Databases and spreadsheets: A guide to good practice

22nd February 2023

The research leading to these results has received funding from the European Community's Horizon 2020 Programme (H2020-INFRAIA-2018-1) under grant agreement n° 823914.

For any questions regarding this Guide, please contact: contact@ariadne-infrastructure.eu

(PDF/A-1 (ISO 19005-1), created in MS Word (Version 2016) from the docx format.

Contents

1. Introduction to databases and spreadsheets 4
1.1 What are databases and spreadsheets? 5
Spreadsheets 6
Databases 6
2. Things to consider when creating databases and spreadsheets 8
2.1 General considerations 8
Avoid embedded material 8
Check data consistency and documentation 8
Formatting data and cells 8
Date 9
Boolean data type 9
Decimals 9
Currency 9
3. Archiving databases and spreadsheets 10
3.1 Deciding which files to archive 10
3.2 Deciding how to archive 10
Significant properties 10
General checks 10
Naming files 11
File formats for archiving 12
Checks after migration 13
Other formats 13
3.3 Metadata and Documentation 14
3.4 Structuring data 15
4 File format 16
Common file formats for databases 16
Common spreadsheet file formats 22

Databases and spreadsheets: A guide to good practice

Acknowledgements: Based on Archaeology Data Service's Guides to Good Practice on Databases and Spreadsheets (https://archaeologydataservice.ac.uk/help-guidance/guides-to-good-practice/). The information has been reworked, updated, and supplemented by the Swedish National Data Service (https://snd.gu.se/en) to be more appropriate to data of non-archaeological origin.

1. Introduction to databases and spreadsheets

Although databases and spreadsheets have different functions, in many cases these different types of application are used in a similar way to collect and store data (i.e., in rows and columns). From an archival perspective, the similarities become more apparent when the main properties of the formats are considered. For both databases and spreadsheets, these properties are the data values entered and the structure (tables and sheets) in which the data is entered. From this perspective, both databases and spreadsheets can be treated (and archived) in a similar way.

This guide aims to provide an overview of what to consider when preserving the most common features of databases and spreadsheets. The guide highlights their similarities and how they can be handled in a similar way, but also their differences and which additional elements, functions and processes need to be documented. The guide does not provide detailed information on the design of databases and spreadsheets beyond that which affects their preservation.

1.1 What are databases and spreadsheets?

The definition of what a database is varies depending on who you talk to. A database consists of a collection of information that is related and organized in such a way that it is easy to search for and retrieve individual pieces of information but also to be able to modify the information. For those working with database technology, a database should also have a schema (an explicit description of the type of data) and be consistent or logically coherent (not contain contradictions). The word *database* can refer to both the information stored and the software (database management system) that will be used to interpret the stored data structure. Many database management systems are very complex and may consist of entire systems with different applications. Examples of database management systems are Microsoft Access, MySQL, Microsoft SQL Server, Oracle, and PostgreSQL. One of the advantages of this type of database is that the user decides on the structure and function. Without going into detail, users can change the logical structure (for example increase the number of columns) without having to rewrite the application. You can also change the physical storage structure without having to rewrite the applications you use. Some other advantages of using this type of database management system are that several users can work on the same database simultaneously; information entered is stored directly in the database and not in the computer's primary memory (which may be lost if the power fails); different interfaces can be created depending on user needs; different users can be given different permissions to protect against unauthorized access.

In this document we mainly deal with the types of databases that have predefined software in which users enter the data they want to be there and then use the functions that are predefined. These functions may be used either via command lines or via menus. The data is then stored in files that may be exported and imported as needed. These programmes range from the simplest type of spreadsheet (Excel) to computer programmes for statistical analysis (SPSS, STATA, etc.).

Although at the most basic level there are similarities between databases and spreadsheets, in that they contain tabulated data with values organized in columns and rows, there is a big difference in terms of function. Spreadsheet applications, originating from paper accounting spreadsheets, are intended to be used for mathematical data (for example financial statements) with fast calculations

and processing. Database applications are designed to store large volumes of data of various types with advanced search/analysis[1] and reporting functions for this data.

Spreadsheets

Spreadsheets are the simplest form of database and usually consist of one or more sheets of tabulated data. In addition to data, spreadsheets may contain formulae, images, charts, and tables. For example, additional values (such as the total of a column) can be created using various formulae. Charts and tables can be generated using data and the chart/table can then be placed in an existing sheet or exist individually as a new sheet. The formatting of cells or their values may also be an important element of a spreadsheet and can be used to convey information or to highlight parts of information. Data input and the use of a spreadsheet can be restricted to some extent by locking cells and cell-specific calculations (for example by rounding values or by specifying values in specific formats such as currency or time).

Spreadsheet applications, like many word processing applications, have increasingly supported XML-based file formats in recent years. Applications such as Microsoft Office, OpenOffice and WordPerfect Office now support both the Office Open XML (OOXML) and OpenDocument Format (ODF) file formats.

Databases

Unlike spreadsheets, most database applications allow, and often require, specification of the field length (number of characters) and the type of data to be recorded (numerical, etc.). In contrast to spreadsheets, which have largely the same fundamental design and management of data, databases may be divided into several different types based on their architecture.

