Instructions for submitting research data consisting of statistical data and documents

Instructions on how to submit research data, which contains data from statistical files and documents in the form of text, audio, video or images

The Danish National Archives, July 2020

Version 1.0

Content

0. Reading instructions

Public authorities are required to hand over data and documents worthy of preservation to a public archive.

This guide describes how research data is submitted in the case of a study that contains both data from statistical files and documents in the form of text, audio, video or images. It can be a research project that has collected data via a questionnaire survey and at the same time conducted qualitative interviews with selected respondents.

The archive decides whether data from a research project can be submitted together in accordance with this guide or whether data must be submitted in two separate submissions.

A. The guide's target group and application

This guide is intended for those who submit research data consisting of both statistical files and documents.

B. Reference to other instructions

In addition to this guide, the Danish National Archives has prepared other guides that are important for the production and submission of a information submission package:

- Sample information submission package with statistics data FD.18005
- Quick guide on the creation and testing of an information submission package with ASTA
- ASTA user guidance
- Instructions for creating an information submission package with data from spreadsheets or CSV files
- Instructions for executive order on archival versions no. 128 (see appendix 9 on information submission packages)
- Guide to Create archiveIndex
- Guide to Create contextDocumentationIndex
- Guidance on converting documents to TIFF
- Guidance on UTF-8

The following instructions are important for submitting the documents:

- Spreadsheets for information about the digital documents (Template)

All guidance material can be accessed from the Danish National Archives' website www.sa.dk.

C. Legislation and legal precepts

Information about legislation etc. can be found on the Danish National Archives' website www.sa.dk.

D. Definitions

(1) Submission provision

Before the submission of the archival version is started, the receiving archive prepares a provision which determines the content of the submission. The submission provision is a requirements specification for the ***content*** of the submission, while the executive order on the archival version determines the ***format*** for the submission.

Information submission packages with data from statistical files generally consist of context documents to be submitted in the Danish National Archives' archive formats, extraction of data and metadata from the statistics files to be submitted, as well as two index files in xml format that contain overall metadata about the submitted data and context documents.

Archive formats The Danish National Archives uses 6 archive formats: TIFF, JPEG2000, MP3, WAV, MPEG2 and MPEG4.

Documents in the submission are data, e.g. journals, qualitative interviews, transcribed interviews, video recordings for field studies, etc.

Context documents are documents that describe the submitted data/documents.

1. Submission format

A research project consisting of both statistical files and documents must be submitted as follows:

1. Statistical files are submitted as a information submission package, cf. the provisions in Appendix 9 in Executive Order no. 128 on archival versions
2. Digital documents are converted to archive format and submitted as a copy of a folder structure containing the documents
3. Records of information about all digital documents located in the folder structure are recorded in a spreadsheet that complies with this guide's metadata standard for document collections

The archive converts the information submission package, the digital documents and the document records in the spreadsheet into a single archival version, cf. the provisions in Executive Order no. 128 on archival versions, Appendices 1-8. The submission format is specified in more detail below.

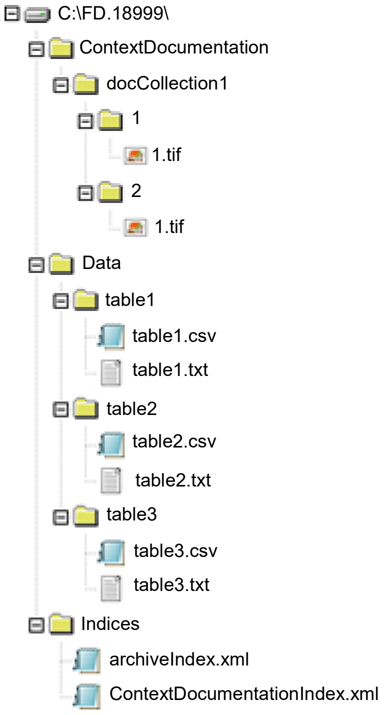
A. Statistics files are submitted as an information submission package

This guide deals with data that was originally created and processed in statistical programs such as SAS, Stata, SPSS or R and stored in one of these statistical file formats or collected data recorded in a spreadsheet and used for statistical analysis in e.g. Excel.

