GUIDE DE BONNES PRATIQUES

Gestion et valorisation des données de recherche

ARNOUD Pierre-Yves (OTELo), JACQUEMOT-PERBAL Marie-Christine (Inist-CNRS)

V1.1

01/02/2016

Relecteurs : AUCLERC Apolline (OTELo – LSE), BEGUIRISTAIN Thierry (OTELo-LIEC), LEGUÉDOIS Sophie (OTELo – LSE), MONTAGERS-PELLETIER Emmanuelle (OTELo – LIEC), RIPAMONTI-CHENOT Elodie-Denise (OTELo – LSE)

INTRODUCTION

L'objectif de ce guide est de rappeler les bonnes pratiques pour la gestion de vos données tout au long de leur cycle de vie (cf. figure 1) de leur production à leur diffusion et valorisation. La gestion des échantillons peut s'avérer aussi nécessaire pour une réutilisation ou un partage.

La mise en œuvre de ces pratiques aura des bénéfices pour vous mais aussi pour vos partenaires dans la mesure où il vous sera plus facile de :

- retrouver/réutiliser vos échantillons et données ;
- comparer/analyser vos données ;
- les intégrer à d'autres jeux de données ;
- les valoriser par leur diffusion ;
- faciliter leur réutilisation dans d'autres projets ou par d'autres chercheurs.

La mise en œuvre de bonnes pratiques peut être catégorisée en trois groupes d'actions.

A- Identifier, stocker, responsabiliser

B- Faciliter l'analyse des données et l'intégration de données hétérogènes

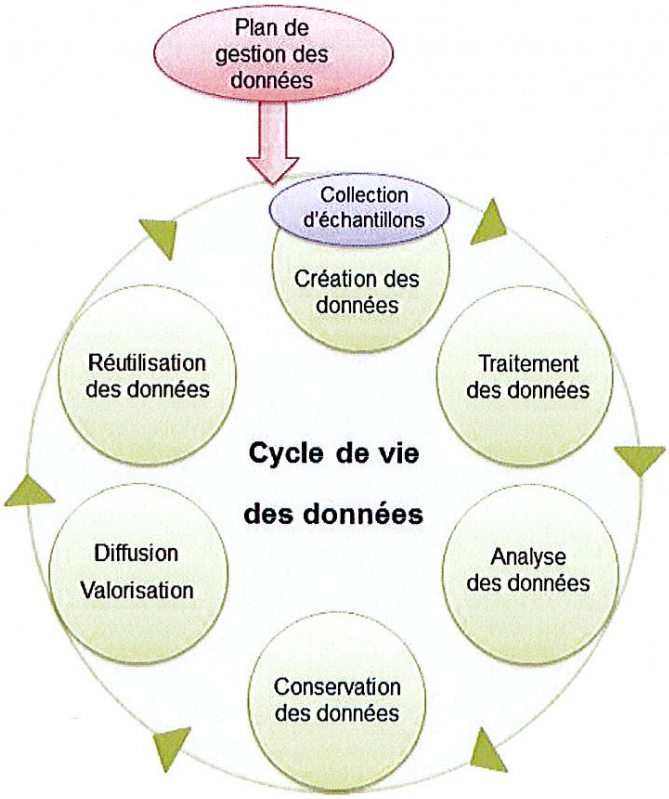
C- Diffuser, réutiliser et conserver vos données

Un plan de gestion de données (PGD ou DMP pour *Data Management Plan*) permettra de recueillir l'ensemble des informations concernant les données produites et/ou réutilisées ainsi que les choix

que vous aurez effectués tout en suivant les recommandations mentionnées ci-dessous. (voir page 22)

PGD : Document dont la rédaction est initiée au commencement d'un projet de recherche, qui décrit les données et comment elles seront partagées et conservées pendant et après le projet. Un PGD représente également un gage de qualité des données. Il est encouragé voire exigé (dans le cadre du pilot Open Data) par la Commission européenne dans le cadre du programme de financement H2020.

Figure 1 : Cycle de vie des données



1. IDENTIFICATION DES ECHANTILLONS

L'identification et la description de vos échantillons faciliteront leur réutilisation et l'analyse croisée de données produites par vous-même ou par différentes personnes à partir d'un même échantillon. Pour cela, il s'agit de convenir d'une nomenclature, ce qu'on appelle convention de nommage.

La nomenclature des échantillons doit contenir *a minima* les éléments suivants :

- localisation (p. ex. nom du lieu, de la parcelle ou de la station) ;
- la date de prélèvement (yyyymmdd ou yyyy-mm-dd[1]) ;
- le type d'échantillon (p. ex. sol, terre, eau).

Exemple

Échantillon d'eau (W) prélevé à Joeuf Abattoir (JOAB) le 7 mai 2015 dans le cadre du projet Mobised : **JOAB_20150507_W.**

Il peut aussi être utile de signifier d'autres éléments comme par exemple une modalité de prélèvement ou d'analyse.