A *hierarchical database* is an older type of database model in which data is stored in a tree structure, i.e., a parent-child relationship, in which each record consists of a 'parent' which may then have one or more 'children' with similar structures. The system is very fast at producing individual records and it is easy to add/remove information, but the model is memory intensive, slow at free data search and relationally rigid.

Rectangular databases are like spreadsheets in that tabulated data are organized in horizontal rows containing data about the object under investigation, and vertical columns representing a particular type, value, or attribute to be recorded for the object. In rectangular databases, there may be a freer definition of what data is and how data are recorded in the system. At the same time, there may be duplicated information in the different values.

Relational databases resolve these and similar problems by requiring a data structure that is predefined by grouping data with similar attributes in separate tables, which are then linked together by certain key fields. The combination of one or more key fields can generate a key. There are

[1] Whether they are called search functions or analysis functions varies depending on the software. The level of sophistication of these functions also varies between applications.

different types of keys, such as primary[2] and foreign[3] keys. Unlike spreadsheets and many rectangular databases, most database applications require a strict definition of the field length and the type of data (numerical, etc.) to be recorded.

Object-oriented databases are created to store complex objects such as multimedia files and CAD objects. Data are not normally stored in the database but as attributes in objects. When an application makes a call to an object database, the search is not always made directly in the database. The application asks the object to perform a certain routine and return a result. Common data are also stored in the object database to facilitate searches.

Like tables generated from spreadsheets, databases may consist of more than just data values and metadata. Forms, which are used to enter data or perform calculations, are often the only way users interact with the database and may therefore be seen as part of it but separate from the actual data. Similarly, queries and results or reports resulting from the use of the database may be seen as 'non-data' components of the database. It may therefore be important to store some components of these materials with the database.

[2] A primary key is the smallest set of columns in a database that you need to know the value of in order to find a unique record (tuple) in a database. The columns included in the key must be defined. For example, the 'identity' key may consist of information from the columns Name, Year of birth, Month, and Day. It may also, as in Sweden, consist of a value: the personal identity number. This is because it is unique to each individual.

[3] If there is a foreign key (i.e., a reference to a primary key in another table/database), this should be indicated. For example, the identity key from the example in the note above may be used as a foreign key in another database, in which case the Name, Year of birth, Month, and Day data do not need to be entered again.

2. Things to consider when creating databases and spreadsheets

2.1 General considerations

Below are a number of points that should be considered to ensure that data are consistent and easily reusable, and that they remain so throughout the preservation process.

- Where possible, use controlled vocabularies and established keyword lists for data entry in both databases and spreadsheets.
- Be consistent and have meaningful names for tables/spreadsheets and rows/columns. Be aware that tables or spreadsheets may not be stored in a single file. Also keep in mind that some applications have restrictions on how you can name fields (for example in ORACLE you cannot use names for tables that start with a number or names for fields that start with 'desc' or 'date') or the use of spaces when naming a field or column. Although some applications accept this, it is better to avoid it as it may cause problems for future data migration.[4]
- When using spreadsheets, avoid using formatting and page layout to highlight the meaning of certain values as this may be lost when data are migrated or when using different applications.

Avoid embedded material

Many database and spreadsheet applications allow the user to embed other types of files in the file (for example images). While databases usually include links to external files rather than the files themselves, spreadsheet applications such as Microsoft Excel and various versions of OpenOffice Calc allow users to embed tables and charts created from data along with other images. It is recommended that such content is stored and archived as separate files in the same folder as the database/spreadsheet. This ensures that the quality of the files is not lost. It is important to have an archival strategy for the format of the database/spreadsheet so that it can be recreated in the future.

Check data consistency and documentation

Coded or inconsistently entered data pose problems for the reusability in both databases and spreadsheets. Coded fields and data must be adequately documented, and the documentation must be archived with the database or spreadsheet so that the meaning of the coding is not lost. Inconsistently entered data (which can be more easily checked in a database than in a spreadsheet) may result in the meaning of data being lost (for example is 'A' the same as 'a'?) and cause problems when querying the database.

Formatting data and cells

Databases and spreadsheets may contain more than tabulated data, charts, and images. Spreadsheet applications, in particular, allow formatting (for example different colours of text and cells) of data and the cells the data are in. This type of formatting is usually used to highlight data or simplify the reading of data (for example column totals, negative values, etc.) but also to highlight information. This type of formatting may be lost on transfer to other formats, especially to plain text.

[4] Migration may involve transfer between different media but also transfer between different file formats.

Date

There are different ways of writing the date, the order in which the day/month/year are written and whether the year is written with two or four digits. The recommendation is to follow the ISO standard, i.e., yyyy-mm-dd for date and hh:mm:ss for time. See also ISO 8601.[5]

Boolean data type

A type of data in which the value represents something that is either true or false, yes/no, highlighted/not highlighted, etc. The easiest way to export these values is to convert them into simple text values like Y or N, or numerical values like 1 or 0.