The Danish National Archives has laid down a number of provisions for the submission of statistical files in the form of a information submission package for the sake of preservation and future use of data, which all authorities must comply with. These provisions are described in the Danish National Archives' Executive Order no. 128 on archival versions, Appendix 9.

The structure and content of the information submission package are shown in Figure 1 and can be created with the ASTA program. The index files can be created with the Danish National Archives' input programs *Create archiveIndex.exe* and *Create contextDocumentationIndex.exe*. Both the executive order, guidelines, and the programs can be downloaded from the Danish National Archives' website www.sa.dk.

Figure 1 Graphic overview of elements and structure in a information submission package



The **ContextDocumentation** folder contains documents converted into their preservation format, e.g. TIFF. It can be a methods report, a questionnaire or other documents that describe the data contained in the information submission package.

The **Data** folder contains both a data file and a metadata file, both of which comply with the requirements of Appendix 9. Data extracted from the original statistics files must be submitted in a semicolon-separated csv file (**table1.csv**). Metadata extracted from the statistics file, e.g. variable descriptions (variable labels), response categories (value labels), and codes for missing values (missing values) must be submitted as a metadata file in txt format (**table1.txt**)

The **Indices** folder contains two index files with metadata at a more general level. The **archiveIndex.xml** file contains e.g. information such as the name of the data set that is submitted, the name of the researcher who collected the data, the period the data covers, access restrictions to data, etc. The file **contextDocumentationIndex.xml** contains information about the context documents located in the

ContextDocumentation folder, e.g. the document’s title, author and subject categorisation of the document.

B. Digital documents are converted to archive format and submitted as a folder structure

A copy of a folder structure containing the digital documents in archive format is delivered together with the information submission package. The folder structure is placed on the submission medium together with the information submission package with statistical data in a folder named lbnr_Dokumenter, where lbnr is the serial number for the submission issued by the archive.

The Danish National Archives distinguishes between context documents and digital documents. Context documents are considered as metadata that describe the submitted data, e.g. a project description, a methods report, a questionnaire, or variable descriptions. The context documents are placed in the information submission package, cf. item 1. Other documents such as qualitative interviews, audio recordings of interviews, or video recordings for field studies, on the other hand, are considered data. These, like the context documents, must be converted to one of the following six archive formats before submission to the archive:

1. TIFF (.tif) for regular office documents in the form of text, spreadsheets and images
2. JPEG2000 (.jp2) for large drawings, maps and photos
3. MP3 (.mp3) for audio files
4. WAV (.wav) for audio files where high quality is important
5. MPEG2 (.mpg) for video files
6. MPEG4 (.mpg) for video files

The digital documents must comply with the specifications set out in Executive Order no. 128 on archival versions, 5 E-F and H-J.

The digital documents can be converted and placed within the original folder structure, placed in a single folder, or placed in a newly created folder structure that possibly arranges and divides the documents into subjects.

Instructions for converting documents, e.g. “Instructions for conversion to TIFF format” with PDFCreator can be downloaded from the Danish National Archives’ website www.sa.dk.

C. Information about digital documents is recorded in a spreadsheet

Information about the digital documents must be recorded in a spreadsheet A fixed template for this spreadsheet must be used for the registrations and can be downloaded from the Danish National Archives’ website www.sa.dk

The spreadsheet contains columns for recording the 15 pieces of information from the Dublin Core metadata standard. The information SourcePath has been added to this. The metadata standard used in the template as well as descriptions of how the information is filled in are shown in Figure 2.

