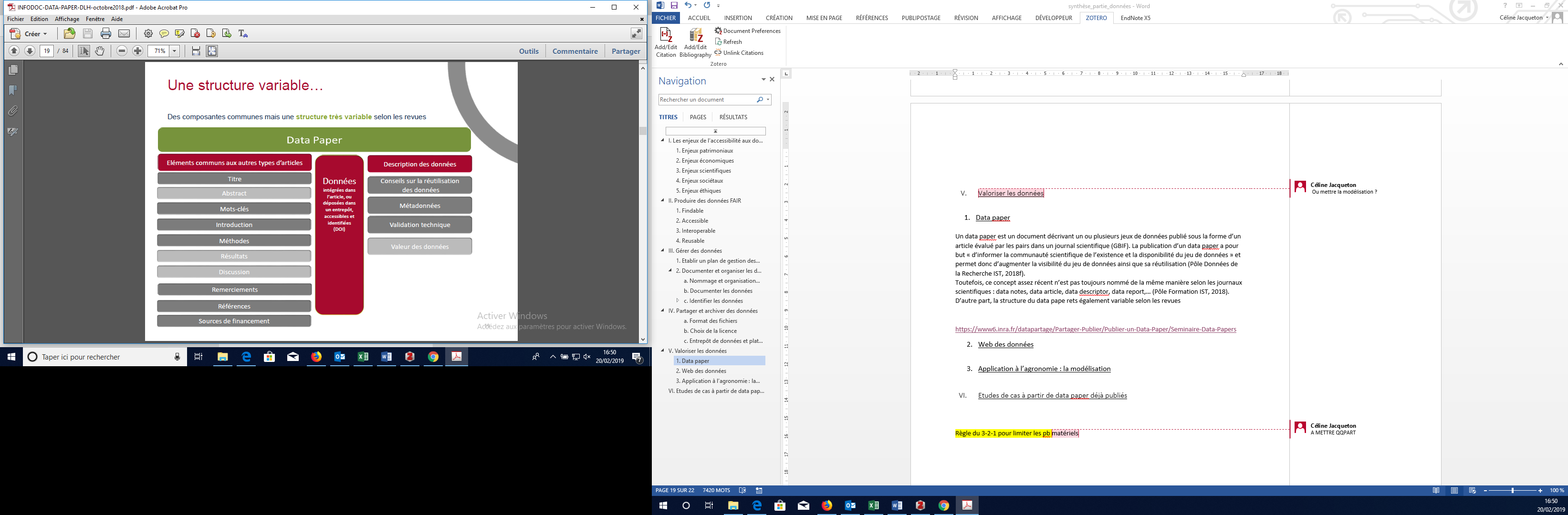


Figure 7 : Structure générale d'un data paper, (Pôle Formation IST, 2018).

En effet, l’abstract n’est pas obligatoirement 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 porte sur la description des données et notamment sur leur réutilisation potentielle. Souvent le contenu comprend un tableau récapitulatif des données après l’abstract. Certains data paper contiennent le jeu de données alors que d’autres renvoient à celui-ci via un lien pérenne. Dans ce cas, le jeu de donnée renvoie lui aussi au data paper également via un lien pérenne.

Beaucoup de revues scientifiques classiques publient des data paper mais il existe également des revues dédiées en libre accès. Certaines de ces revues offrent 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.