Decimals

Some applications automatically round numbers to two decimal places, which may result in the loss of important data. For example, a coordinate written as x=123.456 can be rounded to 123.46. It is possible to reset the number of decimal places so that rounding does not occur. The problem with this is that some applications still do not export the values as they are. For example, previous versions of the Microsoft Access export wizard saw all numerical values as decimals with two decimal places, and thus automatically rounded all numbers to two decimal places. This also applies to later versions if you have formatted the cells to contain numerical values and specify that they should have two decimal places, for example. This problem can be overcome by manually making it clear that decimals should not be rounded and should be exported in their original form.[6] For further instruction on how to handle decimals in certain file formats, see the respective formats in section 4.

Currency

When importing data with currency symbols, the symbol may, when the date file is opened, be modified to the regional currency where you are located instead of the original currency symbol. This causes dollar signs to be converted to euro signs, etc. Therefore, when exporting, it is better to convert currency to a value without a currency symbol and then change it back to currency when importing. However, you should document this in your metadata documentation.[7]

[5] https://www.iso.org/iso-8601-date-and-time-format.html. Accessed 5 December 2022
[6] https://ppp.cessda.eu/doc/D10.4_Data_Formats.pdf p.37. Accessed 30 June 2022
[7] https://ppp.cessda.eu/doc/D10.4_Data_Formats.pdf p.37. Accessed 30 June 2022

3. Archiving databases and spreadsheets

3.1 Deciding which files to archive

Like text documents, databases and spreadsheets normally remain in the same format throughout the creation process. They are also largely self-contained as they usually do not include imported images or other types of media within the file. However, when this is the case, it is recommended that the embedded content is stored separately so that its archiving is optimized.

3.2 Deciding how to archive

The core of databases and spreadsheets is the data tables/sheets themselves, along with documentation and metadata describing the contents of and relationships between tables and sheets. The order and/or layout of rows and columns may also be important (see below). Forms, reports, queries, and macros may also be worth preserving.

Significant properties

The basic elements of a file that should be preserved and maintained are:

- **Values** – The actual data within spreadsheets or databases, including cell headings and the values in the cells. There may be several sheets of data or tables.
- **Graphics** – Figures, charts, and tables in spreadsheets. Databases rarely allow embedded material, but it is possible to link to external files. However, it is important to be aware that this type of functionality is emerging with new types of databases (for example in Microsoft Access .accdb files).
- **Layout** – For spreadsheets in particular, where formatting and colours and certain styles are used to highlight the meaning of parts of the data, it will be important to preserve this additional information in some way. The use of layout functions is common when spreadsheets are used to lay out tabular data. If spreadsheets use features such as formatting, colours and styles, alternative formats, such as PDF/A, should also be used in parallel to preserve the appearance of the data, as formatting is usually lost when data are exported to text-based formats.
- **Relationships** – For databases, but also for spreadsheets, it is important that relationships between tables/sheets are well documented and understandable.

General checks

Apart from ensuring that significant properties of a file are preserved when converting it from one format to another, there are a number of checks that need to be carried out before the conversion takes place. These checks ensure that significant characteristics of a spreadsheet or database are maintained and not lost during the conversion process.

- **Layout and formatting** – As mentioned earlier, especially in spreadsheets, the use of formatting, colours and styles can cause problems when migrating data to comma or tab-separated text files, for example. This is if the features have been used to give additional meaning to the data. Checks should be made for headings that span several rows or columns, and for information that is highlighted by the use of colour, borders, or different fonts. Depending on the type of formatting, data must be edited by hand before the migration to ensure that information is not lost (for example, for spreadsheets, merged cells must be split and the text duplicated for each new cell), or an alternative migration format must be found.

- **Tables and sheets** – Although it should be assumed that databases and spreadsheets are to be preserved in their entirety, each individual file should still be reviewed to assess which tables or sheets should be preserved and migrated. Spreadsheets contain a default number of empty sheets and users often create additional sheets, or tables in a database, to temporarily store data that are not intended for preservation.

- **Formulae, queries, macros** – If a file contains complex formulae or queries that must be preserved as standalone results, they must be defined as migrated versions of the database or spreadsheet. What should then be preserved in the migrated files are the results themselves, while formulae and queries should be preserved separately in a text file so that the spreadsheet functionality can be recreated at a later date.

- **Comments and notes** – When converting from one format to another, comments and notes contained in a file may not be included. This means that they must be identified and saved in a separate file with clear information about which file and which cell the comment/note belongs to.

- **Hidden or protected data** – Sometimes spreadsheets contain hidden or protected cells. These cells must be identified so that information about them and what they contain is not lost.

- **Special characters or delimiters** – Sometimes a database or spreadsheet may contain special characters or common delimiters within the dataset. Delimiters such as pipes, commas or tabs can cause problems in a file when it is to be migrated to text documents with delimiters. Consequently, these types of character need to be identified early so that a strategy can be developed to ensure that they are not lost in a conversion. Apart from various delimiters, special characters, and non-Latin characters such as ampersands (&), em dashes ('—') and inverted commas (" ") may affect the conversion of data and thus the subsequent appearance of the data. Databases and spreadsheets may also contain foreign characters which cannot be exported to a text file unless a specific character set (for example UTF-8) is specified.

