Karolinska Institutet

Department of Medical Epidemiology and Biostatistics (MEB)

Guideline

Version 4

2013-10-22

Guidelines for Documentation and Archiving

of Computer Media Files in Research Projects

Revision history

2013-10-22	Anna Johansson and Bill Österlund for the Data Management Council (version 4).
2011-10-10	Anna Johansson for the Data Management Council (DMC) (version 3).
2006-11-14	Denny Rönngren and Anna Johansson (version 2).
2005-10-28	Anna Torrång and Gunnar Petersson for the Data Management group (version 1).

The latest version of this document can be found at http://intra.meb.ki.se.

Table of contents:

1. Introduction

As we all know, it is of great importance in the research community that research is reproducible. An analysis done today within the department should be reproducible after ten years or more. Also, old data must be comprehensible even when staff leaves and new persons continue on the same research project. The purpose of this document is to provide guidelines on how to document and archive computer files used in research project. It is of importance to the department, to research groups and to individual researchers that staff involved in the research at the department adheres to these guidelines.

The guidelines presented here cover two main areas:

1. How to structure and document your everyday work to make easy archival possible when you are finished with your thesis.
2. How to prepare for the actual archiving of your research.

This document is mainly written for PhD students but also for their supervisors, support persons, principal investigators and other working within research projects. Definition of a project is within this document the entire research work of one PhD student, which results in manuscripts and finally in a thesis.

Aims of these guidelines:

(1) The primary aim of these guidelines is to ensure that research publications and results are reproducible. This also includes publications where the original data and electronic files are not stored on MEB servers.
(2) The secondary aim is to ensure that research data stored at MEB can be read, understood and accessed in the future.

If you choose not to follow this guideline:

You are free to create your own way of documenting your research (i.e. it is not compulsory to follow this guideline). The only requirement is that you document your research in a manner that follows Swedish law and KI general guidelines. If you choose to create your own documentation structure, you must describe it in a readme.txt file stored in the folder your data and files are stored.

If you follow this guideline, you may simply refer to it in your documentation files.

1.1 General laws and rules

All research projects at the department of Medical Epidemiology and Biostatistics (MEB) are subject to Swedish law and rules at Karolinska Institutet (KI) and guidelines for archiving of documents. With documents we refer to all documentation which relates to a research project, e.g. the research plan, ethical application and approval, publications, annotated manuscript, web questionnaires, databases and analysis programs. Guidelines can be found at:

- http://internwebben.ki.se/en/guidelines-research-documentation
- http://internwebben.ki.se/sv/arkiveringskrav-av-forskningsdata

Research projects financed by foreign organizations may have additional rules e.g. NIH.

The documents must be stored for at least 10 years (important documents for much longer, no time limit), and documented and saved in such a way that other staff can reuse the research documents within this time.

This document deals with electronic research documents. Hard copy paper documents are handled by the MEB admin group and the Archiving Plan (Arkivbildningsplan).

Please consult the above document or the MEB admin group for further information regarding hard–copy archiving.

1.2 Responsibilities

Each PhD student is always responsible for documenting and archiving the electronic files in his or her project. The principal investigator or supervisor is responsible for ensuring that this work is done. Students who have defended their thesis, but not yet published all their papers when they leave MEB, are still responsible for ensuring that the archiving of the final papers is done. The archiving can be delegated to another staff member if the student is not present himself/herself to do this in person. If the student is working on data not stored on MEB servers (but elsewhere) then the student is still responsible for making sure that published results can be reproduced, for example by providing a reference to where the original data and analysis files are located.

IT staff are responsible for providing technical advice on the practical procedures for documenting and archiving the electronic files and they are also responsible for maintaining the technical procedures for archiving and restoring the electronic files.

Responsibility for the documentation and archiving of research projects lies with the PI and ultimately MEB (Head of Department). Even in the event that the PI leaves the department, MEB has the responsibility to maintain documentation and copies of the files.

2. Guideline for documenting everyday work

The basis for archiving is a good documentation strategy, which should be implemented during the whole lifetime of the research project. Such a policy, correctly used in everyday work, will yield just a few key documents that clearly describe the study and its related files.