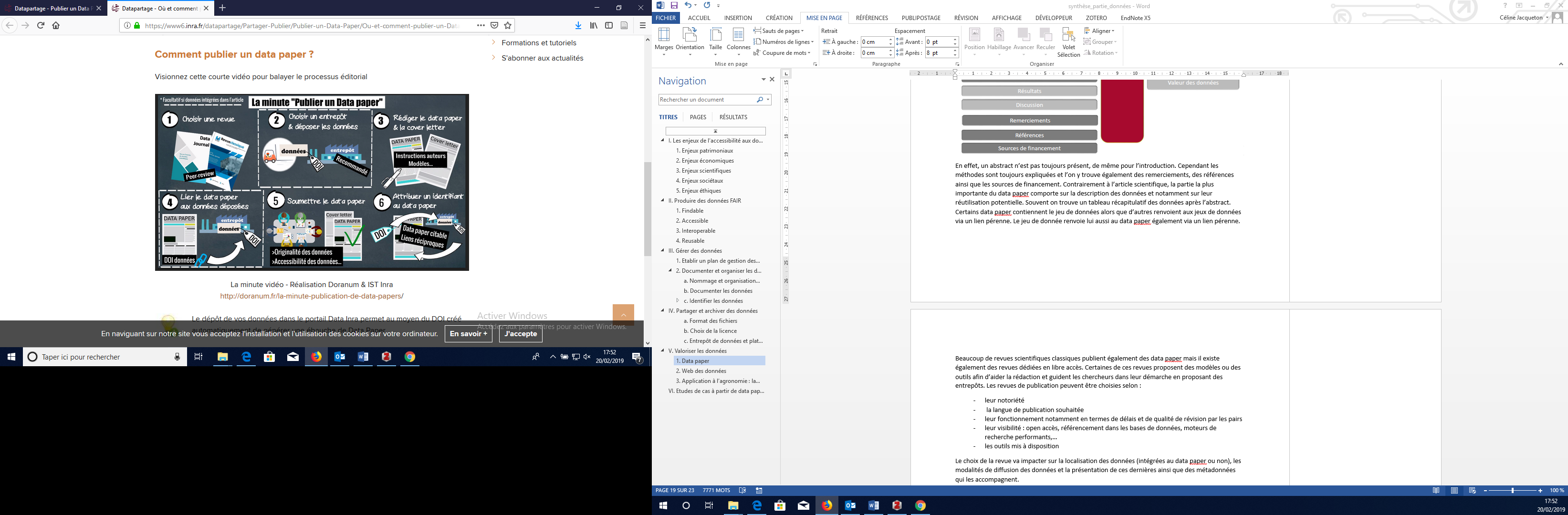


Figure 8 : Démarche de publication d'un data paper, (Pôle Données de la Recherche IST, 2018g).

## 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 de 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, l’utiliser d’ors et déjà peut être une bonne forme de valorisation des données.

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

* Les URI (Uniform Resource Identifier) dont nous avons déjà parlé 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 disponible 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

## 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 en open access et 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. Ces publications sont automatiquement indexée dans PubMed, MEDLINE, Google Scholar et Clarivate’s Web et automatiquement déposés dans PubMed Central qui est une archive ouverte. Le coût de publication s’élève à 1305 euros.

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

Selon le journal la structure d’une telle publication s’articule de la manière suivante :

* Titre (110 caractères maximum)
* Abstract (170 mots maximum, sans références) comprenant un tableau récapitulatif avec des colonnes portant sur le type de conception du jeu de données, le type de mesures, de technologies, de facteurs ainsi que sur les caractéristiques de l’échantillon.
* Eléments de contexte (700 mots maximum)
* Méthodes
* Données
* Validation technique
* Notes d’utilisation
* Références
* (Remerciements)

Afin de faire un choix entre différents journaux, le site présente quelques chiffres clés :

* Facteur d’impact (rapport entre le nombre de citations et le nombre d’articles publiés) sur 2 ans : 5.305
* Facteur d’impact sur 5 ans : 5.862
* Indice d’immédiateté : 0.843
* Note d’influence de l’article : 2.597
* Eigenfactor® score: 0.00855

### Contenu du data paper

Concernant la structure, la partie la plus développée dans ce data paper est la partie sur les méthodes. Le data paper comprend au total 7 figures et 2 tableaux.

L’abstract se compose d’une partie rédigée décrivant le contenu et la structure du jeu de données ainsi que le tableau récapitulatif décrit précédemment.

La partie « background and summary » est en fait une sorte d’introduction où les auteurs présentent le contexte, définissent les mots clés et présentent l’intérêt des légumineuses ainsi que quelques chiffres clés.  
Cette partie démontre également l’intérêt d’un tel jeu de données et le décrit brièvement, notamment à l’aide de figures comprenant une carte et des tableaux récapitulatifs. Les possibilités de réutilisation du jeu de données sont abordées par la suite ainsi que les éventuelles possibilités de le compléter.

