Gestion des Données Ouvertes en Agriculture et Nutrition

Ce cours en ligne est le fruit d'une collaboration entre les partenaires de GODAN Action, y compris Wageningen Environnemental Research (WUR), AgroKnow, AidData, l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO), le Forum Mondial sur la Recherche Agricole (GFAR), l'Institut des Etudes du Développement (IDS), le Land Portal, l'Open Data Institute (ODI) et le Centre Technique de Coopération Agricole et Rurale (CTA).

GODAN ACTION
Global Open Data for Agriculture & Nutrition

UKaid
from the British people

GODAN Action est un projet de trois ans du Département pour le Développement International du Royaume-Uni pour permettre aux utilisateurs, producteurs et intermédiaires de données de s'engager efficacement avec les données ouvertes et maximiser leur potentiel d'impact dans les secteurs de l'agriculture et de l'alimentation. Nous travaillons en particulier à renforcer les capacités, à promouvoir des normes communes et les meilleures pratiques et à améliorer la manière dont nous mesurons l'impact. [www.godan.info]

MODULE 3 : RENDRE LES DONNÉES OUVERTES

LEÇON 3 .1: Gestion des jeux de données

Photo par Neil Palmer (CIAT) Sous licence CC BY -SA 2.0

Objectifs et résultats d'apprentissage

Cette leçon a pour objectif :

- D'expliquer les principes de base de la gestion des données
- D'introduire le processus de préparation d'un plan de gestion des données
- D'expliquer les concepts sur le stockage des données, le versioning et les
  - pratiques de documentation.

À la fin de cette leçon, vous devriez être en mesure de :

- Déterminer les étapes à suivre pour s'assurer que de bonnes pratiques de gestion des données scientifiques sont suivies
- Préparer un plan de gestion des données
- Comprendre le potentiel de l'utilisation de bonnes pratiques en matière De stockage de données et de gestion des versions

Sommaire

Liste des tableaux

1. Gestion des données

De nombreux jeux de données sont fondamentalement éphémères : les données du marché, les données de production et de consommation, et les données météorologiques en sont de bons exemples. Ces données n'ont aucun sens si nous ne connaissons pas le délai qui s'y rattache et si elles ne sont pas mises à jour régulièrement. Les données sur les sols, par exemple, sont similaires, même si elles ne sont pas mises à jour aussi souvent que les données météorologiques et de marché. Elles changent selon l'endroit et au cours des saisons et entre elles.

Afin d'établir et de maintenir la confiance dans des données ouvertes comme celles-ci et d'autres, il est nécessaire de mettre en place des principes et des pratiques stables de gestion des données. De bons principes de gestion des données aident à garantir que les données produites ou utilisées sont enregistrées, stockées, rendues accessibles pour utilisation et réutilisées, gérées dans le temps et/ou éliminées, conformément aux exigences légales, éthiques, des bailleurs de fonds et aux bonnes pratiques. Pour les consommateurs de données ouvertes, la confiance dépend de nombreux facteurs :

- *La connaissance de la source*. La confiance dans les données commence par la connaissance de leur source.
- *La confiance en la source*. Si vous savez que les données proviennent d'une source fiable, vous pouvez vous y fier et tirer les conclusions qui s'imposent.
- *La mise à jour des données*. Même lorsqu'elles proviennent d'une source fiable, les données ne sont pas utiles si elles ne sont pas à jour.
- *La qualité des données*. Les données fiables doivent refléter avec exactitude et précision ce qu'elles mesurent.
- *La pérennité*. Un ensemble de données fiable doit avoir une certaine garantie de disponibilité.
- *L'accessibilité*. Comme les documents, les données ne sont utiles que si elles sont faciles à trouver. (Pour en savoir plus sur les possibilités de d'accessibilité, voir le module 1, leçon 2.1)
- *Documentation et assistance*. Les consommateurs devraient pouvoir avoir accès à une assistance pour les données si nécessaires.
- *Interaction*. Les consommateurs devraient être en mesure de fournir une rétroaction s'il y a un problème avec les données.

La gestion des données est donc un processus qui comprend un large éventail d'activités allant des aspects administratifs aux aspects techniques du traitement des données d'une manière qui tient compte des facteurs énumérés ci-dessus. Une bonne politique de gestion des données définira des objectifs stratégiques à long terme pour la gestion des données dans tous les aspects d'un projet ou d'une entreprise.