- **Links** – For databases, it is important that the relationship between different tables is well documented (see also 3.3) and correct. Checks must be made to ensure that duplicate or orphan records are not present. Worksheets in spreadsheets can be linked together by values in cells being taken from other worksheets. Both databases and spreadsheets may contain links to, or names of, files stored separately outside the database/spreadsheet. If these external files are part of the project, it is necessary to ensure that file names and files are stored correctly.

Naming files

Depending on how the data will be stored, it may be necessary to rename tables in databases and worksheets in a spreadsheet. If files are to be stored as text documents with delimiters, each table and worksheet will result in a separate text file. This is true for both databases and spreadsheets. As far as possible, you should try to keep the original name of the files (the file extension may vary). If multiple worksheets or tables are converted, the new files' names should refer to the original file and also to the individual worksheet(s)/table(s), for example:

- *[databasename]_[tablename].txt,* for example findsdatabase_stoneage.txt
- *[spreadsheetname]_[worksheetname].txt*, for example locationregister_photos.txt

When additional files are created to contain images, queries, notes, or formulae, they should also be named in a logical way so that it is easy to link them to the original file they are associated with, for example:
[spreadsheetname]-[worksheetname]-[chartname].tif.

In some cases, it may be necessary to rename tables by either shortening them or by removing characters that cannot be in a file name. In such cases, the aim is to stay as close as possible to the original name.

File formats for archiving
For the majority of databases and spreadsheets, the recommended format for archiving is delimited text (tab, csv, etc.). However, there are often stylistic or functional elements in databases and spreadsheets that can only be preserved in certain formats. In such cases, it is recommended that an open XML-based format such as .ods or .xlsx is used or that those parts of the file are documented and saved with a text-based export of the data.

Format	Description
Delimited text file (for example .tab, .csv)	Delimited text files are generally the recommended format for archiving databases and spreadsheets, and many applications can export to this format. The contents of the cells are separated by a delimiter (for example a comma) and possibly by quotation marks around the value, and each row is indicated by a line break. Common delimiters are commas (CSV files), tabs and pipes (|). These formats are the most common and have the advantage that they can be opened directly by the most common applications such as MS Excel. This allows the format to be used for both dissemination and archiving. As noted above, the format only stores data, while other information (images, formulae, etc.) should be stored separately. There is no specification for CSV and it is easy to find differences between different CSV formats. The software, system and regional settings affect the appearance of the CSV export of a database or spreadsheet. Export wizards and well-documented definitions of the csv encoding are necessary to ensure that all stored data are converted in the same way. To avoid problems with diacritical[8] characters, such as punctuation, export with Unicode is recommended.[9]
.ods	Works as an archive format for spreadsheets, but embedded charts and tables should be stored separately.
.xlsx	Works as an archive format for spreadsheets, but embedded charts and tables should be stored separately.

[8] Diacritics are small characters added to a letter (above, below or on top of it) that have a distinguishing function (i.e., the pronunciation of the letter is changed).
[9] https://ppp.cessda.eu/doc/D10.4_Data_Formats.pdf p. 36. Accessed 21 February 2023

Checks after migration

After migrating files to other formats, it is important to perform a number of checks to ensure that data have not been lost or corrupted in the process. These checks include:

- Check the number of rows after conversion
- Check the length of the text fields to ensure that the information has not been truncated on account of a character limit
- Check that all worksheets and tables have been exported
- Check that special characters have not been lost.

Other formats

PDF (preferably PDF/A) may be used to disseminate spreadsheet data in some cases. This should only be done when a large volume of information is available via formatting and layout which cannot be acceptably recreated via a CSV file, and which is lost when converted to an XML-based format (ODS or XLSX). In some cases, it may be necessary to attach a PDF along with a CSV file to show the user what the spreadsheet originally looked like and to be able to work with the material at the same time.

XML is a common format for spreadsheets. It can be opened in standard spreadsheet and applications that read text. XML also offers a potentially reliable format for storing database data.

To describe how an XML document is structured, one or more XML schemas (or formerly DTD, Document Type Definition) may be used to clarify which elements and attributes are allowed or required in an XML application. Both forms permit automatic validation of the content of an XML document, but because of the limitations of the DTD syntax, validation cannot be as accurate as using the more modern techniques. Both may contain comments with instructions intended for human reading in which the expected area of use of an element or an attribute may be specified. However, these instructions cannot be checked by machine. Another difference between DTD and XML schemas is that in the latter it is possible to specify the *namespace*[10] and data type.

SIARD[11] (Software Independent Archiving of Relational Databases) is a tool for simplifying the archiving of relational databases. SIARD is written using standards such as XML, SQL:2008 and Unicode and supports archiving of databases based on Oracle 10 or later, MySQL 5.5 or later, SQL Server 2012 or later, PostgreSQL 11 or later, Microsoft Access 2007 or later and DB2 or later. The Swiss Federal Archives develops and makes SIARD available free of charge, but a licence agreement is required.

