Data storage fact sheet manual

Version March 12th 2014

(see Appendix 1, Glossary, for a definition of some of the core concepts that are used in the manual)

Introduction

The data storage fact sheet (DSFS) is a document containing information about (raw, processed, and meta) data that have been stored. A DSFS can be linked to a publication on Biblio (https://biblio.ugent.be/). The aim of the document is to provide basic information about the storage of the data that are described in the publication, to provide information about who can be contacted for additional information about the data, and to provide a checklist for researchers concerning issues that can be relevant when storing data. Researchers can also use it as a tool to document the details of data storage so as to facilitate their own future access to the data as well as possible access by others.

The DSFS is stored as a standard plain text file in order to maximize access (e.g., independence of the software one is using) and durability (e.g., independence of future developments in software). Text files can be imported in various word processors and text editors. After revisions, please store as "unformatted text file" (.txt). Only make changes by adding or deleting text characters (e.g., checking boxes with an X). Do not use word processing functions such as "highlight" or "bold" because they will not be stored when saving the file as an unformatted text file. To avoid automatic formatting, do not change the DSFS in word processors such as Word but use text editors such as WordPad or NotePad++ (windows) or TextEdit (mac).

Using the DSFS

General Information

Although researchers can use a DSFS to document the data storage of unpublished data studies, it was designed to be linked to publications that are posted on Biblio. A publication in Biblio can be linked to one DSFS or multiple DSFSs. Hence, a DSFS can apply to all studies in one publication or to only some studies in a publication. It is also possible to create several DSFSs for a single study, with different DSFSs referring to different datasets in a single study. Researchers are free to decide how many DSFSs are linked to one publication but are advised to always provide clear information about which studies and datasets in which publication a DSFS applies to. When different DSFSs are used for different datasets in a single study, it should also be made clear which datasets that a DSFS applies to. When the same study is reported in multiple publications (e.g., both in a journal paper and in a PhD dissertation), it is best to link different DSFS to the different publications in order to ensure that for each publication there are DSFSs that contain clear information about the studies and datasets within that publication.

Header

The header of the DSFS consists of 4 parts:

- Data Storage Fact Sheet: This is the general title that does not need to be altered
- Name/identifier study: The name or identifier as given by the author of the study. The author can chose whichever name or identifier he or she seems fit.
- Author: This is the name of the person who filled out the DSFS. This is typically (but not necessarily) the main researcher or responsible staff member.
- Date: The date on which the DSFS was created.

1. Contact Details

The section on contact information provides information about the researcher who conducted the study (typically the first author of a publication), the responsible staff member (ZAP), and default contact details:

1a. Main researcher: this part contains the name, postal address and e-mail address of the researcher who conducted the study. Please do not list contact details that will expire within less than six months (e.g., the email address of a PhD student who has decided to leave academia). Feel free to provide multiple contact details with an indication of the time at which they are valid (e.g., also the new email address of a PhD student who will soon move to another university).
1b. Responsible staff member (ZAP): Please provide the name and contact details of the staff member (ZAP) who has academic responsibility for the research (e.g., the supervisor of a PhD or holder of the grant that funded the research). This section does not need to be completed if the main researcher is also the responsible staff member. Again, please provide contact details that remain valid for the foreseeable future.

Default contact: In case unforeseeable changes occur in the contact details of the main researcher or responsible staff member, default contact details are provided that are likely to remain valid in the foreseeable future. These default contact details should not be altered by the author of the DSFS.

Researchers are free to list the contact details of other researchers (e.g., postdoc co-promotors) but are encouraged to always give the contact details of the responsible staff member, in part because these contact details are more likely to remain valid for long periods of time.

2. Information about the datasets to which the sheet applies

First, please provide the official reference to the publication, in line with the current APA guidelines. Because DSFSs are meant to be linked to a publication on Biblio, an official reference will typically be available. However, in case that the DSFS is used for data that is as yet unpublished, researchers are free to list a reference to unpublished data (see APA guidelines).

Second, please describe in a clear manner which studies or datasets the information in the DSFS applies to. For instance, when linked to a journal publication, please provide the numbers of the studies it applies to (e.g., Study 1; Studies 2-4) or state explicitly that it applies to all studies that are reported in the publication. When linked to a dissertation, please also make sure that it is clear which studies from which chapters the DSFS applies to (e.g., Chapter 1, Studies 1-4; Chapter 2, Study 3). When there are different datasets within a study that are stored in different ways (e.g., both digitally stored reaction time data and paper-and-pencil questionnaires), a different DSFS can be created for each dataset. Alternatively, Section 3a (raw data) can be duplicated within a single DSFS, provided that it is made clear which section applies to which dataset.

Because a DSFS will typically be linked to a publication in Biblio, more information about the studies will be available via Biblio (e.g., DOI, Web of Science id) and therefore does not need to be specified in the DSFS.

3. Information about the files that have been stored

During the scientific process from data collection to publication, most often different types of data and meta-data files are generated. In this section, information is provided about which data have been stored, the way in which these data have been stored, and who has access to those data. Whenever possible, researchers are advised to store all the information that is necessary for a reproduction of the results that are reported in the publication to which the DSFS is linked.