Exemple

Échantillon de particules en suspension (SPM) après centrifugation de terrain (FC) à Joeuf abattoir (JOAB) le 7 mai 2015 : **JOAB_20150507_SPM_FC_1**

Il peut être aussi important de bien documenter la méthode d'échantillonnage, le conditionnement des échantillons et leur stockage si vous souhaitez vérifier des résultats, les réutiliser ou les partager. (Ci-dessous un exemple de modèle)

1 Donner les dates sous le format yyyy-mm-dd dans les noms permet de faire des tris informatiques rapides que ce soit dans les tableurs (p.ex. Excel) ou dans les gestionnaires de fichiers.

Modèle

Nom de l'échantillon	Selon la nomenclature établie
Méthode de prélèvement	Description, version ou référence bibliographique,
Date de prélèvement	YYYY-MM-DD (Norme 8601)
Personne responsable	Nom|prénom
Géolocalisation du point de collecte/prélèvement	Coordonnées GPS
Système de coordonnées utilisé	Lambert
Type d'échantillon	Liste contrôlée (p. ex. sol, eau)
Autres caractéristiques (champ répétable)	
Conditionnement	
Stockage	Localisation
Commentaire	

2. IDENTIFICATION DES FICHIERS DE DONNEES

Afin de faciliter la recherche de vos données, utilisez une nomenclature explicite pour vos fichiers.

Voici quelques éléments que nous vous conseillons de prendre en compte pour établir votre nomenclature :

- commencer par une lettre ;
- maximum 30 caractères[2] ;
- pas de caractères spéciaux, d'accent, d'espace ou de point[3], utilisez les lettres de a à z (majuscules et minuscules), les chiffres (0-9), le tiret et le tiret bas (souligné), le numéro ou l'acronyme du projet ;
- la date de création sous la forme : yyyy-mm-dd ou yyyymmdd[4] ;
- le nom du créateur ;
- une description brève du contenu (p. ex. paramètre mesuré, modalité, traitement appliqué) ;
- le numéro de version ;
- extension du format.

Exemple

Relevé de biomasse de la luzerne effectué par Pierre MARTIN (PM) dans le cadre du projet Multipolsite (MPS) sous un format Excel : **MPS_2011-05-30_PM_biomasse_v1.xlsx**

---

2 Afin d'éviter d'éventuels problèmes de transcription du nom lors de sauvegarde automatique sur un serveur ou de changement de système d'exploitation informatique.

3 Selon le système d'exploitation utilisé et la langue de configuration, l'encodage de ces caractères spéciaux est différent. Ainsi cela peut générer des erreurs dans le nom des fichiers lors de la copie du fichier sur un serveur de partage, de l'échange de dossiers entre deux ordinateurs, d'un changement d'ordinateur, d'un échange à l'international.

4 Voir note 1.

3. ORGANISATION D'UN ESPACE COLLABORATIF DEDIE AU PROJET

Un espace collaboratif a pour objectif de stocker, sauvegarder, sécuriser et donner accès aux jeux de données et fichiers produits par l'ensemble des partenaires dans le cadre du projet.

Cet espace dédié à un projet peut par exemple inclure les répertoires suivants :

- documents administratifs ;
- données, documentation et éventuellement les dictionnaires de données et/ou scripts associés avec une sous-catégorisation appropriée, par exemple :
  - données brutes,
  - données corrigées, vérifiées, dérivées,
  - données analysées incluant les données inputs, programmes/scripts, outputs, fichiers de travail intermédiaire, graphiques,
  - données continues (p. ex. sondes, capteurs) ;
- liste et description des échantillons (cf. section 1) ;
- protocoles (versions) et méthodologies utilisées ;
- comptes rendus de réunions ;
- participants au projet avec leurs coordonnées ;
- éléments de valorisation (p. ex. publications, posters et communications orales réalisés dans le cadre du projet) ;
- PGD (versions).

☝ Toujours conserver une copie des données brutes ou originales

Le CNRS (CoRe, MyCore), l'INRA, l'université de lorraine (B'UL, Filers) propose des solutions permettant :

- le stockage et la sauvegarde des données ;
- la conservation des versions successives des fichiers modifiés ;
- l'accessibilité à distance, 24h/24 et 7j/7 ;
- et la sécurité d'accès.

Contact CNRS Centre Est : jean.perruchaud@dr6.cnrs.fr
Contact INRA : dsi@inra.fr

Contact Université de Lorraine : dn-contact@univ-lorraine.fr

4. IDENTIFICATION DES PROPRIETAIRES DES DONNEES

Cas de données produites dans le cadre d'un projet

Dans le cadre d'un projet, la propriété des données est fixée dans l'accord de consortium. S'il n'en existe pas, elle sera précisée dans le PGD.

Complément d'info

https://www.dgdr.cnrs.fr/daj/modele/contrat/textes.htm

Cas d'achat ou de réutilisation de données issues d'un tiers

Dans le cas où vous souhaitez réutiliser des données, il est indispensable de vérifier la source, les licences et les conditions d'utilisation. Celles-ci seront alors mentionnées dans le plan de gestion.

Exemple de données réutilisables :

- données cartographiques IGN ou Open Street Map,
- données du Réseau de Mesure de la Qualité des Sols,
- données scientifiques partagées et disponibles dans des entrepôts (cf. section 9).