La troisième partie du data descriptor concerne les méthodes. Les données publiées étant issues de la littérature, les auteurs indiquent le type de journaux dans lesquels ils ont effectués leurs recherches ainsi que la démarche effectuée pour arriver à leurs résultats.   
Cette partie permet également de décrire l’organisation de la base de données ainsi que d’expliquer certaines valeurs pouvant susciter des interrogations telle que la valeur « NULL ». De plus, ils abordent les sources d’erreurs possibles liées à leur démarche et en donnent une estimation chiffrée.  
Enfin, les fichiers supplémentaires disponibles sont décrits et localisés.

La partie suivante du data descriptor concerne les données en tant que telles. Ici, les auteurs ont décrits les attributs principaux de la table contenant le plus de données expérimentales extraites de la littérature. Les attributs des autres tables sont décrits dans un fichier annexe disponible au téléchargement.  
Cette description inclue la présentation des calculs effectués, la justification des choix pris au vue des valeurs ainsi que la façon de procéder lorsque certaines valeurs manquaient.   
Cette partie comprend également des figures résumant la distribution de plusieurs variables pour un grand nombre d’espèces de légumineuses afin de mieux rendre compte de la diversité du jeu de données.

La cinquième partie de cet article est consacrée à la validation technique des données. Elle sert à exposer les mécanismes mis en œuvre afin de s’assurer de la véracité des données publiées.  
Dans le cas étudié, les auteurs évoquent par exemple que les articles sélectionnés dans la littérature ont été relus plusieurs fois, les données reportées dans les tables ont elles aussi été vérifiées à plusieurs reprises et des requêtes ont été mise en place afin de s’assurer qu’aucune aberration n’en ressortait. Afin de s’affranchir des doutes sur la valeur de certaines données, 17 auteurs ont été contactés.

La partie suivante porte sur les possibilités de réutilisation des données. Les auteurs rappellent brièvement la source de leurs données et proposent plusieurs exemples de réutilisation détaillés avec la marche à suivre. Cette partie permet également d’aborder l’ajout de nouvelles données à la base de données. Pour cela les auteurs indiquent plusieurs possibilités de données à intégrer à leur base en détaillant les solutions techniques. D’autre part un exemple de requête SQL est fourni ce qui permet de faciliter l’utilisation d’une telle base de données par un individu quelconque.

La dernière partie obligatoire du data paper porte sur les références. Ici, les auteurs en donnent 188 (sachant que leur base de données a été créée à partir de 173 articles).  
A la fin de l’article se trouve une partie «  informations supplémentaires » qui permet notamment de donner la licence sous laquelle les données sont publiées.

## 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 data paper est issu du journal Data in Brief proposé par le groupe ELSEVIER permettant 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 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 (dollar américain) soit environ 450 euros 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 qui permet d’attribuer des droits d’édition et de distribution à l’éditeur tout en accordant la conservation des droits d’auteurs.

Les chiffres clés présentés par le site sont les suivants :

* Taux de citation : 0.7
* SNIP (Source Normalized Impact per Paper) : 0.287
* SJR (SCImago Journal Rank) : 0.341
* Taux d’acceptation : 77%

Pour publier un article dans Data in Brief, les données doivent avoir été produites et appartenir à l’auteur ou à son institution.

Le structure d’un data paper publié dans Data in Brief s’articule de la manière suivante :

* Titre, auteurs et labo
* Abstract : purement descriptif, il donne la localisation des données et se doit de citer l’article scientifique lié aux données le cas échéant.
* 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.* |

Tableau 1 : Indications sur le tableau des spécifications, (ELSEVIER, 2018)

* 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).
* Plan expérimental, matériels et méthodes : description complète de la manière dont ont été obtenues les données.
* (Remerciements)
* Références

### Contenu

Tout comme l’exemple précédent, la partie principale de ce data paper est celle portant sur le matériel et méthodes. Il comprend au total 2 tableaux. D’autres issus du même journal ne contiennent pas toujours de figures/tableaux mais certains sont plus fournis avec par exemple 2 tableaux et 3 figures (Oyeyemi et al., 2018).

L’abstract répond aux exigences du journal et permet également dans cet exemple de citer la licence utilisée.

Mis à part la ligne concernant les facteurs expérimentaux, le tableau des spécifications est rempli selon les indications du journal.