An alternative to SIARD is the DBPTK (Database Preservation Toolkit)[12] developed by the European collaborative project E-ARK. This software allows you to add descriptive metadata and to search and navigate the content.

[10] *Namespace*: enables the use of predefined elements and attributes in XML documents. As there are different XML schemas, ambiguities between the definitions of elements and attributes of the different schemas may arise and need to be clarified. This is done by specifying the XML schema used for each element/attribute.
[11] https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html. Accessed 30 June 2022
[12] https://database-preservation.com. Accessed 30 June 2022

3.3 Metadata and Documentation

Both databases and spreadsheets need metadata and documentation on different levels to ensure that data can be preserved and reused. The following elements should be documented and stored in each dataset.

Element	Description
Project title	
Name of database/spreadsheet	

The following information should be provided for each worksheet/table in the spreadsheet/database:

Element	Description
Name of worksheet/table	
Purpose of spreadsheet/table	
Number of rows of data	
Primary key (databases)	Minimum set of columns in a database that you need to know the value of to find a unique record. The columns included in the key must be defined.
Foreign key (databases)	If there is a foreign key (primary key in another table/database), it must be indicated.

The following information should be provided for each column/field in the spreadsheet/database:

Element	Description
Name	Name of the field (database) or column (spreadsheet).
Description	Full description of fields and the codes or terminology used. Codes used in a dataset may be attached in a separate document.
Type of data and length of the field (for databases)	

For databases in particular, it is also necessary to describe the relationships between different tables, either in words or by creating and attaching a relationship diagram.

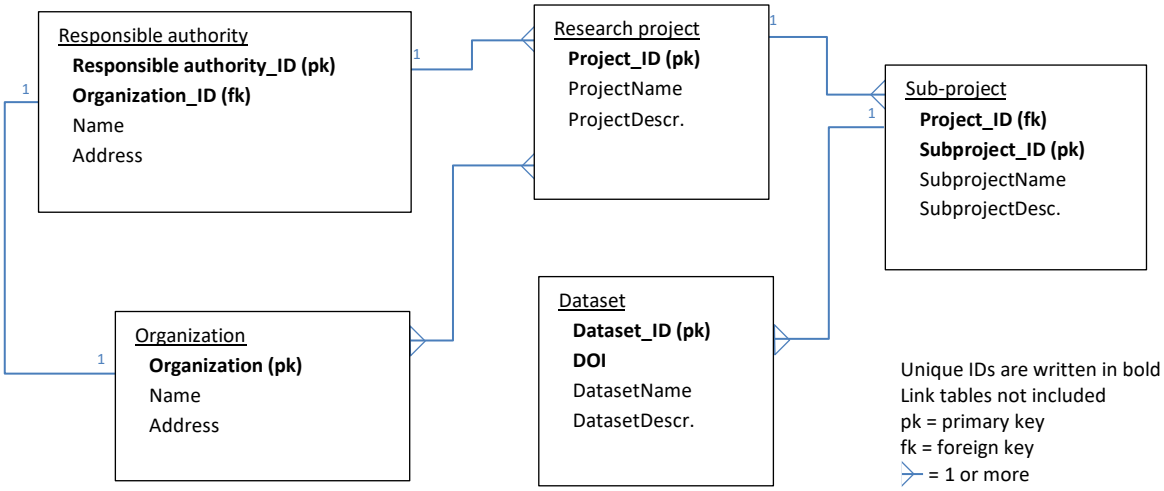

Fig. 1: Example of a relationship diagram.

The documentation should also include information on anything else that should be preserved, such as formulae, queries, macros, and comments. This information may be stored separately, for example in a text file.

3.4 Structuring data

Although the relationships between files should be clear from the filenames, it is an advantage to store related files in a folder. Exported images/charts, etc. may then be stored in a subfolder within the structure.

4 File format

The tables below describe some common file formats used in the creation of databases and spreadsheets. Some of these formats may appear obsolete, but they are described here because it may be necessary to handle older file formats. The associated software for file formats and how, or if, they may be used for archiving is also covered.

Common file formats for databases

dBASE	
File format/ extension	DBF/.dbf (also .dbt and .ndx)
Format	Originally used by dBASE but also used by others (for example ESRI and their ArcInfo). Old proprietary[13] format. Only the latest version (7) is documented (Feb 2011).[14]
Description	The format was developed for dBase but has since evolved into a more general format with a number of (unfortunately incompatible) variants which are often referred to as xBase. Even if the structure is quite simple and well documented[15], it is important to document the software used to create the file.
Recommendations	The format may be used for archiving but on account of incompatible variants (see above) it is important to know exactly how the file was created. dBase, like most applications, supports the export of more suitable formats.