5. ATTRIBUTION DES ROLES ET RESPONSABILITES DE CHACUN

Afin d'assurer le bon déroulement du projet et une réutilisation future des échantillons et des données, il conviendra de définir les rôles et responsabilités de chacun. Pour cela, il est nécessaire de répondre aux questions ci-dessous et de consigner les réponses dans le PGD.

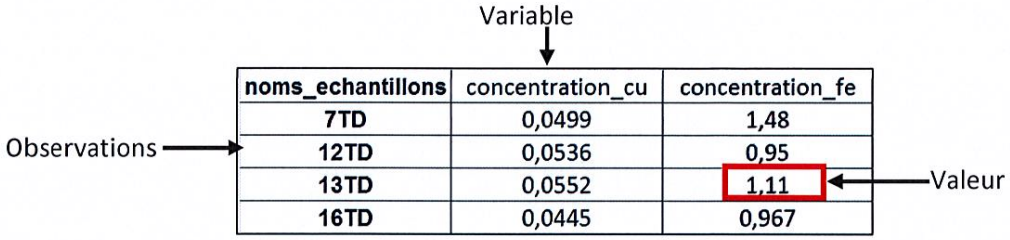
- Qui est responsable de la gestion des échantillons ?
- Qui est responsable pour chacune des activités de gestion de données (capture/collecte des données, description des données, qualité des données, stockage et sauvegarde des données) ?
- Qui est responsable de l'élaboration du PGD, de son application et de sa mise à jour tout au long du projet ?

6. STRUCTURATION DES DONNEES

Une bonne structuration de vos données facilitera les traitements et analyses automatiques avec un gain de temps et de fiabilité. Dans le cas où vous utilisez des tables (p. ex. Excel, LibreOffice calc) pour collecter vos données, voici quelques rappels et conseils.

Principes de base

- Une colonne représente une variable (Figure 2).
- Une ligne représente une observation.
- Chaque cellule contient une seule valeur (donnée).

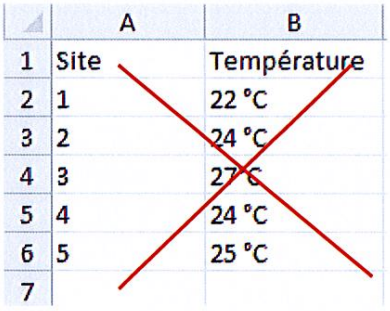


noms_echantillons	concentration_cu	concentration_fe
7TD	0,0499	1,48
12TD	0,0536	0,95
13TD	0,0552	1,11
16TD	0,0445	0,967

Figure 2 : exemple de table de données

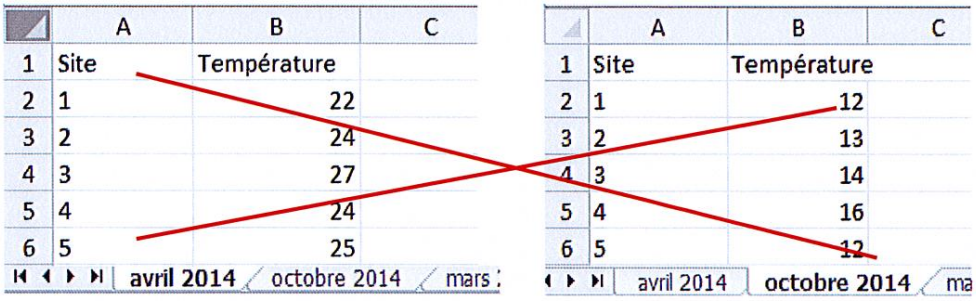
Recommandations

- Ne pas ajouter d'unité ou de commentaires dans les cellules car ils ne sont pas transférables sous d'autres systèmes de gestion de données ou sous certains logiciels de traitement de données (p. ex. R).



	A	B
1	Site	Température
2	1	22 °C
3	2	24 °C
4	3	27 °C
5	4	24 °C
6	5	25 °C
7		

- Utiliser un dictionnaire des données pour préciser les variables mesurées (cf. section 7).
- Saisir seulement du texte et des espaces dans les cellules, pas de retour à la ligne ou de tabulations.
- Ne jamais insérer plusieurs tables de données dans une même feuille.
- Éviter les feuilles multiples dans un même fichier car cela génère un risque d'erreur de saisie et vous obligera à recombiner les données dans un seul fichier pour une exploitation informatisée ou à les séparer en plusieurs fichiers



	A	B	C
1	Site	Température	
2	1	22	
3	2	24	
4	3	27	
5	4	24	
6	5	25	

avril 2014 | octobre 2014 | mars

	A	B	C
1	Site	Température	
2	1	12	
3	2	13	
4	3	14	
5	4	16	
6	5	12	

avril 2014 | octobre 2014 | ma

7. CREATION D'UN DICTIONNAIRE DE DONNEES

Pour une bonne compréhension des données, il est nécessaire d'expliciter les variables mesurées. Ceci sera réalisé par le biais d'un dictionnaire de données défini ci-dessous et qui sera associé à chaque fichier de données et à leur documentation (cf. section 8).