Les points concernant la valeur des données décrivent les différentes réutilisations possibles et en expliquent les intérêts. Une phrase supplémentaire indique que les auteurs sont ouverts à toute éventuelle collaboration.

La partie sur les données comprend une phrase de contexte avant de décrire les données. Les données supplémentaires sont également décrites. Cette partie est essentiellement composée de 2 tableaux descriptifs du jeu de données.

Le matériel et méthode fait référence à l’article associé au data paper où les données et calculs y sont déjà largement décrits. Les lignes suivantes servent à citer la source des données brutes ainsi qu’à expliquer les démarches suivies concernant le traitement des données.

Vient ensuite les remerciements, les liens vers les données supplémentaires ainsi que les références qui sont au nombre de 10 dans cet exemple. Ce chiffre est largement inférieur à celui du document précédent ce qui peut être expliqué par la provenance des données.

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

Cet article est également issu de Scientific data (Nature) mais se différencie du premier exemple car les données ne sont pas organisée en base de données ni issues de recherches dans la littérature. Ces données sont issues d’une expérimentation à long terme sur l’évolution des températures dans les sols. En termes de figures et de tableaux, ce data record en contient 4 de chaque. Les parties sont similaires au premier exemple, toutefois quelques points différent.  
En effet, la partie « background and summary » ne traite pas des possibilités de réutilisation du jeu de données.  
La partie sur les méthodes décrit d’abord le matériel utilisé, expose les méthodes mises en œuvre afin d’éliminer les valeurs aberrantes, les méthodes de contrôle de cohérence, le traitement des données et les méthodes de traitement des valeurs manquantes.

La partie suivante sur le jeu de données est assez brève et sert à détailler le contenu des fichiers disponibles. Cette partie comporte une figure représentant les valeurs moyennes des températures journalières au cours des années 2001 à 2016 pour les 4 compartiments étudiés (à savoir l’air, la surface organique du sol, la surface minérale du sol à 10 et 20cm de profondeur).

Afin de s’assurer de l’exactitude des données publiées, la validation technique comprend la calibration et la vérification des appareils, la vérification de la corrélation des capteurs entre eux ainsi que la vérification qu’il n’y ait pas de biais du au nombre de réplications. De plus, les maximum et minimum de température journaliers sont été comparées avec ceux obtenus dans une station voisines afin d’éviter toute valeur erronée.

La partie sur la réutilisation des données propose des exemples de réutilisation comme c’est le cas dans le premier article étudié. Toutefois, les solutions techniques ne sont pas détaillées. Cette partie permet également d’énoncer les limites du jeu de données.

L’article comporte 39 références et le type de licence est également donné dans la partie « informations supplémentaires ».

Le tableau ci-dessous (*Tableau 2*) propose une comparaison générale des deux revues étudiées dans ces études de cas (à savoir Data in Brief et Scientif data).

