Procedure for data and software integrity

Department of Geoscience & Engineering
Department of Geoscience & Remote Sensing
TU Delft

Jan Thorbecke and Roderik Lindenbergh, 10 September 2015

1. Introduction

This document describes a procedure for storing research data. The main goal of the procedure is to make research results, based on measured, collected and/or numerical computed data, reproducible and accessible. Data which is essential to prove the integrity and ownership or your research and results must be retained in a durable and appropriately referenced form for at least 15 years from last publication. After 15 years the dataset will be evaluated and possibly stored for a longer period.

A secondary goal of the procedure is to provide scientists with reliable ways to exchange research data. Access to retained data may be given to colleagues world-wide. This may give our researchers additional benefit on their data as world-wide usage may lead to further cooperation, while users may also give credits to the owners of the data in the form of a data citation. In addition, storing essential research data in an institutional repository may improve on current practices of data sharing in some TU Delft research labs.

University must pursuit to make all software open-source licensed and operate in an open-source environment. Relying on commercial products make the data and software vulnerable for changes that are not-backward compatible. At the same time a researcher who wants to reproduce the results is not forced to buy an expensive license. Still, it has to be realised that at the time being many researchers use commercial products like Matlab as tool for their research.

Over time computer systems and files may need to be moved, software and technologies may change, file formats and metadata schemas may become obsolete. Data maintenance includes the actions required to ensure that data can be discovered and used by researchers over time.

For storing research data, use will be made of the 3TU.Datacentrum, a cooperation between the three technical universities of The Netherlands. This version of the documents still mainly originates from internal discussions at the Geoscience departments. A next step is to harmonize the department wishes with the possibilities offered by the 3TU.Datacentrum. The 3TU.Datacentrum is interested in our initiative and welcomes cooperation, notably as the Geoscience departments are among the first departments at TU Delft starting to implement a data integrity procedure.

2. Which data to store?

In this chapter we shortly discuss what type of data is distinguished and what data should be stored to ensure data integrity. Metadata will be discussed below.

Collected data

Direct observations consists of data collected directly by the researcher itself. Examples of such data are

- Observations made in a lab during an experiment
- Field observations, e.g. digitized core samples, remote sensing data, seismic passive or active data, collected by the researcher
- Interview records, i.e. a digital transcription of interviews done by the researcher

These type of observations are unique to the researcher and should be stored.

Data obtained from external providers

In notably the Remote Sensing community, researchers will use observations made available by national or international agencies. Examples of such agencies are PDOK, which is a central Dutch facility for distributing digital geoinformation and ESA and NASA, the European and American space agencies, resp. Typically data distributed by such agencies have a clear ID. For that reason it is not necessary for the researcher to store that data, but it is enough to refer to the ID of the data used.

Action: the previous only holds if agencies guarantee the availability of the data they provide for the period for which research reproducibility is guaranteed (e.g. 5 years, 50 years?). This guarantee should be discussed with the agencies.

Often also non-standard products are obtained by researchers from agencies or companies for research purposes. As such data is not reproducible, these should be stored by the researcher in the repository, and agreements should be made with the provider of the data on the type of access for externals (open or restricted access).

Software and data derived by software

Typically, researchers use software to extract information from input observations which results in derived data products. Researchers also design or work with simulation models that generate derived data products. The result of a research may also just be a software product, without input or output data.

It is not required that researchers store derived data products, but in all cases the researcher should store the source code that was used to obtain the derived data products or software product. The source code can be a script in an interpreted language like Python or Matlab, or the source code for a compiled language like C++. The software should preferably be written in a language that does not require a commercial licence like Fortran 90, C or C++. Currently the use of commercial software like Matlab is so widespread that it doesn't seem realistic to forbid that. When commercial software is used a detailed explanation of the method should be a given, with the aim that other researchers could write their own code based on the commercial script and the detailed explanation.

Problem: what should researchers store when they use a GUI to obtain derived data products? Sometimes there may be a log file storing their actions or they could make a movie recording their actions?

3. Metadata requirements

For all data and source code to be stored, metadata should be provided that describes the data in a uniform and computer interpretable way. For different type of data, different metadata is needed as specified below. First metadata is described that is required for all stored data and source code.

General metadata

For all stored data and source code the following metadata fields should be stored

- Title
- Year
- Associated publications
- Publisher
- Names and roles of any team members who worked with the data
- The purpose behind the research
- Implications of the research and future directions
- The owner of the data or software
- List of keywords
- Automatically link ID of the publisher to other ID's (ORCHID, Google Scholar, …)

Collected data

Data collection guidelines and methodologies should be carefully developed before the research begins. The researchers must determine what sort of data will be collected and how this data will be analysed. For data to be considered reliable, data collection should occur consistently and systematically throughout the course of a project.

The collected data contains metadata describing

- Methodology of data collection, measurement plan
- Implementation of this methodology
- How data was collected and analysed in practice
- If unexpected results or significant errors were encountered

For this part of the metadata the researcher may also refer to an accepted publication describing the data collection methodology.

During data collection (electronic) records should attempt to accurately represent the progress of a project and answer such questions as what, how, and why data were collected or amended.