Exemple

Abréviation de la variable	Descriptif de la variable	Type de données	Domaine de valeurs autorisées	Format	Unité	Définition de la variable
temp	Température	Nombre décimal		X,xx	°C	
date_collecte	Date de collecte des données	Date	Norme ISO8601, W3CDTF	YYYY-MM-DD		Date à laquelle l'échantillon de sol a été prélevé
conc_pb	Concentration en plomb	Nombre décimal	>= 0	X,xxxxx	ppm	Si valeur = « » seuil de détection non atteint
species	Nom de l'espèce	Chaîne de caractères	Taxonomie de référence			
mandible_ width	Epaisseur de la mandibule	Nombre			µm	Epaisseur maximale de la mandibule Variable définie dans le thesaurus T-SITA

Recommandations

- **Abréviation de la variable**
    - Harmoniser les noms de variables communes à tous les fichiers et projets afin de faciliter le croisement des données.
    - Utiliser les abréviations les plus communément utilisées (p. ex temp pour température, lat pour latitude).
    - Préférer les noms en minuscules sans espace ou caractères spéciaux afin de faciliter l'exploitation automatique (p. ex. analyses statistiques avec R), le transfert entre les applications (seulement chiffres, lettres ou tiret bas).

> ☝ Il est conseillé pour des données de type date de séparer l'année et le jour dans des colonnes différentes.

- **Format de donnée**

> ☝ Attention à l'écriture des nombres en français et anglais (en anglais, la virgule indique les milliers et le point correspond à la virgule des nombres décimaux en français).

- **Unité de la variable** selon un standard international de préférence
- **Définition de la variable**
    - Définition et méthode de mesure utilisée.
    - Définition d'un code pour les valeurs manquantes auquel cas un commentaire pourra être ajouté dans un champ séparé (flag, abréviation du paramètre manquant).
- **Langue utilisée pour les variables et le dictionnaire de données** : afin d'éviter un travail de traduction au moment d'une publication, il peut être judicieux de privilégier l'usage de l'anglais dès le commencement du projet.

8. DOCUMENTATION DES DONNEES

La description des données produites est essentielle pour valider, reproduire, comprendre et retrouver vos données. Lorsque la description est structurée, on parle de **métadonnées**.

Recommandations

- Associer un fichier de métadonnées à chaque fichier de données.
- Préférer un format tabulé (p. ex. CSV) à un format texte (Read-me file : p. ex. PDF) si possible pour une exploitation automatique des données et le dépôt des données dans des entrepôts (cf. section 9).
- Utiliser des normes, standards, vocabulaires contrôlés (lexique, thésaurus) pour faciliter le partage et l'intégration de données avec votre communauté scientifique.
- Privilégier la langue d'usage de votre communauté de recherche pour faciliter la publication de vos données (cf. sections 9 et 10).

Voici ci-dessous un modèle que vous pouvez utiliser en sélectionnant les champs de métadonnées qui semblent adaptés à votre contexte, à l'entrepôt dans lequel vous déposez vos données et qui apporteront une information minimale et suffisante pour comprendre et reproduire vos données.

Métadonnées	Format
Nom du projet	Texte libre
Titre développé	Texte libre
Nom du porteur du projet	Nom | prénom
Institution	Texte libre
Titre du jeu de données	Texte libre. Le titre doit être explicite.
Identification du jeu de données	Nom du fichier (cf section 2)
Description du jeu de données	Description simplifiée du contexte de production de données
Date de création	YYYY-MM-DD (Norme ISO8601, W3CDTF)
Responsable du fichier	Nom | prénom
Laboratoire de rattachement du responsable de fichier	Texte libre ou liste contrôlée
Adresse mèl du responsable de fichier	

Producteur des données	Nom |prénom
Laboratoire de rattachement du producteur de données	Texte libre ou liste contrôlée
Adresse mèl du producteur de données	
Langue des métadonnées	fr ou en (Norme ISO639-1)
Thématique scientifique	Catégorie issue des Thèmes INSPIRE (cf. Annexe I) Mots-clés issus de thésaurus ou classifications
Nom du point de collecte ou d'observation	Texte libre
Géolocalisation du point de collecte/prélèvement	Coordonnées GPS
Système de coordonnées utilisé	Lambert
Date/heure de collecte	YYYY-MM-DD ou YYYY-MM-DDThh :mmTZD (TZD : désigne zone horaire Pour la France +01 :00 en hiver, +02 :00 en été) Norme ISO8601, W3CDTF
Echantillons	Décrit dans la section 1
Nom du protocole*	
Version du protocole*	
Description du protocole*	Description ou référence bibliographique
Paramètres du protocole*	Paramètres appliqués dans la méthode
Composants du protocole*	Instruments, logiciels ou scripts

*Les champs concernant les protocoles peuvent être répétés si plusieurs protocoles ont été successivement appliqués (p. ex. échantillonnage, préparation des échantillons, mesure, traitement des données, ...)

9. DEPOT DANS UN ENTREPOT DE DONNEES

Au moment de soumettre un article ou un *data paper* (ou publication de données, cf. section10), l'éditeur ou le financeur peut exiger ou recommander le dépôt des données sous-jacentes[5] dans un entrepôt reconnu ou *a minima* accessible.

Un **entrepôt de données** est un réservoir de données de recherche, brutes ou dérivées, qui peuvent être retrouvées et réutilisées grâce à une description par des métadonnées. Un identifiant pérenne peut être attribué à chaque jeu de données.

