GUIDE OF GOOD PRACTISES

Research data management and promotion

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

V1.1

June 20, 2017

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

INTRODUCTION

The aim of this guide is to provide the data management good practices to be used all along data life cycles (see Figure 1) from their production on to their distribution and the promotion of their potential. Managing samples may also prove necessary for re-use and sharing.

Implementing these practices will be useful for you and also for your partners, as this will make it easier to:

- locate and re-use your samples and data;
- compare your data;
- merge it into other datasets;
- promote it through availability;
- facilitate re-use in other projects or by other researchers.

There may be three types of actions to implement good practices.

A- Identifying, storing and assigning responsibilities

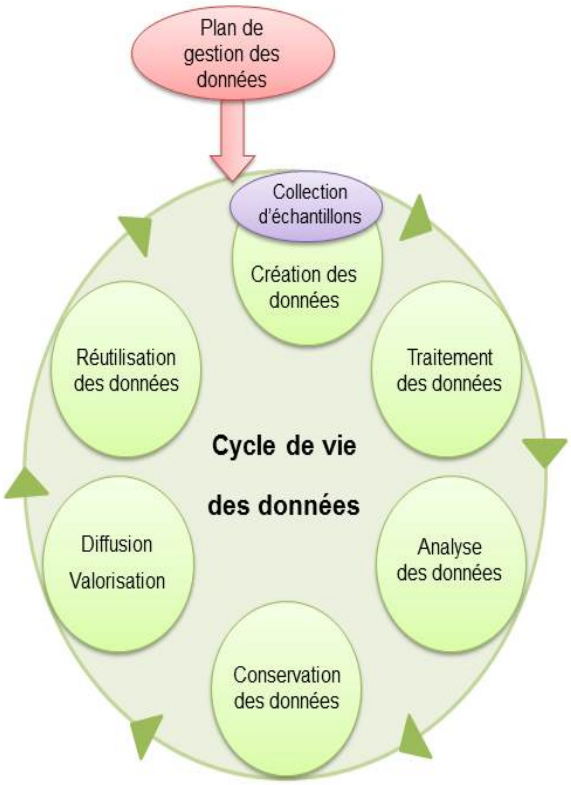
B- Facilitate data analysis and heterogeneous data integration

C- Distributing, re-using and preserving your data

A data management plan (DMP) will help you to record all information about the data produced/re-used and the options made, and to comply with the guidelines below. (see page 22)

DMP: A document, the drafting of which begins at the start of research project, which describes the data and the way it is going to be shared and preserved during and after the project. A DMP is also a guarantee of data quality. The European Commission promotes and even requires (as part of the Open Data pilot) DMPs under its H2020 funding programme.

Figure 1: Data life cycle



Plan de gestion de données = Data management plan ; Collection d'échantillons = Sample collection ; Création de données = Data creation ; Traitement des données = Data processing ; Analyse des données = Data analysis ; Conservation des données = Data preservation ; Diffusion Valorisation = Distribution / Promotion ; Réutilisation des données = Re-use of available data

1. IDENTIFYING SAMPLES

Identifying and describing your samples will facilitate the re-use and cross-analysis of the data you or different people obtained from the same samples. To do so, a nomenclature, or naming convention has to be defined.

The sample naming convention used must at least include the following elements:

- location (e.g., the name of a place, land plot or sampling station);
- sampling date (YYYYMMDD or YYYY-MM-DD[1]);
- sample type (e.g., soil, earth, water).

Example

Water sample (W) collected in Joeuf Abattoir (JOAB) on May 7, 2015 as part of Mobised project: **JOAB_20150507_W.**

It may be useful to provide other elements, like a collection or analysis method.

Example

Suspended particle material (SPM) after field centrifugation (FC) in Joeuf abattoir (JOAB) on May 7,2015: **JOAB_20150507_SPM_FC_1**

It may also be worthwhile to thoroughly document the sampling method, the packaging and storing of samples if you intend to check results, re-use or share them. (You will find an example for a model below)

1 Giving dates in a yyyy-mm-dd format in names enables quick computing sorting, either in spreadsheet software (e.g., Excel) or in file managers.