The overall guideline is as follows:

1. Create a good file and folder structure from start. This will help you to manage your files, and will probably make it easier for you to archive your final results.
2. Continuously document your work. This includes documenting datasets used, commenting program code, saving copies of data request submissions etc.
3. Summarise, document and comment files when manuscripts are submitted for publication, e.g. comment the analysis programs, which code was used to produce the different results, tables and figures in the manuscript, etc. Always make sure you (and others) can backtrace the submitted results to the programs that created them.
4. Prepare your work for archiving when a manuscript is accepted and in press. This is the important step to guarantee the reproducibility of your results; after all, you want to reproduce the results exactly as they were published. If you need to continue to work on the same files, then you need to do the archiving preparation first, before you modify the files.
5. Archive your files (i.e. final move of archive folder) when your employment at MEB ends.

2.1 Good folder structure

In order to make it easy to find project files, we recommend that you create subfolders for different types of files (Figure 1)

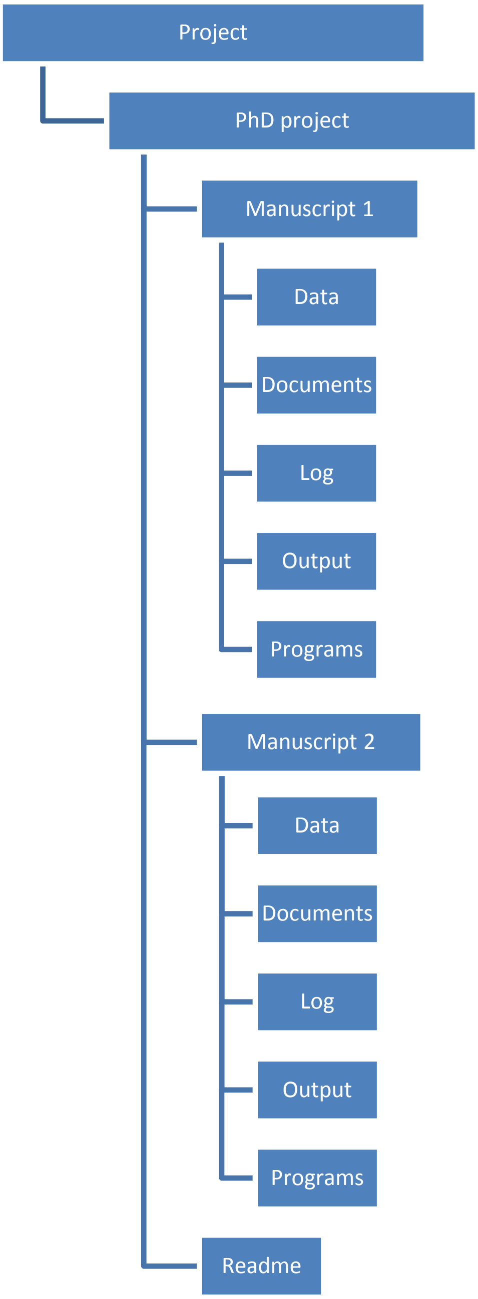

Figure 1. A recommended folder structure for files in a research project.

The main idea is to divide your projects into smaller manageable parts, for instance, keep all your raw data in one separate folder and all the documents in another folder. Program files, analysis output, and log-files should also be stored in separate folders. For projects with very few files, you might put programs, logs and output in the same directory, if the folder does not become too cluttered by this. See Figure 1 for our suggestion on how to structure your folders. A read-me-file with details regarding folder structure and valuable information about the project should be in the top folder of project.

If you use several software programs for data analysis, you may want to save the program code files from each software program in different folders, e.g. Programs/SAS, Programs/Stata. In the Documents folder you may wish to create subfolders for Documents/Data documentation and Documents/Manuscript versions, etc.

Here we give a brief explanation of what the folders are expected to contain:

- Data—data files (e.g. SAS datasets, Stata datasets, R datafiles)
- Documents—documentation files such as the research plan, final version of the ethical application and approval of that, the analysis plan, logbook-file, codebook-files, manuscript versions, important communication, etc.
- Log—log files from data preparation and analyses (e.g. SAS logs (.log))
- Output—output files from data preparation and analyses (e.g. SAS output (.lst), Stata output, R output)
- Programs—program files from data preparation and analyses (e.g. SAS programs, Stata do-files, R programs)

2.2 Documenting your work

Documentation is necessary for understanding your work. Good documentation benefits everyone, including yourself when you have not worked on your project for a while. Everyone knows how hard it can be to remember which file was used for what analysis or where they are located. Of course, there is a trade–off between the time spent on writing documentation versus time saved getting up to date with a project—try to keep it at a reasonable level.