Recommandations

- Choisir de préférence :
  - un entrepôt disciplinaire s'il existe,
  - sinon un entrepôt institutionnel ou autre facilité locale,
  - ou bien un entrepôt multidisciplinaire européen (cf. complément d'info ci-dessous).
- Associer les fichiers « dictionnaire de données » (cf. section7) et métadonnées décrivant les données (cf. section 8) que vous déposez.
- Associer si besoin les scripts ou logiciels nécessaires pour reproduire les données.

Complément d'info

- Comment sélectionner un entrepôt ?
  - https://www.openaire.eu/repository/ordp/select-rep
  - https://www.dataone.org/best-practices/identify-suitable-repositories-data
  - http://www.inist.fr/formations/Deposer-ses-donnees-dans-un-entrepot/story.html

- Consulter des catalogues d'entrepôts si vous n'en connaissez pas :
  - multidisciplinaires : re3data (Registry of Research Data Repositories);
  - spécialisés en sciences de la vie : biosharing.

[5] Données sous-jacentes : données nécessaires à la validation des résultats présentés dans les publications scientifiques

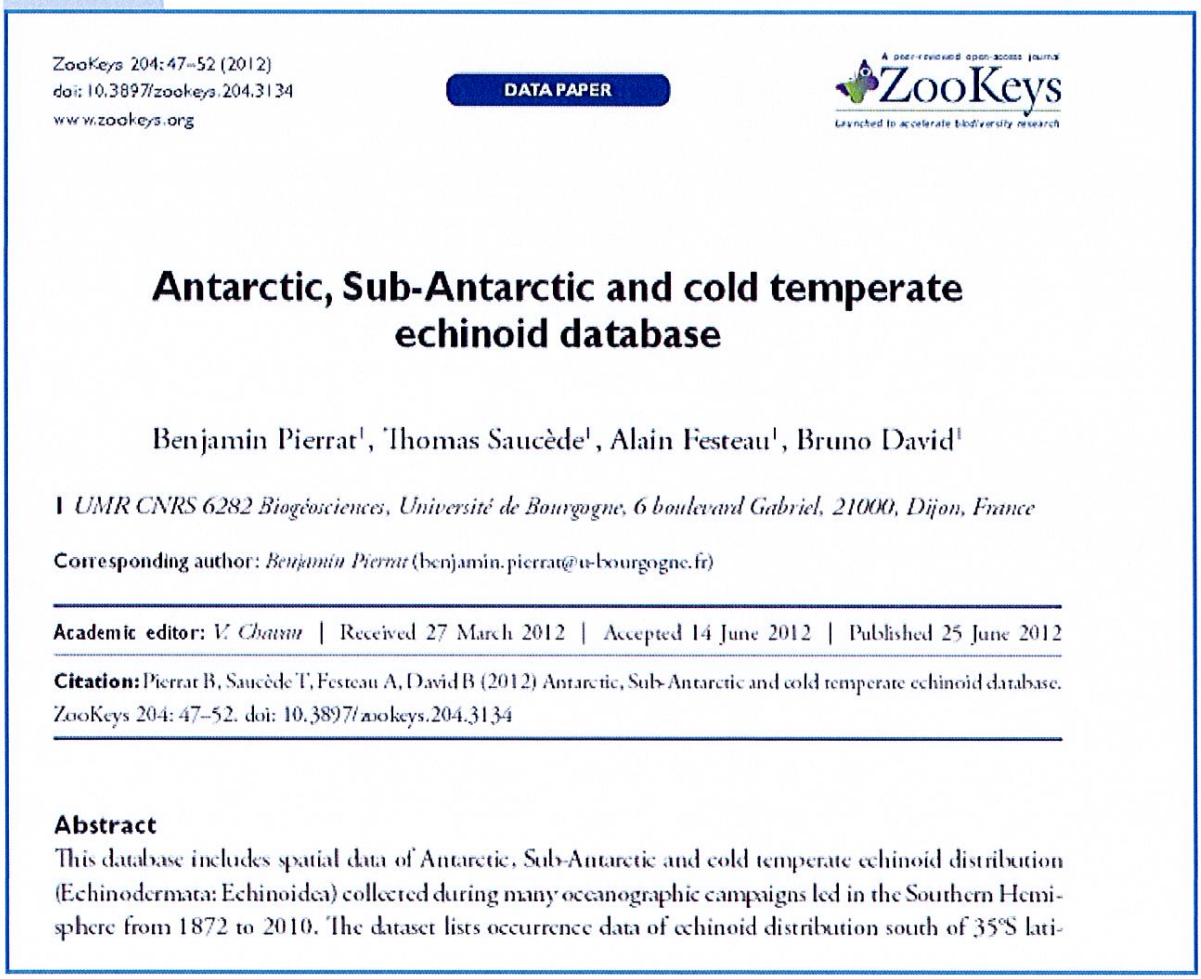
10. PUBLICATION DANS UN DATA PAPER

Vos jeux de données et/ou bases de données peuvent être valorisés par le biais d'un *data paper* ou publication de données.

Le ***data paper*** est une publication qui décrit des jeux de données et leur contexte de production et est revu par des pairs. Ce type de publication est citable au même titre qu'un article scientifique et sera rendu accessible par l'association d'un identifiant pérenne (p. ex. DOI[6], cf. section 11).