Model

Sample name	As per agreed naming method used
Sampling method	Description, version or bibliographic reference
Sample collection date	YYYY-MM-DD (Standard No. 8601)
Person in charge	Surname, name
Positioning system coordinates of collection/sampling site	GPS coordinates
Coordinate system used	Lambert
Sample type	Controlled list (e.g., soil, water)
Other characteristics (repeatable field)	
Packaging	
Storing	Location
Notes	

2. IDENTIFYING DATA FILES

To make data searches easier, use an explicit naming method for your files.

Here are a few elements you should take into account to create your naming method:

- begin with a letter;
- do not use more than 30 characters[2];
- do not use special characters, spaces or dots[3], use upper-case or lower-case a-z letters, digits (0-9), hyphens or underscores, the project number or acronym;
- use the date of creation in the following formats : yyyy-mm-dd or yyyymmdd[4] ;
- the data creator's name;
- type a brief content description (e.g., measured parameters, method, applied treatment);
- the version number;
- the format extension.

Example

Biomass sampling of alfalfa collected by Pierre MARTIN (PM) as part of Multipolsite (MPS) project in an Excel format file: **MPS_2011-05-30_PM_biomasse_v1.xlsx**

2 To avoid any possible name transcription problem during automated backups on a server or when changing a computer operating system.

3 The encoding of these special characters varies depending on the operating system and the configuration language used. Thus, this may generate file name errors when copying files on a shared server, when exchanging files between two computers, when changing one's computer, or during an international data exchange transaction.

4 See note 1.

3. ORGANISING A PROJECT-DEDICATED WORK SPACE

A collaborative work space's purpose is storing, saving, securing and providing access to the datasets and files obtained by all partners in a project.

For example, the project-dedicated space may include the following directories:

- Administrative documents;
- Data, documents and possible data dictionaries, associated scripts with an appropriate subcategory, e.g.,
  - Raw data,
  - Corrected, checked or derived data,
  - Analysed data together with input data, prg, programmes/scripts, outputs, intermediate work file, graphs,
  - Continuous-flow data (e.g., from probes, sensors) ;
- List and description of samples(see section 1) ;
- Protocols (versions) and methods used ;
- Minutes of meetings;
- Project participants with their contact details;
- Promotion material (e.g., journal articles, posters and conference contributions made during the project) ;
- DMP (versions).

☝ Always keep a copy of the raw/original data.

CNRS (CoRe, MyCore), INRA, University of Lorraine (B'UL, Files) propose solutions to :

- Store and save data;
- Keep the successive versions of the modified files;
- 24/7 remote access; and,
- Secured access.

CNRS Center-est Contact : jean.perruchaud@cnrs.fr

INRA Contact : dsi@inra.fr

University of lorraine Contact : dn-contact@univ-lorraine.fr

4. IDENTIFYING DATA OWNERS

Data produced during a project

In a project; the ownership of the data is set out in the consortium agreement. Otherwise, this will be specified in the DMP.

Additional info

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

Purchased or re-used third-party data

If you want to re-use such data, you must check the source, the licences and conditions of use. These are specified in the data management plan.

Example of reusable data:

- French national geographical institute (or IGN) or Open Street Map cartographic data,
- Data from the Réseau de Mesure de la Qualité des Sols (soil quality assessment network),
- Shared scientific data, which are available in repositories (see section 9).

5. ASSIGNING ROLES AND RESPONSIBILITIEs

For good project operation and to allow for good re-usability of data in the future, the roles and responsibilities of each person need to be defined. This is why you have to answer the following questions and to record your answers in the DMP.

- Who is in charge of managing the samples?
- Who is in charge for each of the data management steps (data capture/collection, data description, data quality, data storing and backup)?
- Who is responsible of drafting, enforcing and updating the DMP during the whole project?

6. STRUCTURING THE DATA

Well-structured data will facilitate processing and automated analyses and bring about considerable time saving and reliability. . If you use spreadsheets (e.g., Excel, LibreOffice calc) to collect your data, here are a few reminders and tips.

