IST Agropolis
COLLECTION FICHES PRATIQUES

Focus sur :

l'*Open Access* et les Données de la Recherche

Données de la recherche (*Research data*) : « enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. » (OCDE, 2007).

Jeu de données (*Dataset*) : « agrégation, sous une forme lisible, de données brutes ou dérivées présentant une certaine *unité*, rassemblées pour former un ensemble cohérent. » (Gaillard, 2014).

Les données de la recherche dans le cadre de l'*Open Access*

ENJEUX : SCIENTIFIQUES, ÉCONOMIQUES, SOCIÉTAUX

- Répondre à des défis scientifiques complexes, des enjeux de société, qui supposent transdisciplinarité, collaboration des équipes, partage, ouverture et mutualisation des informations, des données et des compétences
- Accroître la visibilité, l'utilisation et l'impact de la recherche au sein et hors de la communauté scientifique
- Favoriser la participation des citoyens et de la société civile : libre accès pour tous aux connaissances
- Faire évoluer le système de l'édition scientifique en permettant aux chercheurs de se réapproprier la diffusion de leur production scientifique
- Satisfaire aux conditions de financement des bailleurs et justifier de l'utilisation des fonds publics
- Assurer la continuité de la recherche, en permettant la réutilisation des données de recherches antérieures, ainsi que la reproductibilité des expériences, le tout dans un souci d'économie, de retour sur investissement et d'innovation
- Assurer la sécurité des données de la recherche et leur archivage à long terme
- Prendre en compte la nécessité de l'interopérabilité des données

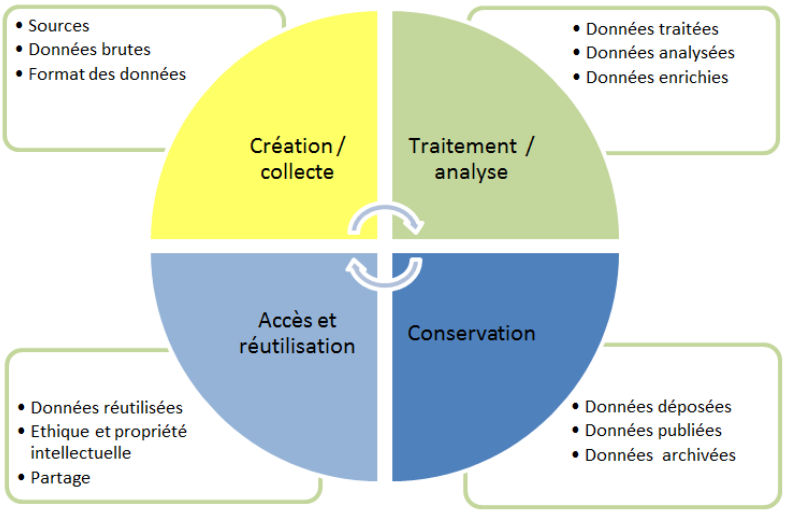
TYPOLOGIE DES DONNÉES DE LA RECHERCHE

Types de données	Exemples
Données d'observation, d'enquête : capturées ou collectées en temps réel, uniques et impossibles à reproduire	enquête niveau de vie de la Banque Mondiale, relevés de concentration en phytoplanctons…
Données expérimentales : obtenues à partir d'équipements de laboratoire, potentiellement reproductibles, parfois coûteuses	chromatogrammes, puces à ADN, cinétique chimique…
Données computationnelles ou de simulation : générées par des modèles informatiques ou de simulation, potentiellement reproductibles	modèle météorologique, modèle de simulations sismiques, modèle bio-économique…

CYCLE DE VIE DE LA DONNÉE

(RESEARCH DATA LIFECYCLE)

La représentation du cycle de vie des données de la recherche est une aide à la gestion des données. Ce cycle doit être décliné dans un Plan de Gestion de Données au début de tout projet nécessitant la création et/ou la manipulation de données. En effet, au-delà du choix de la plateforme, la gestion du cycle de vie des données est un enjeu majeur pour le stockage, la conservation, la pérennisation et la réutilisation de ces données.



Information Scientifique et Technique
AGROPOLIS INTERNATIONAL MONTPELLIER, FRANCE

Contact : cist@agropolis.fr –http://ist.agropolis.fr- Octobre 2016

Les étapes de gestion des données de la recherche

1. DECRIRE LES DONNÉES ET PRIVILÉGIER LES FORMATS NON PROPRIÉTAIRES

Le contexte de la production des données de recherche doit être documenté de manière précise et intelligible sous la forme d'un document et de métadonnées, précisant la paternité, le contenu des données, la méthodologie et les contraintes ou limites.