Les métadonnées produites (cf. section 8) permettront de produire ce *data paper* et faciliteront la soumission du *data paper* à l'éditeur.

Exemple



ZooKeys 204: 47–52 (2012)
doi: 10.3897/zookeys.204.3134
www.zookeys.org

DATA PAPER

ZooKeys
Launched to accelerate biodiversity research

Antarctic, Sub-Antarctic and cold temperate echinoid database

Benjamin Pierrat[1], Thomas Saucède[1], Alain Festeau[1], Bruno David[1]

1 *UMR CNRS 6282 Biogéosciences, Université de Bourgogne, 6 boulevard Gabriel, 21000, Dijon, France*

Corresponding author: *Benjamin Pierrat* (benjamin.pierrat@u-bourgogne.fr)

Academic editor: *V. Chaum* | Received 27 March 2012 | Accepted 14 June 2012 | Published 25 June 2012

Citation: Pierrat B, Saucède T, Festeau A, David B (2012) Antarctic, Sub-Antarctic and cold temperate echinoid database. ZooKeys 204: 47–52. doi: 10.3897/zookeys.204.3134

Abstract

This database includes spatial data of Antarctic, Sub-Antarctic and cold temperate echinoid distribution (Echinodermata: Echinoidea) collected during many oceanographic campaigns led in the Southern Hemisphere from 1872 to 2010. The dataset lists occurrence data of echinoid distribution south of 35°S lati-

[6] DOI : Digital Object Identifier

Complément d'info

Dedieu L. 2014. Rédiger et publier un *data paper* dans une revue scientifique en 5 points. Montpellier (FRA) : CIRAD, 7 p. http://coop-ist.cirad.fr/aide-a-la-publication/rediger/rediger-et-publier-un-data-paper/1-qu-est-ce-qu-un-data-paper

☝ En 2012, l'Aeres mentionne que « La constitution et la mise à disposition de bases de données, de logiciels, de corpus ou d'outils de recherche "est considérée" comme une production scientifique de rang A ».

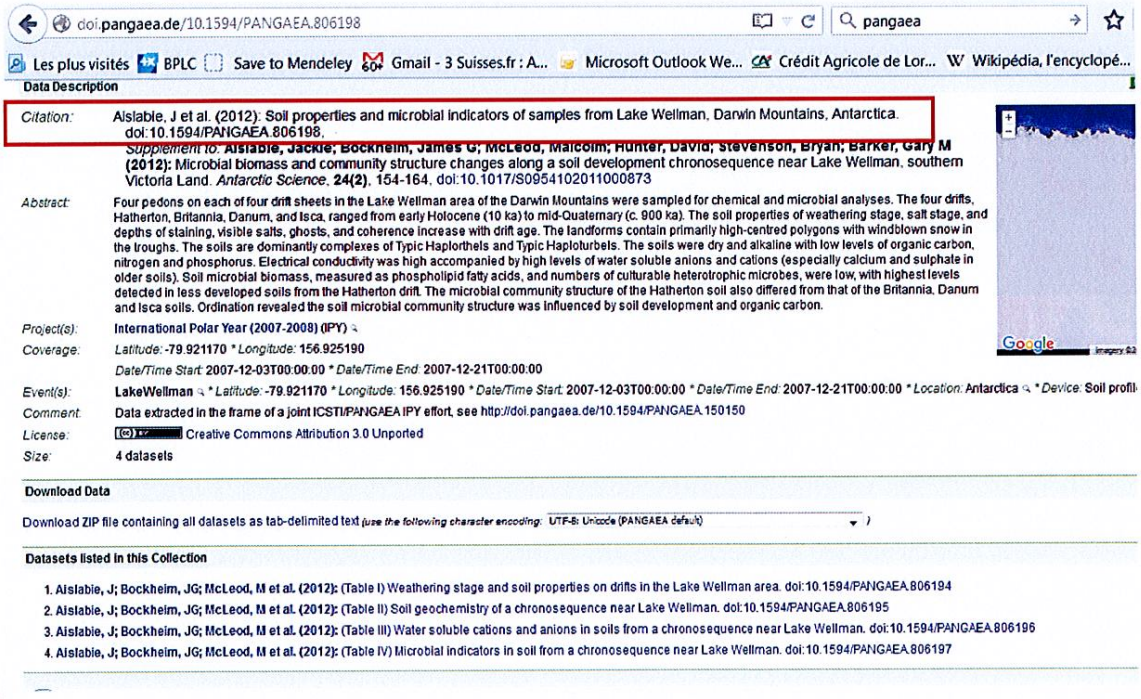
Source : Critères d'identification des chercheurs et enseignants-chercheurs "produisant en recherche et valorisation" AERES, 2012. www.aeres-evaluation.fr/content/download/18835/298036/file/Crit%C3%A8res%20Identif%20Ensgts-Chercheurs%2001102012.pdf

11. ATTRIBUTION D'UN IDENTIFIANT PERENNE

Un identifiant pérenne est un code normalisé (une chaîne de caractère) permanent associé à vos données.

Un identifiant pérenne permet :

- d'identifier de manière univoque vos données ;
- d'y accéder même si l'URL est modifiée ;
- de les citer ;
- de les relier aux publications.