A general guideline is that one of your colleagues, who is not familiar with you project, should be able to sit at your computer and understand the organisation of your work. For example, after being directed to the main project folder your colleague should be able to navigate through your project without assistance from you and, for example, reproduce your analyses. When you are documenting your work, put yourself in the shoes of the person who will be continuing work on your project after you leave. Even if there are no immediate plans for someone else to continue your work, someone will probably do so. The data you leave behind after completing your thesis probably contain enough information for additional scientific papers. New hypotheses arise every day and it may just happen that your data are perfect for investigating a newly formed scientific hypothesis.

There are several MEB documents describing how to document your work, these include:

- MEB Data and Variable Standard
- Take good care of your data

A good start is to have a readme–file at the root of each folder which describes what the files and subfolders located there contain.

Here follows a list of some of the files that should be created and what they are supposed to contain. The purpose of these files is to provide semantic information on what your files and datasets contain.

Create the following documentation files:

- **Readme–files**—information about what a folder contains (see Appendix 1 for example of a project readme–file). A readme–file is simply a textfile (.txt) with basic information regarding the folder content. It should always contain information about who created the files and folders, and when. The easiest way to create a readme–file is to use the Notepad program which can be found in Start menu on all PC computers.
- **Logbook–file** —detailed information about data files (datasets) you have created, from what source they were derived, and what program files were used to create this data file. Logbooks should be stored in the documents folder for each manuscript.
- **Codebook–files** —detailed information about one data file, such list of variables and description on coding of each value. Create a separate codebook-file for each data file. Should be stored in the documents or data folder.
- **Annotated questionnaire, or similar** —electronic version of the questionnaire with the variable names written at each question should be saved. If you have other abstraction forms, annotate them in similar manner. The point being to link the electronic version of data back to the original source.
- **Analysis plan** —detailed description of each manuscript, e.g. hypotheses, data sources, outcome and exposure variables, data management, statistical analyses, contact persons. It may also contain references to program files and output files, meeting minutes and decisions, etc.

Logbooks, codebooks, annotated questionnaires and analysis plans can be created using Word, or similar.

When you prepare to submit your results to a publisher, you must keep track on exactly which files and steps were used to produce the final result. It is a crucial requirement for the necessary reproducibility of your results. This is also an opportunity for preparing the archiving step as well. Try to keep only the files that are actually useful, discard old working copies and such.

3. Archiving your work

Archiving is the process of documenting and storing your work in such a way that someone else can rerun your analysis and understand the results. Someone else in this case, is of course someone with the same academic background as yours. The purpose of archiving is actually twofold—to make your results reproducible and to make it possible to continue working on your project at some point of time in the future. According to this purposes, archiving your work be done at typical endpoints such as:

- When your paper is published
- When you are finished with your PhD thesis
- When you are leaving MEB
- When the project "dies"

The documents to be archived should be organised in a standardised way for each manuscript in a research project. When the time comes for archiving you should contact someone in the data management group to get approval of your folder structure and documentation. If the research project is finished (for example after dissertation) the folder structure (i.e. the whole PhD project folder) will be copied or moved to a central archiving area. If you have followed the recommendation listed so far, this should be an easy transform from your working copies to the final archive copy. Usually there are just some additional steps needed—writing some more documentation, and converting documents and data files to more widely accepted formats.

3.1 Folder structure for archive

The folder structure suitable for archiving is mainly the same as your cleaned working folder (see Figure 1).

Additional to this, every project area should have an archive subfolder. This Archive subfolder should contain a readme subfolder, which should contain information for the whole project, e.g. research plan, project flow chart. There should be a readme–file that describes the basic facts for the project in a mandatory structure (see Appendix 1). The archive subfolder should be divided into subfolders for each manuscript and each manuscript subfolder should have a readme.txt file that describes the basic facts for the manuscript in a mandatory structure.

The programs subfolders should contain a readme–file which describes what the programs do, the input and the output and for what analysis/table/graph in the manuscript. It should also contain information regarding folder path references (in SAS: libnames) used in the programs. Remember that each program file should be properly commented within the program. See Appendix 2, for suggestion on documentation of a SAS program file (same idea applies to other software).

The data subfolder should contain all data files used for the analysis. The storing of data files for each manuscript is mandatory for several reasons. First it must be possible to rerun the analysis and secondly one must be able to rerun it on the same "timestamp" data (databases are constantly updated). If you run into problems with hard drive space, contact the IT–support.