3a. Raw data: In a first subsection, information is provided about the storage of the raw data. First, indicate whether the main researcher has stored the raw data. If this is not the case, please provide a brief explanation for why this is not the case (e.g., the raw data were collected by and are owned by a third party). Second, indicate the platform on which the raw data are stored. Please list multiple platforms if the raw data are stored on multiple platforms. Third, indicate who has direct access to the raw data (i.e., without intervention of another person).

3b. Other files: In a second subsection, information is provided about other files that have been stored. On the DSFS, different types of files are listed that typically are important to allow for a reproduction of the results:

- files describing the different steps via which raw data were transformed into the reported results (e.g., R syntax files, coding schemes)
- files containing processed data, such as a subset of the raw data (e.g. less but identical variables than the raw data), an aggregated version of the raw data (e.g. data set containing cleaned data, aggregated for analysis), or a transcript from audio files
- files containing the outcome of analyses, such as outputs of statistical programs
- files that describe the content of the stored files and how this content should be interpreted; this includes files about the nature of the raw data (e.g., number of files, type, format, content, organization) and the way in which they have been collected (e.g., hardware, software + version) but also files about the content and interpretation of other stored files such as files containing processed data

In addition, the DSFS lists two types of files that provide important information for researchers who might want to re-use the data: First, files can be stored that specify the information that participants were given when agreeing to take part in the study (e.g., a blank copy of the informed consent form). If participants were not asked to provide informed consent, a file can be added that specifies the context in which the data were collected. Second, files that specify legal and ethical provisions can also be stored. These could contain information about who owns the data as well as legal or ethical constrains on how the data can be used. Finally, there is the option to describe other files that were stored.

A minimal use of the DSFS entails that the author of the DSFS indicates for each listed type of file whether such files were stored. However, using the option "specify", authors can also opt to provide for each type of file more information about the files that have been stored (e.g., exact names of the files and location of the files, their format). Providing additional information increases the value of the DSFS as a tool to improve data management.

Finally, authors are asked to specify where the other files are stored and who can access them directly. If this differs for different files, authors can duplicate these sections and specify which information is valid for which type of file.

To guarantee long-term digital preservation, it is advised to store all files using open data formats (if feasible). For more information about open data formats and long term storage see http://www.data-archive.ac.uk/create-manage/format/formats

4. Reproduction

DSFS authors are given the option to indicate whether the results that were reported in a publication have been reproduced by others. Typically, these other persons are co-authors of the publication, but they could also be statistical consultants or fellow researchers at other universities. Reproduction has occurred when the other person was able to reproduce the reported results starting from the raw data. Reproduction is possible only for certain types of research but if it is possible, it can provide researchers (including the main researcher) with an indication of the quality of data storage and the validity of the reported results.

Appendix 1: Glossary

Data: All items that have been generated from or are used in a certain study.

> ***Raw data*** (or primary data) refers to the unprocessed source information as it is provided by a research instrument.
>
> ***Processed data*** refers to data that are derived from the raw data after processing these raw data. Processing implies each form of operation on the raw data, such as deleting certain raw data, calculating averages based on raw data, or making transcripts based on raw data (e.g., audio recordings).
>
> ***Meta-data*** (of data about data) refers to structured and standardised descriptions that present information about the nature and form of the data, as well as information about the collecting, processing, storage and deposition of (raw) data.

Data storage*:* the storage of data (own PC, server, file drawer, archive, etc.) in such manner that data do not get lost. Data storage is a necessary precondition for data deposition and data sharing, but also has to realized for those data that cannot be shared or that are difficult to share. Data storage meets three basic criteria:

1) Confidentiality (data are accessible only for those who are authorized);
2) Integrity (data cannot be deleted or altered, either deliberately or accidentally; data are up to date, correct and complete);
3) Availability (data are easy to access and use by authorized persons within a stable environment).

Data deposition*:* the safe and future-oriented (preferably electronic) storage of data in an archive. Data deposition entails that a record is kept of when and by whom the data where filed in the archive. Authenticity of the data in the archive can be verified by means of – for example - a checksum that is stored at a different location.

Data management*:* All actions that result in keeping the research data safe, traceable, accessible and understandable. Data management is involved in the planning of studies, in collecting, organising, storing, processing, analysing, and archiving of research data, in making research data accessible and potentially available for re-use.

Data management plan*:* A structured document that describes how data will be collected, the magnitude and format of the data, the way data will be stored and processed, and who will have access to the data both during the implementation of the study as well as afterwards. The document captures how researchers manage the collected data before, during and after the study. Such a plan also contains information on the legal and ethical aspects of the data.

Dataset: A set of raw data that are processed in order to obtain results.

***Data sharing**:* Making data available to the research community with the aim of gaining an optimum (re)usability of research information for science and society.

Re-use: The use of data for other analyses than those reported by the researcher in publications or reports.

Reproducibility: The possibility to arrive at the results reported by the researcher in publications or reports, starting from the raw data.