Data Archiving and Networked Services

DANS

Preferred formats

September 2015, version 3.0

>>>

Preferred formats

Data Archiving and Networked Services (DANS)
P.O. Box 93067 | 2509 AB Den Haag
Anna van Saksenlaan 51 | 2593 HW Den Haag
The Netherlands
+31 70 349 44 50 | info@dans.knaw.nl | dans.knaw.nl

Contents

1. Selecting file formats

Digital data are stored in file formats, which are often standard software formats. The software and file format selected will usually depend on the user’s primary purpose.

To create a table, for instance, spreadsheet software will be used more often than a word processor. This is because data tables require specific properties which are better supported by specialized software. This may include the ability to sort data, to use formulas, to set up a filter, and so on. If such information is stored from a spreadsheet application the user may expect the file format to preserve these properties, or ‘significant characteristics’. If the table is created using a word processor it is less likely for the software to support these properties. The word processing application, on the other hand, will be more suitable for formatting an article, for instance using a functional table of contents and page numbers. Such features will not be supported by the spreadsheet application.

When information is stored from a software program, it is usually saved in that program’s standard file format. This is, however, no guarantee that in the future the file contents can be used or displayed in the way that was intended when the file was created. Formats may, for instance, be dependent on particular software. Software can become obsolete or only support certain versions of formats. It is also possible that specific format properties only work in the software used, or even only in one specific version of this software. Files may also be dependent on the use of expensive or exclusive software that not just anyone can access.

To preclude the risk of obsolescence and ensure the accessibility and sustainability of important file properties, a number of measures can be taken. One of these measures is to use file formats that have a high probability of remaining useful for many years.

As a general guideline, DANS believes that the file formats best suited for long-term sustainability and accessibility:

- Are frequently used.
- Have open specifications.
- Are independent of specific software, developers or vendors.

In practice, it is not always possible to use formats which satisfy all of these criteria.

2. Preferred and acceptable file formats

DANS uses two file format categories:

- **Preferred formats** are file formats of which DANS is confident that they will offer the best long-term guarantees in terms of usability, accessibility and sustainability. Depositing research data in preferred formats will always be accepted by DANS.
- **Acceptable formats** are file formats that are widely used in addition to the preferred formats, and which will be moderately to reasonably usable, accessible and robust in the long term. DANS favours the use of preferred formats, but acceptable formats will in most cases also be allowed.

DANS strongly recommends data depositors to supply their data in the preferred formats included in the table in Section 2.1 below.

If your data are stored in other formats than those mentioned below, please contact DANS at info@dans.knaw.nl.

2.1 Overview

This table provides a brief overview of the DANS preferred and acceptable formats. Please refer to the notes below the table for a more detailed explanation of each data type.
The extensions are classified into different types of data, distinguished on the basis of their primary use. There is a lot of overlap: the preferred formats of most data types can in fact be seen as plain text files or markup language.

At the end of this document you will find an explanation of the abbreviations used.

This list may be updated from time to time due to the development of new file formats and the obsolescence of other formats.

§	Type	Preferred format(s)	Acceptable format(s)
2.2	Text documents	• PDF/A (.pdf)	• ODT (.odt) • MS Word (.doc, .docx) • RTF (.rtf) • PDF (.pdf)
2.3	Plain text	• Unicode (.txt)	• Non-Unicode (.txt)
2.4	Markup language	• XML (.xml) • HTML (.html; .xhtml) Note: When valid and complete (see notes) If needed: • Related files: .css; .xslt; .js, .es (see notes)	• SGML (.sgml)

2.5	Spreadsheets	• ODS (.ods) • CSV (.csv)	• MS Excel (.xls, .xlsx) • PDF/A (.pdf) • OOXML (.docx, .docm)
2.6	Databases	• SQL (.sql) • SIARD (.siard) • DB tables (.csv)	• MS Access (.mdb, .accdb), (v. 2000 or later) • dBase (.dbf) (v. 7 or later) • HDF5 (.hdf5, .he5, .h5)
2.7	Statistical data	• SPSS Portable (.por) • SPSS (.sav) • STATA (.dta) • DDI (.xml) • data (.csv) + setup (.txt)	• SAS (.7bdat; .sd2; .tpt) • R(*)
2.8	Raster Images	• JPEG (.jpg, .jpeg) • TIFF (.tif, .tiff) • PNG (.png) • JPEG 2000 (.jp2)	• DICOM (.dcm) (by mutual agreement)
2.9	Images (vector)	• SVG (.svg)	• Illustrator (.ai) • EPS (.eps)
2.10	Audio	• WAVE; BWF (.wav) • FLAC (.flac)	• AIFF (.aif, .aiff) • MP3 (.mp3) • AAC (.aac, .m4a)
2.11	Video	• MPEG-2 (.mpg, .mpeg, …) • MPEG-4 H.264 (.mp4) • Lossless AVI (.avi) • QuickTime (.mov)	• MKV (.mkv)
2.12	Computer Aided Design (CAD)	• AutoCAD DXF v. R12 (.dxf)	• AutoCAD, other versions (.dwg, .dxf)
2.13	Geographical Information (GIS)	• GML (.gml) • MIF/MID (.mif/.mid)	• ESRI Shapefiles (.shp & related files) • MapInfo (.tab & related files) • KML (.kml)
2.14	Images (geo reference)	• GeoTIFF (.tif, .tiff)	• TIFF World File (.tfw & .tif)
2.15	Raster GIS	• ASCII GRID (.asc, .txt)	• ESRI GRID (.grd & related files)
2.16	3D	• WaveFront Object (.obj) • X3D (.x3d)	• COLLADA (.dae) • Autodesk FBX (.fbx)
2.17	RDF	• W3C standards	
2.18	Computer Assisted Qualitative Data Analysis (CAQDAS)	• Formats used in application, processed according to each individual file's data type	• Application's export formats (ATLAS.TI copy bundle; NVIVO export project; …) • QuDEX