3.2 Format of the documents at the final archiving of a project

All files must be readable ten years from now. In the computer world this brings up two problems. Firstly, new software programs and formats are emerging all the time and sometimes they are not backward compatible (i.e. a file created with an old version of a program may not be accessible with a newer version of the same program). Secondly, the data media must be readable after ten years but it might be corrupted or the technical possibility is missing e.g. the equipment is gone and destroyed. The second problem is directed by IT–staff but the first problem concerns each user. These problems and recommendations are also described by Riksarkivet in their advice publications RA-FS 2003:1, 2003:2 and can be found at http://www.ra.se/ra/rafs.html.

The recommended way to be sure that the files will be readable in the future is to save data files as plain text file (.txt). In most programs it is possible to choose: File save as -> text file, or to export to a text format. When you have saved a file as text file, try to open it in Notepad and see that it is readable. If it is not, then contact IT support[1].

The following rules apply to MEB:

1. All files should be stored in their native formats, i.e. in the original filetype. This rule does not apply to Oracle, MySQL and SQL Server relational database files
2. All native files with native formats should also be transformed to a common readable format. See Table 1 for common file types and their suitability for archival at MEB. Note that common file types such as Word .doc files and Excel .xls files are not approved.

Table 1. Approved and non-approved file formats for archiving.

Program	Extension	Approval status
ASCII	.txt	approved
Microsoft Word	.doc	not approved
Microsoft Word	.rtf	approved
Adobe Acrobat	.pdf	approved
Microsoft Excel:	.xls	not approved
Microsoft Excel	.csv	approved
XML	.xml	approved
SAS	.sas,.log,.lis	approved
SAS	.sas7bdat, etc	not approved
STATA	.do	approved
STATA	.dta	not approved
R	.RData	not approved
Microsoft Access	.db	not approved
Filemaker Pro	.fm,.fp3,.fp5,.fp7	not approved

[1] There are several different encodings for text representation in those files, and standard usually referred to are ASCII, different ISO text encodings and Unicode. There are differences in how they handle national special characters (e.g. å, ä, ö), and if your data contains data out of the common English character set, you need to understand the implications of choosing a particular text encoding. IT support will be able to help solve this problem for you, if your file is not readable in Notepad.

In short, all analysis data files should be transformed to text-formatted files like .xml or .csv files regardless of their native (program specific) format. Normally this is not a problem because many programs already have the capability to store data in several formats. A practical guideline: If you can read the file with Notepad it is OK.

A word of caution: Be careful with text strings, native letters and date formats, they tend to create problems. If the text-file is of non-delimited format it should be very well documented. Always check your output! If in any doubt, please contact IT support.

3.3 Procedure for archiving

The process of archiving your work can be summarised as follows:

1. Clean your folders. Create a folder structure that suits the project, see Figure 1 for an example. In this process it is important to clean among the files–delete old "irrelevant" files.
2. Provide additional documentation if you have not done this already. Create simple text files named readme.txt, which describe the contents of the folders. Readme–files should also contain information about who is responsible for the data, contact persons and contact information, date of archiving and other information that is good to know about for persons who want to use the archived material. See template example in Appendix 1.
3. Contact someone in the data management group (preferably the same person who did your last follow–up), for a final check of the documentation and folder structure.
4. Contact IT support for the actual moving of files to an archiving disk area. If the project is finished (for example after dissertation) the folder structure will be moved to a central archiving area.

Checklist of research project files to be archived:

1. Research Plan and the final version of Ethical Application and the Approval.
2. If raw data were extracted from a registry or database at MEB, you need to provide the exact source and the code and definitions used (e.g. extraction description including the script and log for extraction.).
3. If raw data were extracted from registry or database constructed at MEB which are not one of the standard cohorts, you need to provide full documentation of the application and database.
4. If raw data were obtained from a National Registry (SoS or SCB), a copy of the original data request, showing definition of variables.
5. Copy of the analysis data and extracted data used in the analysis and the program code used to create it. Logbook-files and codebook-files for the data.
6. The final Analysis Plan, which in detail describes the hypotheses, variables and the analysis methods.
7. Program code that will produce the published results when run on the analysis data set.
8. Output and results from the last run of the analysis.
9. Electronic copy of manuscript.
10. Important e-mail correspondence to and from the journals.

In addition to this checklist there are KI general rules about what you must archive. See the links in the introductory chapter.

4. How are these guidelines implemented at MEB?

This document is published on the MEB intraweb.