Une politique de gestion des données est un ensemble de principes de haut niveau qui établissent un cadre directeur pour la gestion des données. Une

politique de gestion des données peut être utilisée pour aborder des questions stratégiques telles que l'accès aux données, les questions juridiques pertinentes, les questions de gérance et de garde des données, l'acquisition de données et autres questions. Comme elle fournit un cadre de haut niveau, la politique de gestion des données devrait être souple et dynamique. Cela lui permet de s'adapter facilement aux défis imprévus, aux différents types de projets et aux partenariats potentiellement opportunistes tout en maintenant son orientation stratégique. La politique de gestion des données aidera à informer et à élaborer un plan de gestion des données, qui sera discuté plus en détail dans cette leçon.

Afin d'atteindre les objectifs et les normes de gestion des données, toutes les parties concernées doivent comprendre leurs rôles et responsabilités connexes. Les objectifs de la délimitation des rôles et des responsabilités en matière de gestion des données sont les suivants :

- Définir clairement les rôles associés aux fonctions
- Établir la propriété des données tout au long de toutes les phases d'un projet
- instiller la responsabilité des données
- Veiller à ce que des mesures adéquates et convenues de la qualité des données et des métadonnées soient tenues à jour de façon continue.

La qualité appliquée aux données a été définie comme l'aptitude à l'emploi ou l'utilisation potentielle. De nombreux principes de qualité des données s'appliquent lorsqu'il s'agit des données sur les espèces et des aspects spatiaux de ces données. Ces principes s'appliquent à toutes les étapes du processus de gestion des données, en commençant par la collecte et la saisie des données. Une perte de qualité des données à l'une ou l'autre de ces étapes réduit l'applicabilité et les utilisations auxquelles les données peuvent être adéquatement affectées.

Tous ces facteurs influent sur la qualité finale ou l'aptitude à l'utilisation des données et s'appliquent à tous les aspects des données. Des normes de qualité des données peuvent être disponibles pour :

- la justesse
- la précision
- la résolution
- la fiabilité
- la répétabilité
- la reproductibilité
- la devise
- la pertinence
- la capacité de vérification
- l'intégralité
- la mise à jour

La qualité des données est évaluée en appliquant des procédures de vérification et de validation dans le cadre du processus de contrôle de la qualité. La vérification et la validation sont des éléments importants de la gestion des données qui aident à assurer la validité et la fiabilité des données. Ceci est développé en détail dans le module 2, leçon 2.2.

Certains des principes de gestion des données décrits dans cette unité s'inspirent également des principes de la gestion des données de recherche (GDR). De bons principes de GDR peuvent être appliqués avec succès à des initiatives de données ouvertes, garantissant ainsi :

- Une visibilité et une découverte accrues des ensembles de données
- La facilitation de la pérennité, du partage et de la réutilisation des ensembles de données
- La compréhension qu'ont les utilisateurs de données du contexte du contenu et des limites des ensembles de données
- La réduction du risque de perte de données en les gardant en sûreté et en les sécurisant
- L'interopérabilité des ensembles de données et l'échange des données
- La conformité aux attentes et aux politiques des bailleurs et/ou de l'institution
- L'offre d'occasions de collaboration avec d'autres personnes susceptibles d'utiliser les données

1.1. La diversité des données

Les données peuvent être considérées comme le niveau le plus bas d'abstraction à partir duquel l'information et les connaissances sont dérivées. Les données peuvent également être considérées comme le niveau auquel les mesures ont été recueillies à l'origine, par exemple les réponses individuelles à une enquête ou à un recensement, les mesures horaires de la température, de la vitesse et de la direction du vent, le nombre et le prix des actions négociées dans chaque transaction d'achat/de vente d'actions, etc.

Cependant, le mot " données " n'a pas le même sens pour tout le monde et dans des contextes différents. Différentes disciplines ont et utilisent un langage spécifique à la discipline autour des données de recherche sur le sujet[1].

Tableau 1 Différents types de données rencontrées dans différents contextes et disciplines

[1] http://mantra.edina.ac.uk

Types de données		
Sens Général	Sciences sociales	Données physiques/agricoles
● images ● vidéocassette ● cartographie/données SIG ● mesures numériques	● réponses au sondage ● Transcriptions des groupes de discussion et des entretiens individuels ● indicateurs économiques ● données démographiques ● sondages d'opinion	● Mesures générées par des capteurs/ instruments de laboratoire ● modélisation informatique ● simulations ● Observations et/ou études sur le terrain ● spécimen

Les données peuvent donc être générées à des fins différentes et par des processus différents dans une multitude de formats numériques. La classification suivante a été établie par le Réseau d'Information pour la Recherche :