Core principles

- One column represents one variable (Figure 2).
- One row represents one sample collection.
- One cell contains only one value (data item).

Variable

Sample →

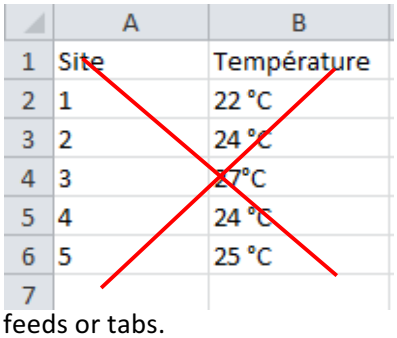
noms_echantillons	concentration_cu	concentration_fe
7TD	0,0499	1,48
12TD	0,0536	0,95
13TD	0,0552	1,11
16TD	0,0445	0,967

← Value

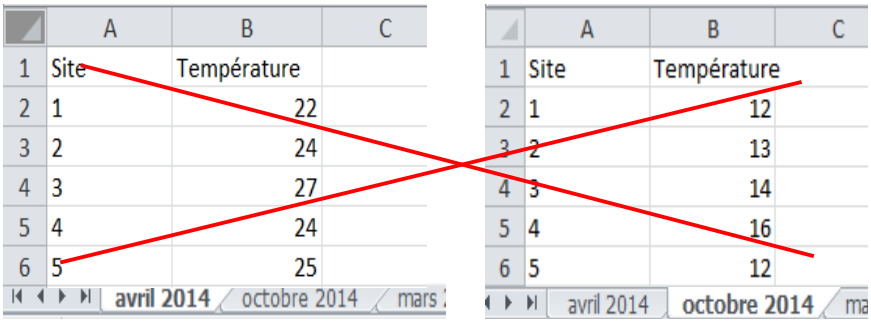
Figure 2: Model of a data table

Recommendations

- Do not add units or comments in the cells, because they cannot be transferred into other data management systems or fed into other data management computer programmes (e.g., R).



- Use a data dictionary to specify the variables measured (see section 7).
- Only type text and spaces in the cells, do not enter line feeds or tabs.
- Never insert several data tables in the same spreadsheet.
- Avoid several spreadsheets in the same file, because this may facilitate input mistakes and you may have to recombine data in one single file for computer processing or to break them down into separate files



7. CREATING A DATA DICTIONARY

To have your data easily understood, you must plainly explicit the variables measured. This will be achieved by means of a data dictionary as the one below which will be associated with each data file and its documentation (see section 8).

Example

Variable abbreviation	Variable description	Data type	Domain of authorised values	Format	Unit	Variable definition
temp	Temperature	Decimal number		X,xx	°C	
date_collecte	Data collection date	Date	ISO8601 Standard, W3CDTF	YYYY-MM-DD		Date when soil sample was collected
conc_pb	Lead concentration	Decimal number	>= 0	X,xxxxx	ppm	If value = « » detection threshold not reached
species	Name of species	Character string	Reference taxonomy			
mandible_ width	Mandible thickness	Number			µm	Maximum mandible thickness. Variable defined in T-SITA thesaurus

Recommendations

- **Variable abbreviation**
    - Standardise all variables common across all files and projects to facilitate data cross-referring.
    - Use the most usually used variables (e.g., temp for temperatures, lat for latitudes).
    - Preferably use lower-case names without any space or special character to facilitate automated processing (like statistical analysis using R), cross application transfers (numbers, letters or underscore only).

☝ It is advisable to separate years and days in different columns.

- **Data format**

☝ Beware number transcription in French and English (in French, dots correspond to thousands and commas to decimal numbers).

- **Preferably comply with an international standard for variables**
- **Variable definition**
    - Definition and method used for measurements.
    - Definition of a code for missing values, in which case a comment may be added in a separate field (flag, abbreviation of the missing parameter).
- **Language used for the variables and the data dictionary**: to avoid extra translation work when about to publish results, it may be reasonable to rather use English right from the start of the project.