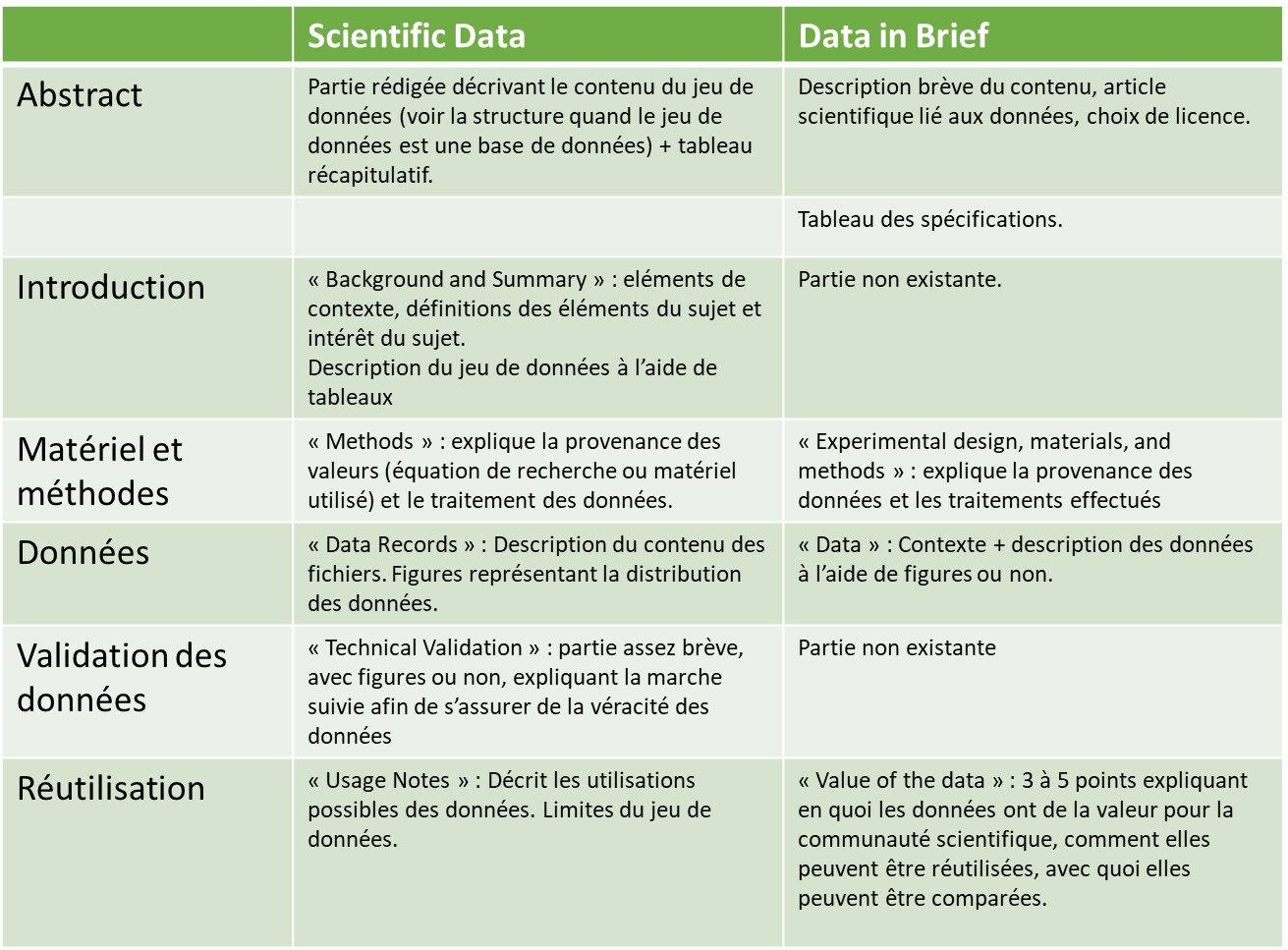


Tableau 2 : Comparaison du contenu attendu pour les journaux Scientific Data et Data in Brief