Filemaker Pro	
File format/ extension	FP7/.fp7 and earlier versions (MF5/.fp5, FM3/.fp3, FM/.fm)
Format	Filemaker Pro 11 and earlier versions. Proprietary file format.
Description	Files created in FileMaker Pro (.fp7) 11, 10, 9, 8, 7 can be opened directly in and converted with Filemaker Pro 12 and later versions (.fmp12). Earlier versions than this require multi-step conversions.[16]

[13] Proprietary file formats are file formats that have restrictions (usually set by the owner) on how they can be used, modified, or copied.
[14] http://www.dbase.com/Knowledgebase/INT/db7_file_fmt.htm
[15] http://www.clicketyclick.dk/databases/xbase/format/index.html . Accessed 30 June 2022
[16] https://support.claris.com/s/article/Converting-older-FileMaker-Pro-files-to-the-fmp12-file-format-1503693002275?language=en_US . Accessed 30 June 2022

	Files created with Filemaker Pro 6, 5, 4 or 3 must first be converted to .fp7 format before they can be further converted to .fmp12 format.[17] If the files were created with Filemaker Pro 2.x or Filemaker Pro 1, the files must first be converted to Filemaker Pro 6 before being converted to .fp7 format and then converted to .fmp12 format.[18]
Recommendations	The file format .fp7 and earlier formats are not recommended for archiving but Filemaker Pro supports the generation of a number of more suitable formats.[19]
File format/ extension	FMP12/.fmp12
Format	Filemaker Pro 12 and later versions. Proprietary file format.
Description	FileMaker Inc., a subsidiary of Apple Inc., was converted into Claris FileMaker Inc. in 2019, and the software is now called Claris FileMaker. The software is available for multiple platforms and operating systems. Although the file format has been the same since Filemaker Pro 12 (.fmp12), features are added and lost,[20] so you should make sure your files are up to date and that you document the version of the software you used.
Recommendations	The file format .fp7 is not recommended for archiving but Filemaker Pro supports the generation of a number of more suitable formats.[21]

Microsoft Access	
File format/ extension	MDB/.mdb
Format	A proprietary Microsoft format used for Access databases (2003 and earlier variants).
Description	A proprietary Microsoft format now replaced by the .accdb format. Files saved in the .mdb file format are opened with the version of Access in which they were saved, i.e., with the relevant database engine (Jet Database Engine). There have been several updates of Jet (Access 2.0 uses version 2.5, Access 95

[17] Ibid.
[18] Ibid.
[19] https://support.claris.com/s/article/Exporting-data-from-FileMaker-Pro-1503692934804?language=en_US. Accessed 10 August 2022
[20] https://support.claris.com/s/answerview?language=en_US&anum=000034874. Accessed 10 August 2022
[21] https://support.claris.com/s/article/Exporting-data-from-FileMaker-Pro-1503692934804?language=en_US Accessed 10 August 2022

	uses version 3.0, Access 97 uses version 3.5, Access 2000 - 2003 uses version 4.0), which means that there have been several changes to the .mdb format. Problems were mainly noticed between Jet versions 2 and 3. It is possible that files created in Access, based on versions of Jet earlier than version 3, cannot be opened.
Recommendations	The .mdb file format is not recommended for archiving but Access supports the generation of a number of more suitable formats.
File format/ extension	ACCDB/.accdb
Format	A proprietary Microsoft format used for Access databases (2007 and later).
Description	The .accdb file format was introduced with ACCESS 2007 and was also used in ACCESS 2010. Although the format is the base format for Access 2007 and 2010, files created in Access 2010 are not fully compatible with Access 2007.[22]
Recommendations	The .accdb file format is not recommended for archiving but ACCESS supports the generation of a number of more suitable formats.[23]

OpenDocument Base	
File format/ extension	ODB/.odb
Format	An XML-based database format originating in OpenOffice.org. The standard was first defined with OpenDocument 1.2, although it was used before that in OpenOffice.org.
Description	An .odb file consists of a number of compressed folders of xml files. One of the files contains the actual data. Other files contain metadata documentation, information about the application used and a document template. An .odb file also consists of a machine interface, which can retrieve information from other data, as well as a database engine (in this case an HSQL database engine[24]).
Recommendations	The .odb file format is not recommended for archiving but several applications that use the format (for example OpenOffice/LiberOffice) support the generation of more suitable formats.

[22] https://support.microsoft.com/en-us/office/convert-a-database-to-the-accdb-file-format-098ddd31-5f84-4e89-8f44-db0cf7c11acd Accessed 10 August 2022
[23] https://support.sas.com/documentation/cdl/en/acpcref/63184/HTML/default/viewer.htm#a003102702.htm Accessed 10 August 2022
[24] http://hsqldb.org/ Accessed 10 August 2022

SAS	
File format/ extension	SAS/.sas
Format	SAS (Statistical Analysis System). A proprietary, cross-platform file format.[25] SAS is not a programme but a collection of computer programmes for data processing, statistical analysis, and reporting.
Description	A command-driven and menu-driven programme. SAS filenames may be up to 32 characters long. Spaces and non-alphanumeric characters (alphanumeric: A-Z and 0-9) are not allowed except for underscores (_). Filenames must start with a letter or underscore, followed by numbers and letters.[26]
Recommendations	The .sas file format is not recommended for archiving but SAS supports the generation of a number of more suitable formats.[27] These include export to ASCII where you simultaneously create a setup file which also includes a definition of variables. Make sure afterwards that the export worked properly.
File format/ extension	XPORT/.xpt
Format	SAS Transport
Description	When you create an SAS transport file (.xpt), missing data are not included and are deleted if you do not use SAS alpha missing codes. This in turn poses a problem because the difference between different types of missing data (such as 'natural wastage' versus 'refusal to answer') is not reflected when all missing data values are merged. Distinguishing between the different missing data values may be important in secondary analysis. When preparing to create an SAS transport file, the recommendation is therefore not to use SAS alpha missing codes but to create separate files. [28] Another problem is what is called SAS Proc Formats (value labels), which are not stored in an SAS transport file. SAS Proc Formats, on the other hand, may be saved in separate programme files or stored in SAS directory files, which are