8. DOCUMENTING THE DATA

The description of the data is crucial to be able to validate, reproduce, understand and retrieve your data. In the case of a structured description, it is referred to as **metadata**.

Recommendations

- Associate a metadata file to each data file.
- Preferably use a tab-delimited file format (e.g.,. CSV) instead of a text file (Read-me file: e.g., PDF) if possible for an automated data processing and their uploading in a repository (see section 9).
- Use standards, controlled vocabularies (lexicons, thesauri) to facilitate data sharing and integration within your scientific community.
- Preferably use the language commonly used in your research community to facilitate the publication of your data (see sections 9 and 10).

Here is a model you can use by choosing the metadata fields suitable to your context, the repository where you upload your data, and which will convey minimum and sufficient information to help others understand and reproduce your data.

Metadata	Format
Project name	Free text
Full title	Free text
Project lead name	Surname, name
Institution	Free text
Dataset title	Free text. The title must be explicit.
Dataset identification	Filename (see section 2)
Dataset description	Simplified description of the data production context
Creation date	YYYY-MM-DD (ISO 8601 Standard), W3CDTF
Person in charge of the file	Surname, name
Affiliation laboratory of the person in charge of the file	Free text or controlled list

Email of the person in charge	
Data producer	Surname |name
Affiliation laboratory of the data producer	Free text or controlled list
Email of the data producer	
Metadata language	fr or en (ISO639-1 standard)
Scientific subject	Category as per INSPIRE Themes (see Annex I) Keywords from thesauri or classification schemes
Name of the collection or measurement location	Free text
Positioning system coordinates of collection/sampling site	GPS coordinates
Coordinate system used	Lambert
Collection date/time	YYYY-MM-DD or YYYY-MM-DDThh :mmTZD (TZD : refers to a Time Zone For France, +01 :00 in winter, +02 :00 in summer) ISO8601 Standard, W3CDTF
Samples	As described in section 1
Protocol name*	
Protocol version*	
Protocol description*	Description or bibliographic reference
Parameter of the protocol*	Parameters used in this method
Protocol components*	Instruments, software or scripts

*The required fields about protocols can be repeated if several protocols have been implemented consecutively (e.g., sampling, sample preparation, measurements, data processing, etc.)

9. Depositing in a data repository

When you are ready to submit an article or a *data paper* (or a data publication, see section10), the publisher or research funder may require or recommend that the underlying data be deposited in a [5]recognised - or at least accessible - repository.

A **data repository** is reservoir for raw or derived research data which can be retrieved and re-used because they are described with metadata. A persistent identifier can be assigned to each dataset.

Recommendations

- Preferably choose:
  - A subject-based repository,
  - Otherwise an institutional repository, or a local facility, or
  - A European multidisciplinary repository (see "Additional info" below).
- Attach the "data dictionary" (see section7) and metadata files describing the data (see section 8) you are depositing.
- If necessary, attach the scripts and software needed to reproduce the data. .

Additional

- How to select a repository?
  - 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

- Search repository registries if you do not know any :
  - Multiple disciplines: re3data (Registry of Research Data Repositories);
  - Life sciences: biosharing.