Métadonnées

Littéralement, une métadonnée est une donnée sur une donnée. Les métadonnées sont un ensemble structuré servant à décrire une ressource quel que soit son support.

2. SÉLECTIONNER LES DONNÉES À CONSERVER À LONG TERME

En s'appuyant sur les critères NERC par exemple.

3. ORGANISER ET STOCKER LES DONNÉES

Conserver les données mais aussi les métadonnées et les logiciels dans un entrepôt de confiance (*trusted repository*)

A minima : 3 copies sur 2 supports différents, dont 1 copie à distance.

Attribution d'un DOI - *Digital Object Identifier*

- Permet l'identification unique et pérenne d'un objet numérique et sa citation
- Facilite l'accès, le partage et la réutilisation des contenus ainsi que le lien entre publications et données
- *DataCite Metadata Store (MDS)* est la plateforme de création de DOI et d'enregistrement des métadonnées associées (https://mds.datacite.org)
- L'INIST-CNRS est l'agence officielle d'attribution de DOI en France

4. CHOISIR LES LICENCES POUR LA RÉUTILISATION DES DONNÉES

Ex. : licence *Creative Commons*, ODbL…

5. PARTAGER ET RÉUTILISER LES DONNÉES

Définir la période d'embargo, prendre en compte les exigences des financeurs, attribuer un identifiant pérenne pour les données de la recherche (DOI).

Le plan de gestion des données (PGD)

Le plan de gestion des données (ou *Data Management Plan* – DMP) est un document formalisé, rédigé au démarrage d'un projet de recherche, qui couvre tout le cycle de vie des données. Il décrit la façon dont les données seront obtenues, traitées, organisées, stockées, sécurisées, préservées, partagées… au cours et à l'issue du projet et aide à la mise en place de bonnes pratiques de gestion. Le PGD n'est pas un document figé, il évolue et est mis à jour pendant toute la durée du projet de recherche.

➲ POURQUOI RÉDIGER UN PGD ?

- Pour identifier les risques liés à la gestion des données, assurer la sécurité et la préservation des données sur le long terme
- Pour identifier les responsabilités, les rôles de chacun dans la gestion des données, planifier les ressources et compétences nécessaires à cette gestion
- Pour donner accès à des données fiables afin d'assurer la reproductibilité de la recherche et permettre à d'autres de comprendre et d'utiliser les données
- Pour répondre aux exigences des financeurs comme : *Research Councils* (Etats-Unis), *National Science Foundation* (NSF), *National Institutes of Health* (NIH), Horizon 2020, *Australian Research Council* (ARC), *National Health and Medical Research Council* (NHMRC)

EXEMPLE DE TRAME D'UN PLAN DE GESTION DE DONNÉES (PGD)

1) Informations administratives	▪ Nom et identifiant du projet ▪ Description du projet, agence(s) de financement ▪ Nom et identifiant éventuel du responsable principal de projet ▪ Contact pour les données de projet ▪ Date de la 1ère version ▪ Date de la dernière mise à jour ▪ Politiques associées au projet, incluant les instructions ou recommandations de l'agence de financement et des institutions participantes
2) Collection de données	▪ Description des données, incluant le type de données, le format et le volume ▪ Jeux de données préexistants qui seront utilisés ▪ Méthodes de collecte et de création des données ▪ Système d'organisation, de nommage et de gestion des répertoires et des fichiers ▪ Processus d'assurance qualité mis en œuvre

Information Scientifique et Technique
AGROPOLIS INTERNATIONAL MONTPELLIER, FRANCE

Contact : cist@agropolis.fr –http://ist.agropolis.fr- Octobre 2016

AGROPOLIS INTERNATIONAL

avec le soutien de Couperin.org
ADBS
couperin.org
Consortium Unitié des Etablissements Universitaires et de Recherche pour l'Accès aux Publications Numériques