- **Observationnelles :** données capturées en temps réel, généralement uniques et irremplaçables, par exemple les images du cerveau, les données d'enquête
- **Expérimentales :** les données provenant de résultats expérimentaux, par exemple à partir d'équipement de laboratoire, souvent reproductibles, mais qui peuvent être coûteuses, par exemple des chromatogrammes, des micro-essais
- **Simulation :** les données générées à partir de modèles d'essai où le modèle et les métadonnées peuvent être plus importants que les données de production du modèle, par exemple les modèles économiques ou climatiques
- **Dérivées ou compilées :** résultant du traitement ou de la combinaison de données "brutes", souvent reproductibles mais coûteuses, comme les bases de données compilées, le text mining, les données de recensement agrégées, etc.
- **Référentielles ou canoniques :** un conglomérat (statique ou organique) ou une collection de plus petits ensembles de données (examinés par des pairs), très probablement publiés et curés, par exemple des banques de données génétiques, des bases de données cristallographiques.

Dans la section suivante, nous décrirons les considérations de base en matière de gestion des données à prendre en compte au cours du cycle de vie des données, c'est-à-dire depuis leur création et leur stockage initial jusqu'au moment où elles deviennent obsolètes et sont supprimées.

1.2. Métadonnées et documentation

Tous les ensembles de données devraient être identifiés et documentés afin de faciliter leur identification ultérieure, leur bonne gestion et leur utilisation efficace, et d'éviter de recueillir les mêmes données plus d'une fois. Pour fournir une liste précise des ensembles de données détenus par une organisation, un catalogue de données devrait être compilé. Il s'agit d'une collection de métadonnées découvrables pour chaque ensemble de données, sous une forme permettant aux utilisateurs de s'y référer. Ces métadonnées devraient fournir des informations sur le contenu, l'étendue géographique, l'actualité et l'accessibilité des données, ainsi que les coordonnées des personnes à contacter pour obtenir des informations complémentaires.

Tous les ensembles de données, une fois catalogués, devraient également être documentés sous une forme détaillée à laquelle les utilisateurs peuvent se référer lorsqu'ils utilisent les données. Ces métadonnées détaillées devraient décrire le contenu, les caractéristiques et l'utilisation de l'ensemble des données, en utilisant un modèle standard de métadonnées détaillées.

Les métadonnées

Les métadonnées, ou " données sur les données ", expliquent votre ensemble de données et vous permettent de documenter des renseignements importants pour les fins suivantes :

- Retrouver les données plus tard
- Comprendre ce que sont les données plus tard
- Partager des données (tant avec les collaborateurs qu'avec les futurs utilisateurs de données secondaires).

Cela devrait être considéré comme un investissement en temps qui vous évitera bien des ennuis par la suite.

Exemples

- Dublin Core
- Darwin Core
- FGDC (Comité fédéral des données géographiques)
- DDI (Data Documentation Initiative)
- ABCD (Accès aux données des collections biologiques)
- CSDGM (Norme de contenu pour les métadonnées géospatiales).

Une distinction qui est souvent citée lorsqu'il s'agit de gestion des données est celle des données par rapport aux métadonnées (c'est-à-dire les données sur les données). Il y a un certain nombre de distinctions spécifiques auxquelles celles-ci pourraient se référer :

- *Les métadonnées comme schéma*. Lorsque nous recueillons des données tabulaires, nous devons savoir à quoi font référence les " colonnes " dans les données. Même dans les données non tabulaires, certaines informations de schéma sont utiles pour interpréter les données. De nombreux formats de données incluent des façons de spécifier les métadonnées de schéma, par exemple XSD pour XML, RDFS pour RDF, et DDL pour les bases de données.
- *Les métadonnées bibliographiques*. Les bibliothécaires et les documentalistes utilisent les métadonnées pour décrire des documents (livres, articles, images, etc.) depuis des siècles et ont déterminé des structures d'enregistrement et de recherche pour ce type de métadonnées. Ce type de métadonnées comprend l'information sur la provenance (auteur, données de publication), les dates, la taille de la publication (par exemple le nombre de pages) et s'applique également aux ensembles de données (par exemple DCAT27 et VoID28).
- *Le vocabulaire partagé*. L'alignement de différents ensembles de données est un défi dans tout environnement distribué. L'utilisation d'un vocabulaire commun est un outil clé pour la gestion de ces ensembles de données. Le vocabulaire est utilisé dans le contenu des données, plutôt que de décrire les données en soi. Par exemple, l'AGROVOC (pour AGRiculture VOCabulary)[2] propose (entre autres choses) la terminologie à utiliser pour se référer aux produits agricoles, par exemple lait, produits dérivés du lait, gras, etc. Agroportal[3] et le registre VEST[4] donnent accès à un lexique important lié à l'agriculture.