5 Underlying data : those data necessary to validate the results discussed in scholarly publications

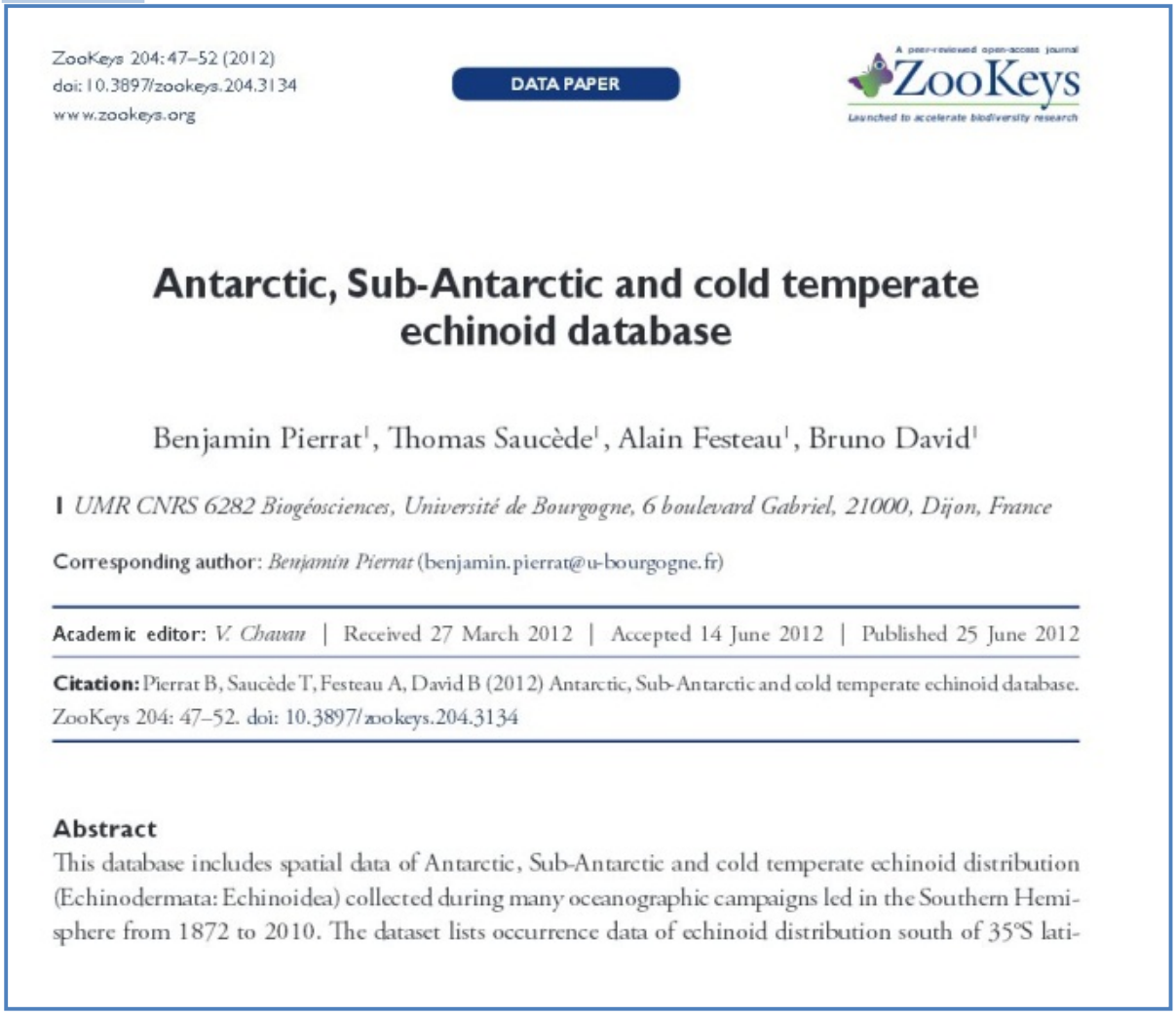
10. PUBLISHING IN A DATA PAPER

You can promote your datasets or databases by means of *data paper* or a data publication.

A ***data paper*** is a kind of peer-reviewed publication which describes a dataset and its production context. This type publication is as citable as a scholarly article and associated with a persistent identifier will become accessible. (e.g., DOI[6],see section 11).

The attached metadata (see section 8) will enable this *data paper* to be produced and will facilitate the submission of this *data paper* to a publisher.

Example



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

DATA PAPER

ZooKeys
A peer-reviewed open-access journal
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. Chavan* | 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

Additional

Dedieu L. 2014. Rédiger et publier un *data paper* dans une revue scientifique en 5 points. (Drafting and publishing a data paper in a scholarly journal in 5 steps) 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

☝ In 2012, the Aeres agency made it public that "Producing and making databases, software, corpora or search tools" shall be considered as " a rank-A scientific output".