Recording information: Record anything that seems relevant to the project, its data, and the standards of the project. At a minimum, records should include the following information:

- Date and time of the data acquisition
- Used materials, instruments, and software, including version
- Precise instrument settings and use

- Identification number(s) to indicate the subject and/or session. For example, If there are repeated experiments: each experiment must have its own ID (maybe the time stamp can be used for that purpose?).
- Data from the experiment and any pertinent observations from the data's collection. For example, when the data has some special features (e.g. on a warm day) this should also be described.
- The format of the data
- The dimensions of the data
- The spatial domain of the data (if applicable)
- Used reference systems like geoids

When transferring records from written to electronic format, use a double entry system to reduce rates of incorrectly entered electronic data. ****no idea what this means??****

Data obtained from external providers

- Provider of the data, preferably including contact details
- ID of the data, if applicable

All recording information, as specified for Collected Data above.

Software and data derived by software

For software and software scripts the following metadata is required:

- Programming language, including version
- Programming platform, if applicable
- Supported input and output formats
- Dependencies on software libraries
- A man page in ascii format, including syntax, options and examples.
- Supported Operating Systems, including version and distribution
- A minimal test data set on which the functionality of the software is demonstrated

The software must preferably be run on an open source operating system. The open source operating system can be stored with the software itself to be able to reproduce the data many years (>10) after it has been created.

All software must be documented, in the man pages, in an accepted publication, or in an additional technical report. The documentation should be stored in a format that will remain valid after many years. e.g. no Microsoft Word or PDF, but plain ASCII files (and for example typeset in LaTeX)

To discuss:

- *are all programming languages and data formats required?*
- *Should submissions to the repository be moderated, i.e. always approved by a lab-leader or professor?*

4. Data storage

Requirements:

- Integrity (accuracy and completeness), of all data is maintained during transmission and storage;
- If not specified differently, the University has data ownership;
- If not specified differently, access rights are open to everybody;
- Each user is responsible for all transactions resulting from the use of his/her log-in username and password to the data repository.
- Data protection to avoid unwanted corruption of data;
- Availability of data meets University requirements for operations and can be restored after loss or damage by accident, breach of security, or natural disaster;
- Retention, how long must we store the data? Use of hierarchical storage (slowly moving data from hard disk to tape/DVD/cloud);
- Store data as much as possible in prefered data formats supported by 3TU

Discussion: here the focus is on reproducibility of published research results. Data storage solutions can also be used to improve data sharing.

5. Storing data at 3TU.Datacentrum

Research data will be stored at the 3TU.Datacentrum. From their website:

3TU.Datacentrum offers the knowledge, experience and the tools to archive research data in a standardized, secure and well-documented manner. It provides the research community with:

- *A long-term archive for storing scientific research data*
- *Permanent access to, and tools for reuse of research data*
- *Advice and support on data management*
  3TU.Datacentrum currently hosts thousands of datasets. To see examples please visit: http://data.3tu.nl.

Current implementation at 3TU data-center

(Status: spring 2015, to be updated)

- Simple data sets are easily uploaded through a web interface. The upload-size is currently limited to 4GB.
- Complex or large data sets require help from 3TU data-experts.
- All uploaded data is checked for consistency before it is placed into the database.
- Access to data is unrestricted, with a possible limit of 2 year to protect a publication.
  - to get access to the data a user must make a (free) login account to 3TU
- List of prefered formats:
  - NetCDF
  - .
- The prefered formats are automatically converted to the latest available version.
- The data center is mainly focussed on data sets and has less experience with storage of software and processing scripts.
- Direct links from web-page to doi-data-object is possible.

- Courses are available to assist users with data handling and storage.
- Indication of costs: €3700 per TB for 15 years

Data storage example

An example of one our data sets stored at the 3TU.Datacentrum can be found by following this citation:

Beril Sirmacek (2015)
Delft Windmill Interior and Exterior Laser Scanning Point Clouds.
TUDelft.
Dataset.
http://dx.doi.org/10.4121/uuid:daea472d-2ca5-4765-9f1b-bd3200de4b41

6. Further research data initiatives in the Netherlands

KNAW

The KNAW wrote a report on Responsible research data management and the prevention of scientific misconduct: http://www.knaw.nl/en/news/publications/responsible-research-data-management-and-the-prevention-of-scientific-misconduct/@@download/pdf_file/20131009.pdf

They also offer EASY, the online archiving system of Data Archiving and Networked Services (DANS). From the website: “EASY offers access to thousands of datasets in the humanities, the social sciences and other disciplines. EASY can also be used for the online depositing of research data”
https://easy.dans.knaw.nl/ui/home

Open Earth

OpenEarth is a free and open source initiative, organized by Deltares to deal with Data, Models and Tools in earth science & engineering projects, currently mainly marine & coastal:
https://publicwiki.deltares.nl/display/OET/OpenEarth
They also provide a page with links to different data sharing initiatives:
https://publicwiki.deltares.nl/display/OET/Data+sharing+initiatives

7. Research data initiatives outside The Netherlands

For a later version of this document explicit references to existing standards for data and software storage should be included.

Re3data.org

In August 2012 re3data.org – the Registry of Research Data Repositories went online with 23 entries. Two years later the registry provides researchers, funding organisations, libraries and publishers with over 1,000 listed research data repositories from all over the world making it the largest and most comprehensive online catalog of research data repositories on the web. re3data.org provides detailed information about the research data repositories, and its distinctive icons help researchers easily identify relevant repositories for accessing and depositing data sets.
http://www.re3data.org/

Inspire

Insipre stands for Infrastructure for Spatial Information in the European Community:
http://inspire.ec.europa.eu/

University of Melbourne

The University of Melbourne has a research data management plan:
http://researchdata.unimelb.edu.au/

University of Bieleveld

https://data.uni-bielefeld.de/en/researchdata