Every PhD student should, on arrival, be shown this document and informed of the importance of good data management practice. Once per year, in connection with the compulsory annual review conducted for all postgraduate students, the data management group conduct a review of data management aspects for postgraduate students working on projects where MEB is responsible for archiving (i.e., students working with data solely at other organisations, such as SMI, will not be reviewed). This "data management review", for those students who require it, is a component of the compulsory annual review and the MEB director of postgraduate studies not approve the annual review until the data management review is successfully completed. The data management review is an opportunity for students to get advice from experts on how to organize their data more effectively with a view towards making the final archiving as simple as possible.

For questions or feedback on these guidelines, please contact the Data management council. See MEB intraweb for contact information.

5. Questions & Answers

Working with large datasets, such as SAS/Oracle dataviews:

If you are working with SAS/Oracle dataviews or similiar, rather than SAS datasets, then it is not possible/reasonable to save and archive each dataview according to the guidelines. Instead, keep the SAS program files that use the dataviews. Document the program files (by commenting, and describe them in the logbook and analysis plan) and describe which database was used and which views were used as data sources. Also, add dates on when the views were accessed for analyses (time-stamp) if possible. If you generate a permanent analysis dataset from the dataviews then store and archive that file according to the guidelines. If it is too large to be saved permanently or archived, then document the program that created it and save only programs and documentation. Give a reference in your documentation to the original location of the files.

GWAS or other genetic datasets:

If you are working with GWAS or sequencing data, follow the documentation guideline in Appendix 3 for how to store and document this type of data. We recommend using the analysis plan, logbook and suggested folder structure for this type of projects as well. If you want to use your own folder structure, then you may do so. But then describe it in a readme file (stored in the top folder) so others can understand it.

There is too much overlap between analysis plan, logbook and codebooks:

The point of documentation is not to create extra work or double work. The point of documentation is to save time – for you and for staff in future! Use these guidelines efficiently, that is, do not do double work. It is for example possible to have the logbook as a separate chapter within the analysis plan (instead of as a separate file), if this makes the work less and the analysis plan is still readable. The important point is to make your files and results traceable.

If you have complicated data merges then a separate logbook is recommended. The guidelines provide suggestions, but you have a freedom to implement them in a way that suits your project. If you choose to create your own structure of documentation, you must describe how it works (for example in a readme file).

Only create codebooks for datasets that you suspect may be used by others in future or datasets that will be stored permanently in your archived folders.

PhD Projects that are part of large projects on Project disks:

The guidelines are written as if you have a PhD folder where all your studies are stored together. Most Phd students have projects which are part of a large project, which may already have a Project folder on P:\ or other project disks. If your study is part of a larger project, then keep your folder on that project disk together with the rest of the large project.

Maintain the same documentation files as recommended in the guidelines.

For archiving, create an Archive subfolder in the larger Project folder. When your study is published move your folder to this Archive folder and archive it there temporarily.

If the project has other documentation routines than suggested in this guideline, then make a note about it in the readme file and describe where certain key project files can be found.

When should you archive and when should you contact IT to move files to archive:

You should clean up files and prepare your study folder for archiving when the paper is published. This work is easiest done when the work is fresh in your mind, that is, when it has just been published. (I.e. It is **not** easy to archive all four papers at the end of your thesis work!) If your study is part of a larger project on P:\, then you may at this point move the cleaned folder to the Archive subfolder of the Project.

You should contact IT-support when you leave MEB **if** all your study folders are ready to be archived. If not all of your papers are published yet, you should wait until they are or delegate to another MEB staff to archive your files after publication. It is best to archive all your folders together. If you leave MEB before all your papers are published, you need to inform the person you delegate the archiving to and your supervisor about the files and folders and what is left to do before it can be archived. If your folders are part of a larger Project on P:\ then you should contact IT-support to move them but also (if your supervisor or PI request it) save copies of the archived files in the Project folder. Inform your supervisor and the PI of the project about the location of your folders and files. If you have any questions about archiving and how to proceed, you can always ask for help and advice from the IT staff.

Can I create my own way of documenting:

Yes you can, as long as you follow KI rules and Swedish law.

The guidelines provide suggestions in accordance with KI rules and Swedish law, but you have a freedom to implement them in a way that suits your project. If you choose to create your own way of documentation, you must describe how it is set up (for example in a readme file). If you choose to use the documentation structure as described in the guidelines (i.e. using files such as analysis plan, logbook, codebooks etc.), then you simply refer to the guideline in your readme file: “This project follows the MEB Guidelines for documentation and archiving of research files.”