Le contenu du fichier

Pour que d'autres puissent utiliser les données, ils doivent pouvoir comprendre le contenu du jeu de données, y compris les noms des paramètres utilisés, les unités de mesure, les formats et les définitions des valeurs codées. Au début du fichier, inclure dans l'en-tête des descripteurs pour relier le fichier de données à l'ensemble de données (par exemple, le nom du fichier de données, le titre de l'ensemble de données, l'auteur, la date du jour, la date de la dernière modification des données du fichier et les noms de fichiers correspondants). Les autres lignes d'en-tête doivent décrire le contenu de chaque colonne, y compris une ligne pour les noms des paramètres et une pour les unités de paramètres. Dans le cas des ensembles de données qui sont vastes et complexes et qui peuvent nécessiter beaucoup d'information descriptive sur le

[2] FAO. 2016. *AGROVOC Multilingual agricultural thesaurus*. Available at: http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus.
[3] IBC Agroportal. Available at: http://agroportal.lirmm.fr/
[4] VEST Directory, Agricultural Information Management Standards (AIMS). http://aims.fao.org/vest-registry

contenu des ensembles de données, cette information peut être fournie dans un document distinct lié plutôt que sous forme d'en-têtes dans le fichier de données lui-même.

- *Les paramètres.* Les paramètres rapportés dans les ensembles de données doivent avoir des noms qui décrivent leur contenu, et leurs unités doivent être définies pour que les autres comprennent ce qui est rapporté. Utilisez des noms de paramètres communément acceptés. Un bon nom est court, unique (au moins dans un ensemble de données déterminé) et descriptif du contenu des paramètres. Les en-têtes de colonne doivent être construits de manière à pouvoir être facilement importés par divers systèmes de données. Utilisez une majuscule cohérente et n'utilisez que des lettres, des chiffres et des traits de soulignement - sans espaces ni décimales - dans le nom du paramètre. Choisissez un format cohérent pour chaque paramètre et utilisez ce format pour l'ensemble des données. Dans la mesure du possible, essayez d'utiliser des formats normalisés, tels que ceux utilisés pour les dates, les heures et les coordonnées spatiales.

Toutes les cellules de chaque colonne ne doivent contenir qu'un seul type d'information (par exemple, du texte, des chiffres, etc.). Les types de données les plus courants sont les suivants : texte, numérique, date/heure, booléen (Oui/Non ou Vrai/Faux) et commentaires, utilisés pour stocker de grandes quantités de texte.

- *Les champs codés*. Les champs codés, par opposition aux champs en texte libre, comportent souvent des listes normalisées de valeurs prédéfinies parmi lesquelles le fournisseur de données peut choisir. Les collecteurs de données peuvent établir leurs propres champs codés avec des valeurs définies à utiliser uniformément dans plusieurs fichiers de données. Les champs codés sont plus efficaces pour le stockage et la récupération des données que les champs en texte libre.

- *Les valeurs manquantes. Il existe plusieurs options pour traiter une valeur manquante. La première est de laisser la valeur vide, mais cela pose un problème car certains logiciels ne font pas la différence entre un blanc et un zéro, ou un utilisateur pourrait se demander si le fournisseur de données a accidentellement sauté une colonne. Une autre option est de mettre un point où le nombre irait. Cela indique clairement qu'une valeur doit être présente, bien qu'elle ne précise pas pourquoi les données sont manquantes. Une dernière option consiste à utiliser des codes différents pour indiquer les différentes raisons pour lesquelles les données sont manquantes.*

1.3. Sécurité et stockage

Le partage efficace des données dépend d'un solide réseau de confiance entre les fournisseurs de données et les consommateurs. L'infrastructure pour le partage des données ne sera pas utilisée si les parties qui fournissent et utilisent les données ne font pas confiance à l'infrastructure ou à l'autre partie. Si des données sensibles doivent être partagées, il doit y avoir des dispositions dans la plate-forme pour assurer la sécurité de ces données. Que les données soient fermées ou partagées avec des individus ou des organisations spécifiques, elles devront être stockées d'une manière contrôlée. En fonction de la sensibilité des données, cela comprendra une certaine garantie de sécurité, par exemple contre le piratage informatique. Dans les cas les plus extrêmes, les exigences de sécurité pour les données partagées en agriculture pourraient être aussi sévères que pour les données partagées dans l'armée. Ces principes ne sont pas propres aux données agricoles et ont été étudiés en profondeur.