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-

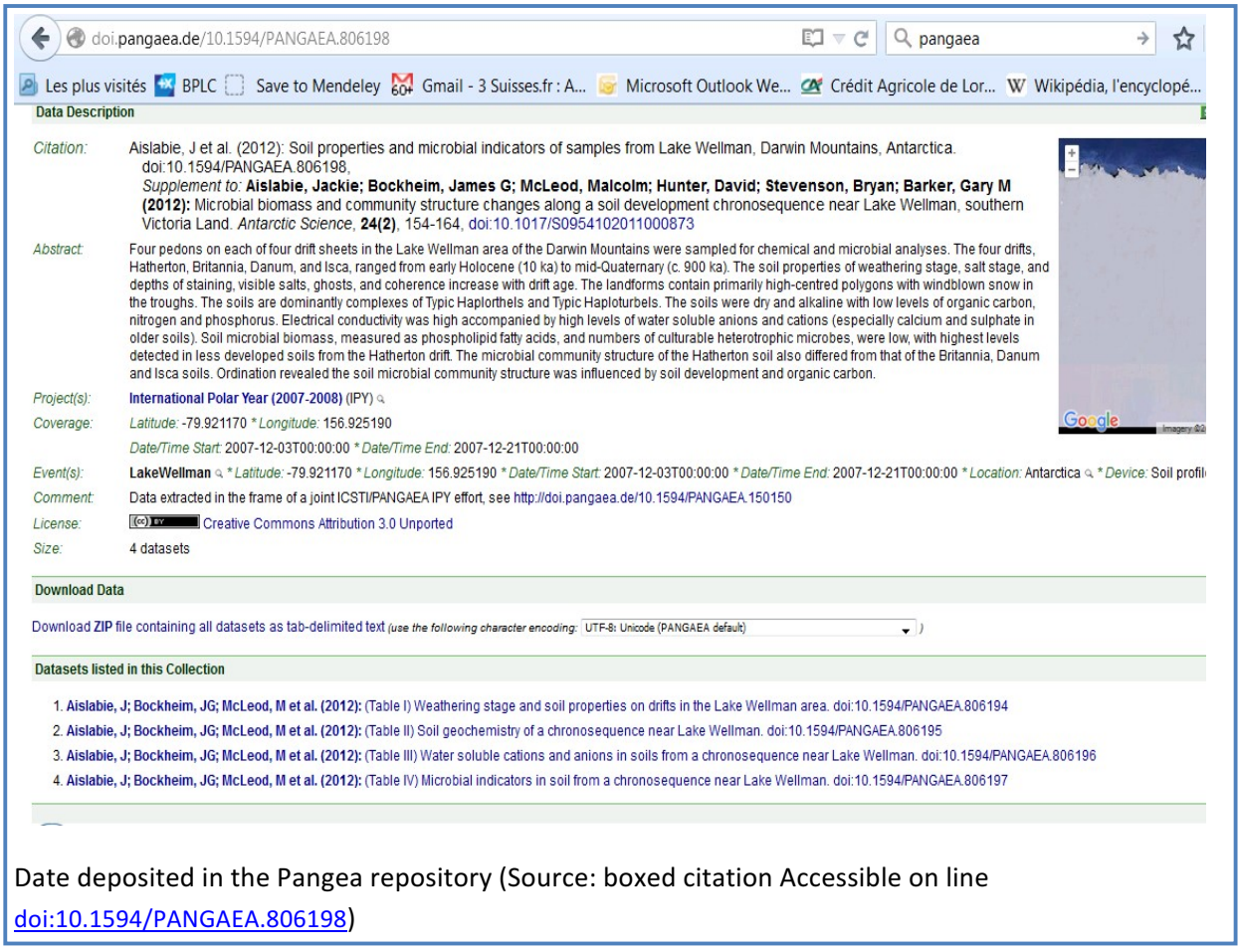
11. ASSIGNING A PERSISTENT IDENTIFIER

A persistent identifier is standardised code (a character string) permanently linked to your data.