(*) Under examination

For each data type, a brief overview will be given of the preferred format chosen, the use of the data, and any conversion possibilities.

This document is intended as a guide for data depositors. It is far from the only list of recommendations regarding file formats in the world. There are numerous other sources and wikis about formats and risks. At DANS, we have evaluated several existing documents based on our own experiences with the file formats DANS has encountered.

In preparing this document we have consulted the following sources:

- http://guides.archaeologydataservice.ac.uk/g2gp
- http://www.digitalpreservation.gov/formats/index.shtml
- http://www.loc.gov/preservation/resources/rfs/index.html
- https://www.archivematica.org/wiki/Significant_characteristics

This is a dynamic document. A working group within DANS is responsible for monitoring file formats and updating the recommendations based on new developments.

2.2 Text documents

PDF, or Portable Document Format, developed by software giant Adobe, has a subtype called PDF/A which is specifically designed for long-term sustainability. PDF/A is the international standard for (formatted) text documents. A PDF/A file is a stand-alone document: all fonts and images are included in the file, so it is not dependent on other files on the computer to correctly display its content.

There are several types of PDF/A. PDF/A-1a is recommended for text documents that were created entirely on a computer (“born digital”). For digitized documents PDF/A-1b is a suitable format.

The Adobe Reader can be downloaded for free, although many computers will already have software installed with which PDF files can be opened.
Adobe software for creating PDF files is not free, but several free software bundles, such as OpenOffice and IrfanView, also provide PDF support. There are also free applications that can “print” documents to a PDF document, such as Bullzip PDF Printer.

To create a PDF file, the default settings need to be adjusted in order to generate the correct type of PDF/A.

2.3 Plain text

Plain text files usually have the extension TXT. These files can be easily opened with various software applications. Text files can use various character sets, for example to represent the Latin alphabet, punctuation marks and special characters. DANS is confident that the Unicode character set, using Byte Order Mark and UTF encoding, ensures that all characters in all computing environments are represented correctly.

2.4 Markup language

Standardized General Markup Language (SGML) and Extensible Markup Language (XML) are markup languages used for text documents and datasets, both to present them to people and to enable data exchange between computers.

XML, SGML and HTML are popular markup languages. If valid and complete (see below), these formats are considered to be adequate, accessible and sustainable.

Apart from these formats there are XML-based or SGML-based formats that can only be read by special software. Such files cannot be accepted without further verification; please check with DANS.

XML is a variant of SGML: all XML files are SGML files. Since XML has a much stricter syntax, it is easier to validate. HTML (Hypertext Markup Language) is another variant of SGML; it is primarily intended for the presentation of rich text (and layout) and hyperlinks to other documents.

In addition to "regular" HTML there is also XHTML, which is HTML under the stricter rules of XML.

SGML and XML are hardly being further developed. HTML has recently seen its latest version 5 officially recognized as W3C standard. As Web technology continues to develop, it is expected that HTML will continue to be developed further.

XML, HTML and SGML are common and suitable markup language formats, provided the file formats are *valid* and *complete*.

Validity
Valid markup language documents are both well-formed and comply with the rules that apply to the file formats.

Well-formed documents require that the content is defined in a particular manner. Well-formed XML complies with syntax rules that state, among other things, that the character set used is also the character set specified; that no prohibited characters are used in the file; that there is one root tag and that each <tag> is correctly terminated with a </tag>.

The rules governing the content of a markup document are described in a DTD (Document Type Definition) or (XML) schema file. At the top of XML and HTML documents there is a reference to the DTD or schema used. This reference should really lead to the schema file itself. Ideally, the schema should be attached, unless it is available at a reliable public service.
If a non-standard schema or DTD file is used, the data depositor should consult DANS beforehand.
Through schemas and DTDs, entirely new "file formats" can be defined, such as SVG (Scalable Vector Graphics, for vector images), TEI (Text Encoding Initiative, used to format and annotate text), and MathML (for mathematical formulas).
The World Wide Web Consortium (W3C) manages the specifications for HTML and XML, and provides a Markup Validator that can validate both XHTML and HTML. In addition, it can validate a number of other formats, such as MathML and SMIL: http://validator.w3.org/

Completeness
Markup language may be based on the use of other file formats, either in separate files or within one file. All files associated with an XML/HTML/SGML file must be included. Common markup language related files are XLST stylesheets, CSS definition files and JS/ES scripting languages.