Les concepts de base qui sous-tendent ces principes sont que les services devraient être difficiles à compromettre, qu'une interférence devrait être facile à détecter et que l'impact d'un tel scénario peut être limité. Pour les données ouvertes, cette préoccupation est beaucoup moins importante, mais pour établir la confiance chez les fournisseurs de données, un certain soutien à la sécurité des données doit être en place.

Voici quelques considérations importantes concernant les aspects physiques du stockage et l'archivage des ensembles de données électroniques/numériques :

- *Le matériel et les logiciels du serveur*. Quel type de base de données sera nécessaire pour les données ? Faudra-t-il mettre en place une infrastructure physique ou l'infrastructure est-elle déjà en place ? Une importante base de données sera-t-elle nécessaire ? (Voir la leçon 3.3 sur la création de dépôts de données ouvertes pour les plates-formes logicielles.) Qui supervisera l'administration de ce système?

- *Infrastructure de réseau.* Pour les données ouvertes, la base de données doit être connectée à Internet pour être accessible. Quelle quantité de bande passante est nécessaire pour desservir le public cible ? À quelles heures de la journée doit-elle être accessible ?

- *Taille et format des ensembles de données.* La taille d'un ensemble de données doit être estimée de façon à ce que l'espace de stockage puisse être correctement pris en compte. Les types et les formats doivent être identifiés de manière à ce qu'il n'y ait pas de surprises sur le plan des capacités de la base de données et de la compatibilité.

- *La maintenance et la mise à jour de la base de données*. Une base de données ou un ensemble de données devrait avoir des procédures de mise à jour soigneusement définies. Si un ensemble de données est en ligne ou en cours, il comprendra des ajouts, des modifications et des suppressions, ainsi que la fréquence des mises à jour. Le versioning sera extrêmement important lorsque vous travaillez dans un environnement multi-utilisateurs.

- *Les exigences en matière de sauvegarde et de restauration des bases de données.* Les exigences relatives à la sauvegarde ou à la restauration d'une base de données en cas d'erreur de l'utilisateur, de défaillance d'un logiciel ou d'un support, ou de catastrophe, devraient être clairement définies et convenues pour assurer la pérennité des bases de données. Les mécanismes, les calendriers, la fréquence et les types de sauvegardes, ainsi que les plans de reprise appropriés devraient être spécifiés et planifiés. Cela peut inclure les types de supports de stockage pour les sauvegardes sur site et la nécessité d'une sauvegarde hors site.

1.4. Accès aux données et leur diffusion

La production de données est une chose, leur diffusion en est une autre. Les données ouvertes sont utiles lorsqu'elles peuvent être livrées entre de bonnes mains (ou sur la bonne machine) et dans un contexte où elles peuvent être très utiles. Dans certains cas, il peut s'agir d'un laboratoire de recherche sur l'efficacité d'un traitement (par exemple, l'efficacité d'un herbicide pour traiter certaines infestations). Parfois, les données doivent être livrées sur le terrain pour aider les petits exploitants à prendre des décisions éclairées sur les variétés de cultures à cultiver ou les traitements à appliquer. Il doit y avoir une variété de canaux de diffusion de données, adaptés à chaque situation. La "mise au point" des canaux de diffusion des données peut devenir une opportunité commerciale pour les intermédiaires de données dans la mesure où les données sont totalement "ouvertes". Un intermédiaire peut fournir des services pour personnaliser la livraison des données pour la vaste gamme de clients qui pourraient exister pour les données. Les données ouvertes créent la possibilité d'un marché, où multiples sources de données pertinentes sont disponibles.

Pour être disponibles, les données doivent être stockées de manière à être accessibles. Même à l'ère moderne du déploiement de cloud, les données et les applications sont stockées quelque part sur du matériel, même si celui-ci est virtualisé. Une stratégie de partage des données à l'échelle mondiale doit préciser où elles seront stockées et quels accords de niveau de service (SLA) seront maintenus (temps de disponibilité, débit, contrôles d'accès, etc.).

Les éléments suivants devraient être fortement pris en compte au moment de décider de la façon de diffuser les données :

- L'accès aux données devrait être assuré conformément à la politique de l'organisation en matière de données et aux législations/lois nationales sur l'accès à l'information
- L'accès aux données devrait être accordé sans porter atteinte aux droits d'auteur, à la propriété intellectuelle des données ou à toute obligation statutaire ou ministérielle en vigueur.