##### Conclusion :

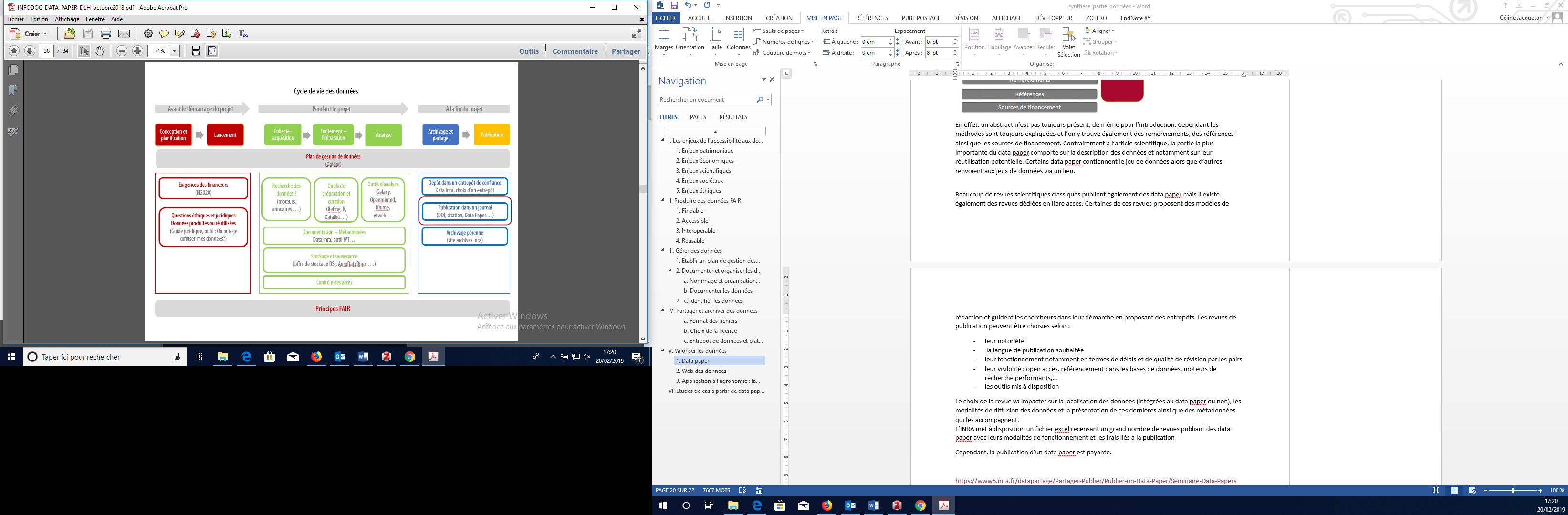
Afin de s’inscrire dans une démarche Open science et donc de produire des données FAIR, la première étape indispensable est l’établissement d’un plan de gestion des données qui décrira les grandes lignes des différentes étapes de leur cycle de vie. Dès la collecte des données, il est nécessaire de définir les données pouvant être partagées de celles devant rester confidentielles. Le nommage et l’organisation des fichiers ainsi que la documentation des données qu’ils contiennent à l’aide de métadonnées, idéalement issues de standards, doivent également être réalisés dès la collecte. Il est nécessaire que les fichiers soient dans un format le plus pérenne possible adapté aux données qu’ils contiennent. Tout au long de la collecte et du traitement des données, les fichiers doivent être stockés et sauvegardés dans un entrepôt de stockage par exemple. A la fin du projet, les données doivent être archivées de manière pérenne dans un entrepôt prévu à cet effet ou une plateforme d’archivage. Il est possible de valoriser son travail en publiant ses données dans un journal ou encore sur le web des données (annexe 2).

Concernant notre projet, les données sont déjà acquises et tout le travail de gestion repose donc sur la collecte des informations nécessaires. Les fichiers devront être nommés et organisés selon les principales règles énoncées dans les parties correspondantes. Les métadonnées seront décrites à l’aide de standards tel que celui établit par Dublin Core. Toutefois, ces standards sont assez généraux. Afin d’augmenter le potentiel de réutilisation des données, il sera demandé aux contributeurs au projet de fournir des informations sur les méthodes utilisées pour l’acquisition de leurs données ainsi que sur le matériel. Idéalement nous utiliserons du vocabulaire contrôlé ou des thésaurus provenant de l’AgroPortal. Il existe notamment un thésaurus : « anaeeThes » qui pourrait correspondre. L’organisation des données et donc par conséquent le format de fichier restent encore à définir. Les fichiers seront à minima des fichiers Excel en format csv mais les données pourront également, de façon plus ambitieuse, être déposées dans une base de données ouvertes au langage SQL. Le choix de la licence et des entrepôts de stockage et d’archivage n’est pas encore établit à ce stade, bien que DataInra semble être en terme d’entrepôt la solution la plus simple et la plus adéquate au projet. Enfin, concernant la valorisation des données, celles-ci seront publiées dans un data paper, l’objectif étant de viser le journal Scientific Data en premier lieu. La publication sur le web des données semble être un peu ambitieuse au vue de nos compétences en la matière. Toutefois, ces données seront doublement valorisées car une partie d’entre elles serviront à calibrer les modèles de décompositions des résidus végétaux disponibles sur la plateforme Virtual Soil. Une collaboration avec la communauté STICS est également envisagée en termes de modélisation. Cependant, la communauté n’étant pas Open Science, cette collaboration passera au second plan.

##### Annexes

|  |  |
| --- | --- |
| 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)

  
Annexe 2 : Schéma récapitulatif du cycle de vie des données FAIR – selon (Pôle Formation IST, 2018)