When filling in the spreadsheet, please note:

- A row in the spreadsheet is filled out per document in the folder structure.
- Columns must be retained in the spreadsheet in the order specified even if they are not filled out.
- Column names appear in row 1 of the spreadsheet.
- The column descriptions from the metadata standard in Figure 2 appear from row 2 in the spreadsheet.
- Adding own or edited column descriptions in row 3 of the spreadsheet is allowed so that they more

accurately describe the contents of the column.

- The information ***SourcePath*** and ***Title*** are required to be filled in for all documents registered in the spreadsheet.
- ***Date*** for the document must be filled in if it exists and must comply with the format YYYY-MM-DD, e.g. 2019-08-12.
- The *information* ***Relationship*** must be filled in if there is a relationship between the registered document and a row in a statistics file which is also found in the information submission package (from item 1 A). If this information is entered (e.g. with a civil registration number or an ID), the column description for the column must be adapted (in row 3 in the spreadsheet) so that it appears from this description which variable in which statistics file in the information submission package this reference is linked to.
- Other information to fill out in the spreadsheet is optional. The recorded information will be used by future users of the material to search for a relevant document.
- Adding additional information about the documents in additional columns inserted after the last column of the spreadsheet is allowed. Column title in row 1 as well as column description in row 3 are mandatory to fill in for the additional columns added.

Figure 2. The guide's metadata standard for document collections
*Information marked with * is mandatory to fill in*

Column name	Column description
SourcePath*	Specification of the full path to the document in the folder structure incl. file name and extension, e.g. S:\ Skilsmisseprojekt\Børn og skilsmisse\Interview A – barn.tif
Title*	(***Title***) Indication of the original title or title subsequently registered on the document. I.e. the title of the image, movie, video, audio recording, or text document.
Creator	(***Creator***) Indication of the name of the person who took the initiative for or created the content of the document. For example, name of author, interviewer, researcher, photographer, test subject or other persons who have produced material that is included as empirical data in the research.
Subject	(***Subject***) Indication of the subject or category associated with the document. For example, keywords that describe the contents of the document, the name of the folder where the document is located, or other categorisations that group the documents.
Description	(***Content description***) Free text description of the document's content. For example, what the document contains, what the film is about, or what the picture shows.
Publisher	(***Publisher***) Indication of the publisher of the document. For example, film company, producer, publisher. Rarely relevant to fill in for research data.
Contributor	(***Contributor***) Indication of the names of persons or other entities that have contributed to the creation of the content of the document. For example, actors in a video, people talking in an audio recording, interviewees or respondents.

Date	(***Production date***) Indication of when the document was created. For example, the creation date of a document, the date when a video was recorded, or the date when an interview was performed. The specified date must comply with the format YYYY-MM-DD.
Type	(***Object type***) Indication of the document type selected from the following fixed list: Audio, Video, Text, Picture.
Format	(***Original format***) Indication of the original format of the document. Here, for example, original media can be specified, such as drawing, photography, VHS tape or cassette. Here you can also specify the digital format the document had before conversion to preservation format, e.g. mov, wma, spreadsheet, word, pdf, docx, or jpeg.
Identifier	(***Identification***) Indication of an original ID for the document used in a given context.
Source	(***Original source***) Indicates a reference to another material from which the document is derived/digitised from. For example, indication of bar code, name, ID or other identification of the material referred to.
Language	(***Language***) Indicates the language used in the document.
Relation	(***Relationship***) Indicates information that constitutes a reference from this document to a row in a statistics file which can be found in the information submission package. Here you can, for example, enter a CPR number or an ID which is also found in the statistics file and which constitutes this reference.
Coverage	(***Size***) Specification of the dimensions of the document. For example, the size of an image where height times width are specified in centimetres (cm) or metres (m), a document size specified in Kilobytes (KB), Megabytes (MB), or Gigabytes (GB) or a video recording play time specified in minutes.
Rights	(***Rights***) Indication of copyright owner, i.e. who it is that grants permission for the use of the document.