2. Plans de gestion des données

Les bailleurs exigent de plus en plus l'élaboration de plans de gestion des données comme condition de financement. [5] Les exigences visent habituellement à permettre le partage des résultats des projets ou de la recherche, y compris les données (et les publications). Par exemple, en décembre 2013, l'UE a publié de nouvelles lignes directrices sur la gestion des données, dans le projet de recherche Horizon 2020: ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Le processus de planification de la gestion des données commence par une liste de contrôle. Une liste de contrôle aidera plus tard à l'élaboration d'un plan de gestion des données. Cette liste de contrôle pourrait comprendre les éléments suivants, en tout ou en partie :

- Quelles données allez-vous collecter ou créer, comment seront-elles créées et dans quel but ?
- Comment gérerez-vous les questions d'éthique ? Comment allez-vous gérer les questions de droits d'auteur et de propriété intellectuelle ?
- Quels formats de fichiers seront utilisés ? Sont-ils non exclusifs, transparents et durables ? Quelles conventions de dénomination des répertoires et des fichiers seront utilisées ? Y a-t-il des normes officielles que vous adopterez ? Quels documents et métadonnées accompagneront les données?
- Comment les données seront-elles stockées et sauvegardées ? Comment allez-vous gérer l'accès et la sécurité ? Qui sera responsable de la gestion des données?
- Y a-t-il des procédures existantes sur lesquelles vous baserez votre approche ? Par exemple, y a-t-il des politiques institutionnelles de

Commented [SS1]:

[5] https://ocw.mit.edu/resources/res-str-002-data-management-spring-2016/workshopmaterials/MITRES_STR002S16_IntroDM.pdf

protection

ou de sécurité des données à suivre, des lignes directrices de gestion des données du service ou du groupe, définies par votre institution ou votre bailleur de fonds, dont il faut tenir compte ?

- Quel est le plan de préservation à long terme de l'ensemble de données
- ? Par exemple, quelles données devraient être conservées, partagées
- et/ou préservées ? Comment allez-vous partager les données, et des
- restrictions sur le partage des données sont-elles nécessaires ?

- De quelles ressources aurez-vous besoin pour exécuter votre plan ? Par
- exemple, existe-t-il des outils ou des logiciels nécessaires pour créer, traiter
- ou visualiser les données ?

Selon la nature du projet, il est possible d'en ajouter ou d'en soustraire d'autres éléments à la liste proposée. Une liste de contrôle plus détaillée est disponible au Digital Curation Centre à l'adresse suivante : www.dcc.ac.uk/sites/default/files/documents/resource/DMP_Checklist_2013.pdf

Il est important de noter : Les plans de gestion des données doivent être tenus à jour en permanence tout au long d'un projet ou d'une recherche.

Table 2 Composantes d'un de gestion des données

Informations administratives	• Nom et identité du projet • Description du projet • Organisme(s) de financement • Personne-ressource pour les données du projet • Politiques connexes • Date de la première version • Date de la dernière mise à jour
Collecte des données	• Description des données, y compris le type, le format et le volume prévus • Réutilisation des ensembles de données existants • Méthodes de collecte ou de création des données • Structures, système de dénomination et de gestion des versions pour dossiers et fichiers • Processus d'assurance qualité

Documentation et métadonnées	● Une liste d'informations dont vous vous attendez à avoir besoin pour que les données soient lues et interprétées à l'avenir. ● Comment vous comptez recueillir ou créer cette documentation et ces métadonnées ● Les normes de métadonnées que vous utiliserez **Quelques exemples de documentation de données** : ● Cahiers de laboratoire et protocoles expérimentaux ● Questionnaires, livres de codes, dictionnaires de données ● Syntaxe du logiciel et fichiers de sortie ● Informations sur les réglages de l'équipement et l'étalonnage de l'instrument ● Schéma de base de données ● Rapports méthodologiques ● Informations de provenance sur les sources de données dérivées Ajoutez des descriptions détaillées pour les collections ou les dossiers, par exemple, ce qu'il y a dans un dossier, d'où il vient, comment il peut être récupéré si nécessaire, tout problème existant, etc.
Éthique et conformité aux lois	**Éthique** ● Détails du consentement requis pour la conservation et le partage des données ● Mesures à prendre, au besoin, pour protéger l'identité de tout participant ● Mesures à prendre, au besoin, pour s'assurer que les données sensibles sont stockées et transférées en toute sécurité **Droits d'auteur et droits de propriété intellectuelle** ● Nom(s) du (des) propriétaire(s) des données ● Licence(s) de réutilisation qui sera appliquée (par exemple une des licences disponibles auprès de Creative Commons ou Open Data Commons) ● Restrictions concernant l'utilisation par des tiers ● Tout retard prévu dans le partage des données, par exemple en attente d'une demande de brevet ou d'un embargo lié à la publication dans une revue.