Must I document my submissions to journals:

Yes, each submission must be documented, even rejections. Make a note in the analysis plan about which files were submitted and when and to which journal. Also make notations about resubmissions and decisions from the journal.

You must keep files related to original submission, reviews and resubmissions. These files must be archived.

My data is not analysed or stored at MEB, but in another department or authority, must I follow MEB guidelines?

Yes, you must still adhere to MEB guidelines for publications which are published during your time as a PhD student at MEB. If you are working on data outside MEB then you are still responsible for making sure that published results can be reproduced. It is important to document where the original data and analysis files can be found (even if in another place), for example by providing a reference, location and contact persons at the other work place. This should be included in the readme-files in your MEB folder. We encourage students working in other places to use the proposed file and documentation structure in the MEB guidelines, if it is compatible with the routines of the other workplace.

APPENDIX 1. Example: Template for project readme.txt file

Readme.txt
Name:
Date:
Address:

This document provides explanations regarding the folders included in my Ar-chive for the period: start date – end date, when I was a PhD student at MEB.

Main tutor: Name:, Department:, Address:, E-mail:
Other tutors: Names

FOLDERS:

1. Archive_name1

This folder includes datasets, programs, and drafts for the case-control study on xxx cancer based on data from the xxx Register. PIs for this study are:
Name, Department, e-mail:
You can also contact: Name, Department, e-mail:

The files logbook.rtf (in former version: "Files.rtf") and code-book_filename.rtf (in former version: "Variables.rtf") describe all programs, original and created datasets, and the format of the variables.

The folder "Manuscript x" includes the drafts for the article xxx et al. in journal xxx year.

The folder "Region_codes" is explained in "xxx.doc"

The folder "Original_data" includes the original files delivered from Statistics Sweden.

2. Archive_name2

3. Archive_name3

4. Archive_name4
etc…

5. Thesis

This folder includes a copy (.pdf file) of my thesis

APPENDIX 2. Example: Template for SAS program

```
*----------      MEB SAS template     ------------------------------------;
* Filename:                                      ;
* Study....:                                     ;
* Author...:                                     ;
* Date.....:                                     ;
* Updated..: “date” by “whom”                                ;
* Purpose..:                                     ;
* Note.....:                                     ;
*---------------------------------------------------------------------------;
* Data used...:                                  ;
* Data created:                                   ;
*---------------------------------------------------------------------------;
*                                             ;
* OP.......: SAS9.1TS1M3                              ;
*---------------------------------------------------------------------------;

*-- External programs ------------------------------------------------------;

*-- SAS macros ------------------------------------------------------------;

*-- SAS formats -----------------------------------------------------------;

*-- Main program ----------------------------------------------------------;


*-- Cleanup ----------------------------------------------------------------;

title;footnote;
proc datasets lib=work mt=data nolist;
  delete _null_;
quit;

*-- End of File ------------------------------------------------------------;
```

APPENDIX 3. Documenting GWAS and biological (laboratory) data

Patrik Magnusson for DMC, MEB, KI, 20110810

General principles for documentation of biological (laboratory) data at MEB, v.1.0

To be kept in mind: what should and can be documented varies a lot depending on type of laboratory analyses that has been undertaken. Genotyping has become more and more standardized over time and is expected to continue that way. Much larger variation exists when it comes to serum/cell measurements or bacterial/viral/parasitic, particularly non-DNA based analyses.

Location

It is recommended that the documentation text should be available in README.txt files located within the folder in which the laboratory data are stored as well as archived.

Recommended content of documentation text

1) EPN number EPN decision for the study (EPN=Etikprövningsnämnd, ethical approval board)

2) Short descriptions (or pointers to such descriptions) of:

a) *samples*
- subjects (ascertainment, age, sex, ethnicity)
- collection (what, when, how, by whom)
- preparation (e.g. extraction)
- storage (method, format, temperature)
- Amount of sample used for analysis

b) *laboratory analyses*
- laboratory (what, when)
- laboratory contact person
- what analysis (purpose/platform/version)

c) *Quality Control (QC) steps*
- removal of subjects (why, how many)
- removal of measurement variables (why, how many)
- QC steps performed when and by whom
- naming of pre and post QC files

d) *Variables included in the lab-data file(s)*

3) Where to find archived version of source (raw, pre-QC) as well as post-QC lab-data.