Example

A persistent identifier helps you to:

- Identify your data unequivocally;
- Make them accessible, even if the URL changes;
- Make them citable;
- Link them to the publication.



Date deposited in the Pangea repository (Source: boxed citation Accessible on line doi:10.1594/PANGAEA.806198)

Additional

See Annex 2. Persistent identifiers for research data

Please contact: Inist, Research Data Impact Development Unit, Information Impact Development and Analysis Section, Information Services Department
Telephone: +33 (0)3 83 50 46 25
Email: contact-donneesrecherche@inist.fr

12. ASSIGNING A LICENSE

It is advisable to assign a license to your data when making them public. A licence is a legal agreement stipulating very precisely what users are allowed to do with your data.

Recommendation for choosing a licence:

- Choose a commonly used license;
- Choose a license that is portable from one jurisdiction to another;
- Choose a license that is the least restrictive as possible in order to facilitate the re-use of your data.

For example, the Creative Commons 4.0, is international and covers both the author's rights and "sui generis" database rights.

Additional info

Dedieu L., Fily M.F. 2015. Rendre publics ses jeux de données scientifiques en 6 points. (Making your scientific datasets publicly available in 6 steps) 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. SELECTING AND ARCHIVING DATA

At the very start of a project, it is important to consider which data to preserve and how long they should be preserved. The following question may help you make your decision.

- Should the projects data be preserved on a long-term basis?
- What are the likely re-uses of this data?
- Which data should be preserved for lawful, political, legal or statutory reasons?
- Which data should be preserved based on data quality or demand criteria?
- What is the cost-benefit ratio for the acquisition of this data?

Additional

NERC data value checklist

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

14. CHOOSING DATA FORMATS

To make it easier to re-use and preserve datasets, it is recommended to use non-proprietary, or widely used, formats to avoid software to become obsolete.

Examples

Not recommended	Preferred formats
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)
MySQL database (.sql)	Comma Separated Values (.csv) or XML

Additional

https://dmptool.org/dm_guidance#formats

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

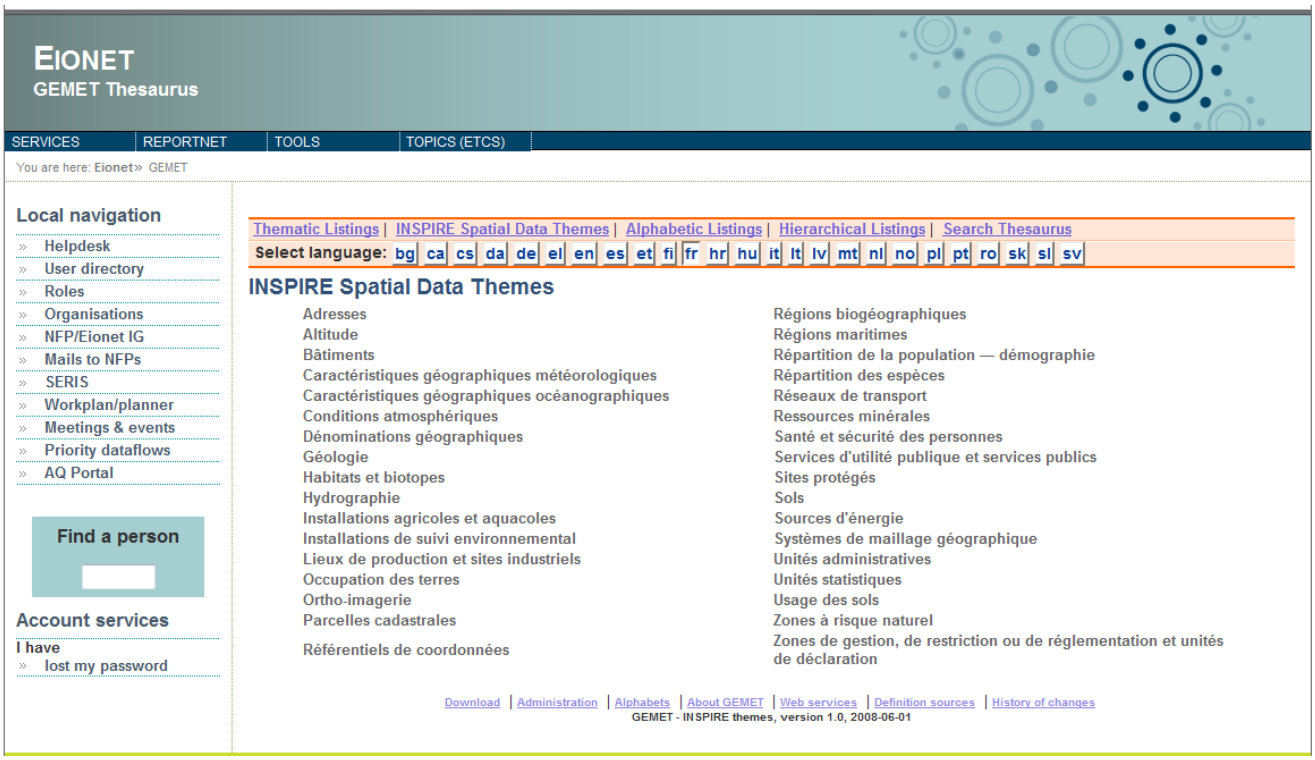
Preserving software versions is to be considered when proprietary formats are used and if the data is to be reproduced over time..