Stockage et sauvegarde	● Le lieu où les données (physiques) seront stockées ● Disposition de sauvegarde ● Personne ou équipe responsable de la sauvegarde ● Procédures de recouvrement **Sécurité** ● Les risques et la façon dont ils seront gérés ● Modalités d'accès ● Toutes les dispositions nécessaires, le cas échéant, pour un transfert sûr et sécurisé des données recueillies sur le terrain
Sélection et conservation	● Détails sur les données à conserver, à partager et/ou à préserver, en particulier en ce qui concerne les exigences contractuelles, légales ou réglementaires ● Utilisations prévisibles des données pour la recherche ● La durée pendant laquelle les données seront (ou devraient être) conservées au-delà de la durée de vie du projet ● Le dépôt ou l'archive où les données seront conservées, ainsi que tous les frais connexes. ● Temps et efforts nécessaires pour préparer les données en vue de leur préservation et de leur partage
Responsabilités et ressources en matière de partage des données	● Personne désignée responsable de la mise en œuvre du plan de gestion des données ● Personne désignée responsable de chaque activité de gestion des données ● Matériel et logiciels requis (tout matériel qui s'ajoute à l'offre institutionnelle existante) ● Expertise ou formation spécialisée supplémentaire requise ● Frais requis par les entreprises de stockage de données

3. Organisation des données

Les fichiers et dossiers de données doivent être étiquetés et organisés de manière systématique afin d'être identifiables et accessibles aux utilisateurs actuels et futurs. Les avantages d'un étiquetage cohérent des fichiers de données sont les suivants :

- Les fichiers de données se différencient les uns des autres à l'intérieur du dossier les contenant
- Le choix du nom des fichiers de données évite la confusion lorsque plusieurs personnes travaillent sur des fichiers partagés.
- Les fichiers de données sont plus faciles à localiser et à parcourir
- Les fichiers de données peuvent être récupérés non seulement par leur créateur mais aussi par d'autres utilisateurs.
- Les fichiers de données peuvent être triés dans un ordre logique
- Les fichiers de données ne sont pas écrasés ou supprimés accidentellement
- Différentes versions des fichiers de données peuvent être identifiées
- Si les fichiers de données sont déplacés vers une autre plate-forme de stockage, leurs noms conserveront leur contexte utile.

Il y a trois principaux critères à prendre en considération en ce qui concerne la dénomination des fichiers et l'étiquetage, à savoir :

- *L'organisation :* elle est importante pour l'accès et la récupération futurs et doit tenir compte des contraintes de dénomination des fichiers du système dans lequel le fichier est stocké.
- *Le contexte* : il peut s'agir d'informations descriptives ou spécifiques au contenu, indépendamment de l'endroit où les données sont stockées
- *La cohérence :* choisissez une convention de dénomination et assurez-vous que les règles sont systématiquement suivies en incluant toujours les mêmes informations (comme la date et l'heure) dans le même ordre (par exemple, AAAAMMJJ).

Éléments communs d'une stratégie de dénomination des fichiers

- Numéro de version
- Date de création
- Nom du créateur
- Description du contenu
- Nom de l'équipe/service/unité associés aux données
- Date de publication
- Numéro de projet

Il est important de tenir compte de ce qui suit lors de l'attribution d'un nom aux fichiers :

- L'utilisation de noms de fichiers génériques qui peuvent entrer en conflit lorsqu'ils sont déplacés d'un emplacement à un autre. Assurez-vous que les noms de fichiers sont indépendants de leur emplacement.
- Les noms de fichiers doivent durer plus longtemps que le créateur du fichier qui a initialement nommé le fichier.
- Si le numéro de projet est limité à deux chiffres, vous ne pouvez avoir que quatre-vingt-dix-neuf projets.

Le contrôle de version

Il est important d'identifier et de distinguer les versions des ensembles de données de façon cohérente. Cela permet de s'assurer qu'il existe une piste de vérification claire permettant de suivre l'élaboration d'un ensemble de données et d'identifier les versions antérieures au besoin. Ainsi, vous devrez établir une méthode qui vous semble logique et qui vous indiquera la version de votre ensemble de données.