3) Documentation et métadonnées	▪ Informations nécessaires pour lire, interpréter et reproduire les données ▪ Organisation de la collecte et de la documentation ▪ Standards de métadonnées adoptés : généralistes (*Dublin Core*, *Data Cite*) ou spécifiques à des domaines (ISA-Tab, FUGE-ML, *MicroArray Gene Expression Tabular*…)
4) Ethique, cadre légal	**4.1 - Ethique** ▪ Détails de l'accord de conservation et de partage des données ▪ Etapes pour la protection de l'identité des participants ▪ Etapes pour la sécurité du stockage et du transfert de données sensibles **4.2 - Droits de propriété intellectuelle et copyright** ▪ Nom de(s) propriétaire(s) des données ▪ Licence(s) pour la réutilisation des données (par exemple, une des licences *Creative Commons* ou *Open Data Commons*) ▪ Restrictions éventuelles d'utilisation par des tierces parties ▪ Délai requis pour le partage de données (embargo lié à la publication dans une revue ou délai d'application d'un brevet)
5) Stockage, sauvegarde, et sécurité	**5.1 - Stockage et sauvegarde** ▪ Lieu de stockage des données ▪ Plan de sauvegarde ▪ Personne ou équipe responsable de la sauvegarde ▪ Procédures de récupération **5.2 - Sécurité** ▪ Risques et leur gestion ▪ Dispositif d'accès ▪ Dispositif éventuel pour le transfert sûr et intègre des données collectées sur le terrain
6) Sélection et conservation	▪ Informations détaillées sur les données qui seront retenues, partagées et/ou conservées, et référence aux obligations contractuelles, légales ou réglementaires ▪ Utilisations prévues des données pour de futures recherches ▪ Durée de conservation des données au-delà du projet ▪ Entrepôt ou archive de conservation des données et responsabilités associées ▪ Temps et efforts nécessaires à la préparation des données pour leur conservation et leur partage
7) Partage des données	▪ Actions à mener pour faciliter la prise de connaissance (*discovery*) des données par d'autres ▪ Conditions éventuelles de restriction du partage des données et détails de leur application dans l'accord de partage de données ▪ Mécanisme de partage de données (*via* un entrepôt, sur demande expresse ou tout autre processus) ▪ Délai de publication ▪ Procédure éventuelle d'obtention d'un identifiant pérenne pour les données
8) Responsabilités et moyens	▪ Nom de la personne responsable de la mise en œuvre du plan de gestion de données ▪ Nom de la personne responsable de chaque étape de gestion des données ▪ Equipements et logiciels requis (en addition à ceux existants fournis par l'institution) ▪ Besoins additionnels d'expertise ou de formation ▪ Charges imposées par les entrepôts de données

La trame de plan présentée ci-dessus est une adaptation de **: *Checklist for a Data Management Plan.*** *V.4.0. Edinburgh, UK: Digital Curation Centre (DCC), 2014,* téléchargeable au format pdf sur http://www.dcc.ac.uk/resources/data-management-plans.
Elle correspond au formulaire en ligne ***Digital Curation Centre's DMPonline tool*** (https://dmponline.dcc.ac.uk/).

➲ EXEMPLES DE MODÈLES DE PLANS DE GESTION DE DONNÉES

- ***BBSRC Data Sharing Policy*** *: Version 1.2. Biotechnology and Biological Sciences Research Council,* 2016, 15 p.
  http://www.bbsrc.ac.uk/about/policies-standards/data-sharing-policy/
- ***ESRC Research Funding Guide***. *Economic and Social Research Council (ESRC), 2016*
  http://www.esrc.ac.uk/funding/guidance-for-grant-holders/research-data-policy/
- ***Full Data Management Plan Template***. *Natural Environment Research Council (NERC), 2012*, 3 p. (.doc)
  http://www.nerc.ac.uk/research/sites/data/dmp/
- ***Guidelines on FAIR Data Management in Horizon 2020*** *: Version 3.0. European Commission,* 2016, 12 p.
  http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- **Réaliser un Plan de Gestion de données**. Université Paris Diderot, 2015, 30 p.
  http://www.univ-paris-diderot.fr/DocumentsFCK/recherche/Realiser_un_DMP_V1.pdf
- ***UCCC's DMPTool***. *University of California Curation Center (UCCC), US*, 2014. https://dmptool.org/

La valorisation des données de la recherche

LES DIFFÉRENTS MODES DE PUBLICATION DE DONNÉES

Publier dans un entrepôt Privilégier un **« entrepôt de confiance » certifié** qui répond aux critères de qualité : -format des données, -qualité des métadonnées, -conditions d'accès et de réutilisation, -identifiant pérenne, -archivage à long terme, …	L'entrepôt de données est un réservoir constitué majoritairement de données de recherche, brutes ou élaborées, qui sont décrites par des métadonnées de façon à pouvoir être retrouvées. ▪ Entrepôts institutionnels : **Edinburgh dataShare** UK ; **Open Data LMU** Allemagne ; **Merritt** USA… ▪ Entrepôts thématiques et disciplinaires : **PANGAEA** (Sciences de la terre et environnementales) ; **BioSharing** (Sciences de la Vie et Biomédecine); **GenBank** (séquences d'ADN)… ▪ Entrepôt pluridisciplinaire : **Zenodo** (Europe), **Dryad**, **Figshare**…
Publier des données comme matériel supplémentaire d'un article	*Supplementary material, supplemental data* etc. : fichier contenant des données complémentaires à la publication. La plupart des revues préconisent un entrepôt.
Publier des données dans un *data paper*	Le *data paper* est un type de publication citable au même titre que les publications classiques : ▪ Dans une revue classique (type d'article : *data paper*). Ex. : **Ecology**… ▪ Dans un *data journal* (revue qui contient exclusivement des *data papers*) Ex. : **Scientific Data**, **Biodiversity Data Journal**, **Gigascience**…
Publier dans le web des données	Schéma de déploiement à 5 étoiles voir le site **http://5stardata.info/en/**