Données entreposées dans l'entrepôt Pangaea (Source : citation encadrée. Accessible en ligne doi:10.1594/PANGAEA.806198)

Complément d'info

Cf. Annexe 2. Des identifiants pérennes pour les données de la recherche

Contacter : Inist-CNRS, Equipe valorisation des données de recherche, Service Analyser Valoriser, Département de l'Offre de Service
Téléphone : +33 (0)3 83 50 46 25
Email : contact-donneesrecherche@inist.fr

12. ATTRIBUTION D'UNE LICENCE

Il est conseillé d'attribuer une licence à vos données lors de leur diffusion. Une licence est un contrat qui indique aux utilisateurs exactement ce qu'ils peuvent faire avec vos données.

Recommandations pour le choix d'une licence :

- licence communément utilisée ;
- licence portable d'une juridiction à l'autre ;
- licence la moins restrictive possible afin de faciliter la réutilisation des données ;

La licence Creative Commons 4.0, par exemple, est internationale et couvre à la fois le droit d'auteur et le droit *sui generis* des bases de données.

Complément d'info

Dedieu L., Fily M.F. 2015. Rendre publics ses jeux de données scientifiques en 6 points. Montpellier (FRA) : CIRAD, 6 p.http://coop-ist.cirad.fr/gestion-de-l-information/gestion-des-donnees-de-la-recherche/rendre-publics-ses-jeux-de-donnees/6-les-principales-licences-de-diffusion-des-jeux-de-donnees

13. SELECTION ET ARCHIVAGE DES DONNEES

Dès le début du projet, il est important de réfléchir aux données à conserver et à leur durée de conservation. Les questions ci-dessous peuvent vous guider dans cette prise de décision.

- Doit-on conserver à long terme toutes les données générées au cours d'un projet de recherche ?
- Quelles sont les réutilisations possibles pour ces données ?
- Quelle sont les données qui doivent être conservées pour des raisons juridiques ou politiques, contractuelles, réglementaires ?
- Quelles sont les données qui doivent être conservées en se basant sur des critères de qualité des données, demandes ?
- Quel est le rapport coût – bénéfice de l'acquisition de ces données ?

Complément d'info

NERC data value checklist

University of Bristol (2013). Research Data Evaluation Guide
http://data.bris.ac.uk/files/2014/02/Research-data-evaluation.pdf

14. CHOIX DES FORMATS

Pour faciliter la réutilisation et la conservation des données, il est conseillé d'utiliser des formats non propriétaires ou largement utilisés afin de prévenir les potentiels problèmes d'obsolescence des logiciels.

Exemples

Format déconseillé	Format préféré
Excel (.xls, .xlsx)	Comma Separated Values (.csv)
Word (.doc, .docx)	Plain text (.txt) If formatting is needed, PDF/A (.pdf)
PowerPoint (.ppt, .pptx)	PDF/A (.pdf)
Photoshop (.psd)	TIFF (.tif, .tiff)
Quicktime (.mov)	MPEG-4 (.mp4)
Base de données MySQL (.sql)	Comma Separated Values (.csv) ou XML

Complément d'info

https://dmptool.org/dm_guidance#formats

http://datacentrum.3tu.nl/en/publishing-research/data-description-and-formats/

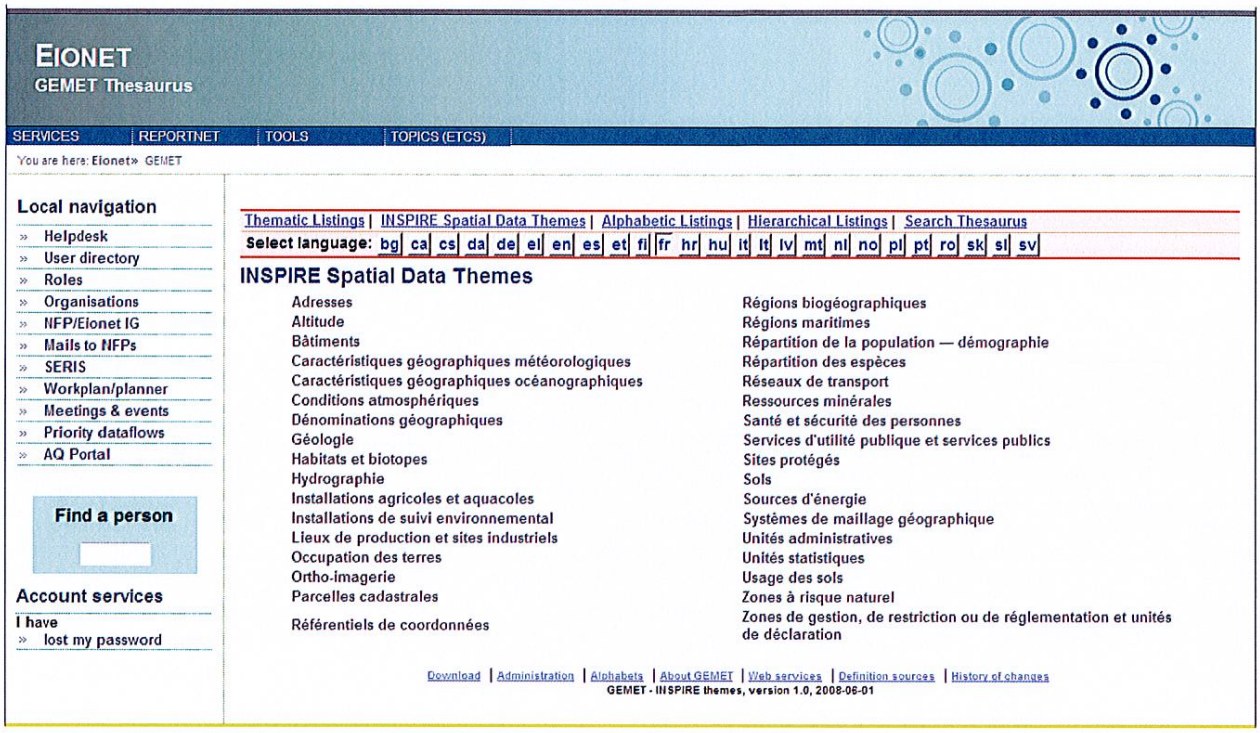
La conservation des versions de logiciels utilisées devra être envisagée s'il s'agit de formats propriétaires et si les données doivent être reproduites dans le temps.