- Une forme courante pour exprimer les versions des fichiers de données consiste à utiliser des nombres ordinaux (1, 2, 3, etc.) pour les changements majeurs de version et des nombres décimaux pour les changements mineurs (p. ex., v1, v1.1, v2.6).
- Il faut éviter de confondre les étiquettes, p. ex. révision, finale, finale2, definitive_copy, car elles peuvent s'accumuler.
- Enregistrez TOUS les changements (mineurs et majeurs)
- Éliminer ou supprimer les versions obsolètes (tout en conservant la copie 'brute' originale)

- Utiliser une fonction de sauvegarde automatique (si disponible) plutôt que d'enregistrer ou d'archiver plusieurs versions.
- Activez le versioning ou le suivi dans les documents collaboratifs ou les utilitaires de stockage tels que Wikis, GoogleDocs, etc.
- Prévoyez l'utilisation d'un logiciel de contrôle de version, par exemple Subversion, TortoiseSVN.

Quelques exemples structurés de maintien du contrôle de version [nom du document] [numéro de version] [état : brouillon/final] :

- Jones_interview_July2010_V1_BROUILLON
- Lipid-analysis-rate-V2_définitive
- 2001_01_28_ILB_CS3_V6_AB_modifié

Résumé

De bons principes de gestion des données aident à garantir que les données produites ou utilisées sont enregistrées, stockées, rendues accessibles pour utilisation et réutilisées (le cas échéant), gérées dans le temps et/ou éliminées,

conformément aux exigences légales, éthiques, des bailleurs de fonds et aux bonnes pratiques.

La gestion des données est donc un processus qui comprend un large éventail d'activités allant des aspects administratifs aux aspects techniques du traitement des données d'une manière qui tient compte des facteurs énumérés ci-dessus. Une bonne politique de gestion des données définira des objectifs stratégiques à long terme dans tous les aspects d'un projet ou d'une entreprise.

Une politique de gestion des données est un ensemble de principes de haut niveau qui établissent un cadre directeur pour la gestion des données. Une politique de gestion des données peut être utilisée pour aborder des questions stratégiques telles que l'accès aux données, les questions juridiques pertinentes, les questions de gérance et de garde des données, l'acquisition de données et autres questions. Les plans de gestion des données doivent être tenus à jour en permanence tout au long d'un projet ou d'une recherche.

Les éléments d'un plan de gestion des données comprendront ce qui suit:

- information administrative
- méthodes de collecte des données et processus d'assurance de la qualité
- documentation et métadonnées
- l'éthique et la conformité aux lois
- stockage et sauvegarde
- sélection et conservation
- responsabilités et ressources en matière de partage des données

Lectures complémentaires

- Arms, C. R., Fleischhauer, C. and Murray, K. (2013). Sustainability of digital formats: planning for Library of Congress collections. Library of Congress, Washington DC, USA. Available at www.digitalpreservation.gov/formats
- Beagrie, N. and Houghton, J. (2014). *The value and impact of data sharing and curation - synthesis of three recent UK studies*. Jisc. Available at: repository.jisc.ac.uk/5568/1/iDF308_-
- Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf
- Charles Beagrie Ltd (2013). *Keeping research data safe: cost / benefit studies, tools, and methodologies focussing on long-lived data*. Available at: http://www.beagrie.com/krds.php (accessed 4 August 2014)
- Digital Curation Centre (DCC). (2010). *Data management plans*. Available at: http://www.dcc.ac.uk/resources/data-management-plans (accessed 4 August 2014)
- Drummond, C.G. (2009). Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning*

- *Workshop*, 26th ICML, Montreal, Canada. Available at: http://cogprints.org/7691
- European Commission (EC). (2017). *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon*
- *2020*. Available at:
- ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa
- _pilot/h2020-hi-oa-pilot-guide_en.pdf (Version 3.2, 21 March 2017)
- GODAN 2016 *A Global Data Ecosystem for Agriculture and Food*. GODAN, Wallingford, UK. Available at: http://www.godan.info/documents/dataecosystem-agriculture-and-food
- Jones, S. (2011). *How to Develop a Data Management and Sharing Plan*. DCC How-to Guides, Digital Curation Centre, Edinburgh, UK. Available at: www.dcc.ac.uk/resources/how-guides/develop-data-plan
- UK Data Archive (UKDA). *Plan to Share*. Available at: www.dataarchive.ac.uk/create-manage/planning-for-sharing (accessed 4 August 2014)