La (ré)utilisation des données et le droit d'auteur

La citation des données permet d'assurer une meilleure visibilité à l'auteur des données, facilite la diffusion de ces données, assure leur pérennité et permet de vérifier et valider les résultats de recherche.

Les standards de citation de jeux de données

5 champs sont obligatoires :

- Auteur
- Année de Publication
- Titre
- Editeur
- Numéro d'identification (DOI)

D'autres éléments peuvent les compléter :

- Version
- Type de données

Remarques :

- La revue ou l'entrepôt du jeu de données peuvent recommander un format de citation
- Le service *DOI Citation formatter* de Datacite vous permet de générer automatiquement une citation à partir d'un DOI **http://crosscite.org/citeproc/**

Exemple :
Claire Loison. (2015). Hydrated DPPC, MD simulation trajectory and related files for UA charmm36 model by Lee et al 2014. Zenodo. **http://doi.org/10.5281/zenodo.16978**

Le statut juridique des données de la recherche

A l'heure actuelle, l'environnement juridique entourant les données reste flou. Les données brutes ne sont *a priori* pas protégées par le droit d'auteur. Sous certaines conditions, le droit protégeant les bases de données peut s'appliquer.
Il est donc important de protéger ses données par des licences prédéfinies.
Un jeu de données peut être protégé par une licence ***Creative Commons,*** une **Licence ouverte** ou une **licence de l'OKF**

Les licences *Creative Commons*	CC0 ZERO	Domaine public / pas de droit réservé / Licence CC0 1.0
	CC BY	Attribution / Licence CC-BY 4.0
La licence Ouverte	LICENCE OUVERTE OPEN LICENCE	Attribution
Les licences de *l'Open Knowledge Foundation* (OKF)	OpenDataCommons.org Legal tools for Open Data	Licence ODC-BY / Attribution
		Licence ODC-ODbL / Base de données
		Licence PDDL / Domaine public

WEBOGRAPHIE

SITE D'INFORMATION SUR LES DONNEES DE LA RECHERCHE

- **Plateforme d'information officielle du Ministère de l'Enseignement Supérieur et de la Recherche**
  http://www.donneesdelarecherche.fr/
- **Gestion et partage des données scientifiques**, INRA Science & impact https://www6.inra.fr/datapartage/
- **Gestion des données de la recherche**
  CIRAD-CoopIST, mise à jour avril 2015
  http://coop-ist.cirad.fr/gestion-de-l-information/gestion-des-donnees-de-la-recherche

TUTORIELS, MODULES D'AUTOFORMATION :

- **INIST-CNRS. (2014).** Une introduction à la gestion et au partage des données de la recherche. Module de sensibilisation en ligne
  http://www.inist.fr/donnees/co/Donnees_recherche_web.html
- **Cosserat F., Ciolek-Figiel A. (2016)** Gestion et diffusion des données de la recherche. Stage URFIST 2016/05/12. Diaporama : 129 slides
  http://www.sites.univ-rennes2.fr/urfist/ressources/gestion-et-diffusion-des-donnees-de-la-recherche?destination=ressources

BIBLIOGRAPHIE

- **Dekkers M., Loutas N., De Keyzer M., Goedertier S. (2013).** Licences pour les données et les métadonnées. Module de formation 2.5. PWC
  http://fr.slideshare.net/OpenDataSupport/licences-pour-les-donnes-et-les-mtadonnes
- **Dzale E., L'Hostis D. (2015).** *Open Science. Gestion et partage des données de la recherche.* Journée de Formation Agropolis, 2015/04/01, Montpellier (France). Diaporama : 211 slides.
  http://prodinra.inra.fr/record/280536
- **OCDE. (2007).** *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics.* 29 p. Version. 1.0 11 December 2013
  http://www.oecd.org/fr/sti/sci-tech/38500823.pdf
- **Silvy C. (2015).** *De l'Open Access à l'Open Data : Enjeux et perspectives.* Séminaire CBGP, 2015/01/06, Montpellier (France). Diaporama : 100 slides
  http://www.ist.agropolis.fr/les-formations/tutoriels/item/de-l-open-access-a-l-open-data-enjeux-et-perspectives