[25] https://support.sas.com/documentation/cdl/en/movefile/59598/HTML/default/viewer.htm#a000986099.htm Accessed 10 August 2022
[26] https://support.sas.com/documentation/cdl/en/lrcon/62955/HTML/default/viewer.htm#a000998953.htm Accessed 10 August 2022
[27] https://support.sas.com/documentation/cdl/en/acpcref/63184/HTML/default/viewer.htm#a003102702.htm Please note that missing values are converted to empty fields when exporting to dBASE (DBF). Accessed 10 August 2022
[28] http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter6.html **Portable software-specific files. Accessed** 10 August 2022

	operating system specific. The best way to solve this is to attach user-defined 'SAS Proc format' and 'format' in separate files.[29]
Recommendations	Not intended for archiving but as a medium for transport between different data environments.[30] SAS support desk suggests conversion to ASCII (CSV).[31] CESSDA suggests conversion to plain text with Unicode encoding.[32]

SPSS	
File format/ extension	POR/.por
Format	SPSS-portable format. Proprietary file format.
Description	A portable format that can be read by other versions of SPSS regardless of the operating system. Variable names are limited to 8 bytes and, if necessary, are automatically converted to 8 bytes. You cannot save data in .por format with the Unicode setting. Data saved with IBM SPSS cannot be read by versions older than version 7.5. Data saved with Unicode encoding cannot be read by versions of IBM SPSS older than version 16.0. Opening data files with variable names longer than 8 bytes in version 10.x or 11.x creates unique versions of the variable names that are 8 bytes long. The original names are automatically saved for use in version 12.0 or later. In versions older than 10.0, the original variable name is lost. When using data with text variables longer than 255 bytes in versions prior to version 13.0, the text variable is broken down into multiple text variables that are a maximum of 255 bytes long.[33] SPSS maintains missing data values when portable files are generated. [34]

[29] *Guide to Social Science Data Preparation and Archiving*. **Portable software-specific files,** p. 48 (https://www.icpsr.umich.edu/web/pages/deposit/guide/). Accessed 11 August 2022
[30] https://ppp.cessda.eu/doc/D10.4_Data_Formats.pdf p. 34. Accessed 11 August 2022
[31] https://ppp.cessda.eu/doc/D10.4_Data_Formats.pdf p. 34. Accessed 11 August 2022
[32] https://ppp.cessda.eu/doc/D10.4_Data_Formats.pdf p. 34. Accessed 11 August 2022
[33] https://www.ibm.com/docs/en/spss-statistics/25.0.0?topic=formats-saving-data-data-file-types. SPSS Statistics (*.sav). IBM® SPSS® Statistics format. Accessed 11 August 2022
[34] *Guide to Social Science Data Preparation and Archiving*. **Portable software-specific files,** p. 48 (https://www.icpsr.umich.edu/web/pages/deposit/guide/). Accessed 11 August 2022

Recommendations	Not intended for archiving but as a medium for transport between different data environments.[35] CESSDA suggests conversion to plain text with Unicode encoding.[36]
File format/ extension	SAV/.sav
Format	Proprietary SPSS format. SPSS originally stood for Statistical Package for the Social Sciences.
Description	Initially command-based, it has evolved to use a graphical user interface (GUI). Syntax commands, SAX Basic and Python are three other ways to use SPSS if you do not want to use menus.
Recommendations	The .sav file format is not recommended for archiving but SPSS supports the generation of a number of more suitable formats. These include conversion to ASCII where you simultaneously create a setup file which also includes a definition of variables. Make sure afterwards that the export worked properly.

STATA	
File format/ extension	DTA/.dta
Format	Developed by STATA as a proprietary, cross-platform file format for rectangular data. The programme can only work with one dataset at a time.
Description	Command-based user interface. The data format has changed over time, but later versions can import data from earlier versions. Data can be saved in the current version, as well as in the format of the last previous version. However, this applies up to Stata 13. Later versions of Stata (Stata 14, file format .dta_118) can only save in their own current format.[37] Text is encoded with UTF-8 in Stata. Variable names may be up to 32 characters long with each variable string automatically terminated by a numerical zero (/0). Each UTF-8 character takes 1-4 bytes of space, so 32*4+1 = 129 bytes must be allocated for each variable name.[38]

[35] https://ppp.cessda.eu/doc/D10.4_Data_Formats.pdf p. 34. Accessed 11 August 2022
[36] https://ppp.cessda.eu/doc/D10.4_Data_Formats.pdf p. 34. Accessed 11 August 2022
[37] http://www.stata.com/help.cgi?dta 'description' and '2. Versions and flavors of Stata'. Accessed 11 August 2022
[38] http://www.stata.com/help.cgi?dta '3. Representations of strings'. Accessed 11 August 2022

Recommendations	The .dta file format is not recommended for archiving but STATA supports the generation of a number of more suitable formats. These include conversion to ASCII where you simultaneously create a setup file which also includes a definition of variables. Make sure afterwards that the export worked properly.