15. PRODUCING A DATA MANAGEMENT PLAN

Structure of a Data Management Plan (DMP)[7]	
Information about the project	Administrative information + see section 4 (p. **Erreur ! Le signet n'est pas défini.**)
Accountability for the data	See section 5 (p. 9)
Resources needed to implement the DMP	Make an estimate of human resource and archiving costs in a repository
Datasets All sections below should be duplicated for each dataset, i.e., a technically homogeneous series of data or deemed to be intellectually consistent.	
Dataset description	See sections 7 and 8 (p. 11 et 13)
Storage, access and data safety	See section 3 (p. 7)
Metadata: data documentation and organisation	See sections 2, 3 end 8 (p. 6, 7 et 13)
Availability of datasets	See sections 9, 10, 11, 12 (p. 15, **Erreur ! Le signet n'est pas défini.**, 18, 19)
Data selection and archival	See sections 13 and 14 (p. 20 et 21)

7 DMP structure excerpted from Cartier A, Moysan M, Raymonet N. Réaliser un plan de gestion de données : guide de rédaction.Retr (VO1 09/01/2015)

ANNEX 1



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

ANNEX 2

Persistent identifiers

for research data

What is a persistent identifier?

A *unique identifier* is unique identification code assigned to an object (or a person) in a way that this object can be identified unequivocally[8]. Most of the time, this identification code is an alphanumeric character string.

A *persistent identifier* (PID) is an identification code assigned to an object once and for all[8]. It can be used and managed on a long-term basis and it will not change even if the object is renamed or moved about. (e.g., moved to another website or repository). Usually, URLs are used as identifiers, but there is not guarantee that they are persistent.

What is the use of persistent identifiers for research data?

- ✓ DF are on the web, even if its URL changed (e.g., changed website, changed institution or repository), and thus avoid hitting on this message: HTTP 404-File not found.
- ✓ To make research data as findable as publications (as persistently as its producer wishes it).
- ✓ To facilitate research data citations.
- ✓ Link research data and scholarly publications.

8 Adapted from Borremans Catherine (2012). Minutes of a Workshop on persistent identifiers (Atelier sur les Identifiants Persistants (PIDs) organised GBIF France held in Paris (MNHN) on June 27, 2012. http://archimer.ifremer.fr/doc/00119/22978/ Retrieved on January 4, 2013

Major identifying systems used for research data

System	Full name	Example
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

* a persistent identifier base on the Handle system

There are other persistent identifier systems like:

- URN : NBN National Bibliographic Numbers (e.g., urn:nbn:nl:ui :32-424171),
- LSID, Life Science IDentifier: an identifier used by the Life Science community.

Illustration of how datasets are linked to the publication in which they are cited