15. PRODUCTION D'UN PGD

Structure d'un PGD[7]	
Informations sur le projet	données administratives + cf. section 4 (p. 8)
Responsabilité des données	Cf. section 5 (p. 9)
Ressources nécessaires à la mise en œuvre du PGD	Estimer le coût des ressources humaines et de l'archivage dans un entrepôt
Jeux de données L'ensemble des sections ci-dessous doivent être dupliquées pour chaque jeu de données c'est-à-dire un ensemble de données techniquement homogène ou intellectuellement cohérent identifié comme tel.	
Description du jeu de données	Cf. sections 7 et 8 (p. 11 et 13)
Stockage, accès et sécurité des données	Cf. section 3 (p. 7)
Métadonnées : documentation et organisation des données	Cf. sections 2, 3 et 8 (p. 6, 7 et 13)
Diffusion des jeux de données	Cf. sections 9, 10, 11, 12 (p. 15, 16, 18, 19)
Sélection et archivage des données	Cf. sections 13 et 14 (p. 20 et 21)

[7] Structure d'un DMP extraite de Cartier A, Moysan M, Raymonet N. Réaliser un plan de gestion de données : guide de rédaction. (VO1 09/01/2015)

ANNEXE 1



EIONET GEMET Thesaurus : http://www.eionet.europa.eu/gemet/inspire_themes?langcode=fr

ANNEXE 2

Des identifiants pérennes

pour les données de la recherche

Qu'est-ce qu'un identifiant pérenne ?

Un *identifiant unique* est un code d'identification unique assigné à un objet (ou une personne) de manière à ce que cet objet puisse être référencé sans ambiguïté[8]. Le code d'identification est le plus souvent une chaîne de caractères alphanumériques.

Un *identifiant pérenne* (Persistent identifier ou PID) est un identifiant qui est assigné à un objet de façon permanente[8]. Il est disponible et gérable à long terme ; il ne changera pas si l'objet est renommé ou déplacé (p. ex. changement de site, d'entrepôt). Les URL sont communément utilisées comme identifiants mais leur pérennité n'est pas assurée.

A quoi sert un identifiant pérenne pour les données de la recherche ?

- ✓ Retrouver l'emplacement de données de la recherche sur le web même si leur URL a changé (p. ex. changement de site, d'institution, d'entrepôt), et éviter ainsi le message « HTTP 404-File not found ».
- ✓ Rendre accessibles les données de la recherche au même titre que les publications (de manière aussi pérenne que souhaitée par le producteur).
- ✓ Faciliter les citations des données de la recherche.
- ✓ Lier les données de la recherche aux publications.

---

8 Adapté de Borremans Catherine (2012). Compte-rendu de l'Atelier sur les Identifiants Persistants (PIDs) organisé par le GBIF France à Paris (MNHN) le 27 juin 2012. http://archimer.ifremer.fr/doc/00119/22978/ consulté le 04 janvier 2013

Principaux systèmes d'identifiants utilisés pour les données de recherche

Système	Nom développé	Exemple
DOI*	Digital Object Identifier	doi:10.1594/PANGAEA.726855
Handle	Handle	hdl:10283/239
EPIC PID*	European Persistent Identifier Consortium	http://hdl.handle.net/11304/3339d230-b988-11e3-8cd7-14feb57d12b9
PURL*	Persistent Uniform Resource Locator	https://treebase.org/treebase-web/search/study/anyObjectAsRDF.rdf?namespacedGUID=TB2:S1975
ARK	Archival Resource Key	ark:/b7272/q6td9v7j

*identifiant pérenne basé sur le système Handle

Il existe d'autres identifiants pérennes, comme par exemple :

- URN : NBN National Bibliographic Numbers (ex : urn:nbn:nl:ui :32-424171),
- LSID, Life Science IDentifier : identifiant utilisé par la communauté des sciences de la vie.

Illustration du lien entre les jeux de données et la publication dans laquelle ils sont cités