Common spreadsheet file formats

Lotus 1-2-3	
File format/ extension	123/.123 and .wk* (for example .wk4, .wks, etc.)
Format	Files created with the Lotus 1-2-3 spreadsheet programme. Binary and proprietary format.
Description	Lotus 1-2-3 was a popular software package in the 1980s and 1990s but was not further developed after 2002. The file format can be read by MS Excel 2000 and OpenOffice Calc where it can be converted to more suitable formats.
Recommendations	The format is not recommended for archiving. Problems may occur when converting files using other programmes (for example formulae used may be counted differently or even not work at all). It is therefore important to have documentation that explains key calculations and other additional features and functions.

Microsoft Excel	
File format/ extension	XLS/.xls
Format	A proprietary binary format used for Microsoft Excel (up to Excel 2003).
Description	Although the .xls format is proprietary and owned by Microsoft, it is widely used and can be imported by a number of third-party applications (for example OpenOffice and Google Docs). The format has been replaced by an XML-based format (.xlsx) in versions of Microsoft Office 2003 and later.
Recommendations	Although the file format is compatible with other open applications, .xls is not the standard format for Excel files (see .xlsx below) and is not recommended for archiving.

File format/ extension	XLSX/.xlsx
Format	Part of the Office Open XML (OOXML) format created by Microsoft. An ECMA[39] and ISO[40] standard.
Description	A format from Microsoft released with Office 2007. Microsoft chose to develop its own specification (OOXML) rather than use the existing ODF format. The format consists of human readable xml files with different types of information contained in a zipped file.
Recommendations	.xlsx is an open and well-documented standard that works for archiving, but embedded material should be stored separately. In simplified terms, the .xlsx file is a zipped archive but you should save the contents in an uncompressed format. Migration to plain text formats should be considered for maximum accessibility, but care should be taken to ensure that decimal numbers[41] are transferred correctly so that valuable information is not lost.

OpenDocument Spreadsheet/Calc	
File format/ extension	ODT/.ods
Format	Part of the OpenDocument package of file formats, which are ISO standards (ISO/IEC 26300:2006[42]) for XML-based document formats. The format is supported and used by several office applications.
Description	Like the OpenDocument Text (.odt) format, the .ods file is basically a packaged file containing several separate files with a document template, text (like XML) and embedded files (for example images).
Recommendations	As the format is an open XML-based format, it is suitable for archiving but should be stored in uncompressed form. If a document contains images and other added content, this should be stored separately in suitable formats. The .ods format may be used as an archiving format in cases in which .csv cannot adequately preserve all the important information contained in a file. It may be worth trying to convert to .xlsx to see which works best. It is worth noting

[39] ECMA-376. Accessed 11 August 2022
[40] ISO/IEC 29500-1:2008. Accessed 11 August 2022
[41] When converting an Excel file to CSV format, you should first note which cell values are in text or numerical format, etc. Once this has been done, convert all values to text format and then save the data as CSV. To ensure that the import from a CSV file is correct: Open a new Excel file, under the Data tab select From Text, locate the correct file, select Import, follow the instructions that come up in the import window. Note that a CSV file will be larger than an XLSX file because Excel compresses its files.
[42] https://www.iso.org/standard/43485.html, last visited 11 August 2022

	that Calc offers the opportunity to preserve character set information when exporting to .csv. This is a good option when converting from .xls to .csv as UTF-8 encoding is necessary.

OpenOffice/StarOffice	
File format/ extension	SXC/.sxc
Format	The format is part of the OpenOffice XML package and is used in OpenOffice 1.0 Calc. Later replaced by OpenDocument Format (.ods).
Description	The .sxc format (now replaced by .ods) is an XML-based format, like its successor. The format is still supported by a number of different applications, including OpenOffice. The format consists of a number of compressed XML files in which the actual data are contained in one file, while images and other information are contained in a separate register.
Recommendations	Although it is an open format and text-based, it is recommended to use a newer .ods format.

Quattro Pro	
File format/ extension	WQ/.wq*, QPW/.qpw
Format	Files created with the spreadsheet programme Quattro Pro (now part of the WordPerfect Office package).
Description	A proprietary format used by the Quattro Pro spreadsheet programme. Older files can be opened with Microsoft Excel 2000 (for Quattro Pro for Windows versions 1-5, Quattro Pro converter needs to be installed) but there may be problems importing the files into Excel. Charts, embedded objects, and macros contained in Quattro Pro files may be lost when imported into Excel. Current versions of Quattro Pro support the OOXML format.
Recommendations	Not recommended for archiving. The files should be exported to XML-based or text-based formats.