Extensible Stylesheet Language Transformations (XSLT)
XSLT is an XML vocabulary for transforming XML files, among others.
XSLT is a widely supported open standard. We can accept it, provided the linked files are archived with it.

Cascading Style Sheets (CSS)
CSS is widely used on the Web to define the layout of markup language documents. There are different versions of CSS. For permanent archiving it must be made clear to which files the CSS files refer and which version of CSS is being used. As browser-specific extensions may exist, the target environment of the files must be known, unless only basic elements have been used. If the CSS files contain links to other CSS files or external files, these links must be working.

JavaScript/ECMAScript (.JS, .es, ...)
JavaScript and similar scripting languages serve many purposes. Browsers read the scripts and execute commands. The basics of JavaScript are well supported and can be archived, but when using script files there may be dependency on external data.

2.5 Spreadsheets

Spreadsheets are primarily used for dealing with tabular data: values in cells, arranged in rows and columns.

However, a spreadsheet is often much more than a flat table. Spreadsheets can be provided with further formatting, for example by using colours in the cells or changing the appearance of the lines between the cells. Also, the structure of a spreadsheet can be of importance. Cells may for example be based on calculations made with values in other cells. With spreadsheets, care should therefore be taken to consider what features are important to maintain; which are the ‘significant characteristics’ of the file.
The OpenDocument Spreadsheet (.ods) format is an open, fairly well supported and robust spreadsheet format which is recommended as the preferred format for the permanent storage of spreadsheets with calculations and/or other (structural) properties.

If a spreadsheet can be considered as or reduced to a flat table of rows and columns, then it is an option to create a CSV (Comma Separated Values) text file from the table. See the “CSV files” section below for an explanation of how to deal with this format.
CSV files are only suitable for the storage of flat tables. CSV does not retain formatting (text or cells), formulas or links to external resources.
Is direct visualization the primary purpose of the spreadsheet? Then the file may optionally be treated as a formatted text file and be submitted in PDF/A format. See the section on text documents for details.

PDF/A is primarily suitable for the presentation of formatted tables. The format offers limited support for features like spreadsheet formulas and links to external resources.

2.6 Databases

Databases exist in various forms, the best known probably being the relational database. Databases are managed by a Database Management System (DBMS). Besides ensuring consistency of data and their processing (reading and writing), the DBMS controls roles and privileges of users (groups) and also offers a range of functions to perform various operations on the data.
The file format is usually linked to the DBMS, but there are independent exchange formats too.

Many DBMSs support the ISO standardized version of Structured Query Language (SQL). This is a language for querying and updating relational databases. Together with the data definition language, used to define and modify schemas, the contents of a database can be stored as a collection of schema and data statements.
The language rarely changes, but various modifications and additions may change along with software updates. When extensions are used, the documentation must show which SQL version has been used.
In SQL it is possible to refer to non-existent or external data without invalidating the file. If SQL is used for data exchange, any references must therefore be supplied, or each reference must be replaced by the data referred to.

For relational databases, SIARD is seen as a suitable and sustainable format. SIARD (Software Independent Archiving of Relational Databases) is intended for archiving relational databases in a way that is as independent of the original DBMS as possible. This format takes into account all the significant characteristics of databases.
SIARD is an open, freely available format, based on clear text formats: Unicode, XML, SQL (1999). This makes it accessible for various tools.
SIARD is a relatively young format. There are tools for converting databases to SIARD and for validating the format, but the possibilities are limited. Some conversion tools such as AccessToSiard and CSV2SIARD can be found here: http://coptr.digipres.org/Category:File_Format_Migration. Using these tools requires the SIARD Suite. This program is explained, with a link to where it can be requested free of charge, on the same website:
http://coptr.digipres.org/SIARD_Suite
Databases can use routines which may be dependent on their own scripting languages or programming languages from the DBMS. When converting to SIARD there is the potential risk of loss of such routines, but this risk is not considered to be very large because such languages are not expected to become obsolete.

dBase, HDF5 and Microsoft Access are alternative database formats that can be considered acceptable, although it is better to convert them to more sustainable formats.

The dBase Table File Format (.dbf) is a proprietary product. The company behind dBase will still support older versions of the format. For versions earlier than 7 no official documentation is made available. dBase formats can be read in several different database applications, including LibreOffice/OpenOffice, MySQL and MS Access. From these applications, it is possible to export the dBase data into other formats.

Hierarchical Data Format (version 5, not compatible with earlier versions) is a common dataset format with the ability to store data in multidimensional arrays, grouped into collections and/or hierarchies. Relationships between data in the arrays can be saved, but the format does not allow for storing structured (descriptive) metadata. The format is open and can be read in a variety of applications, but it is very difficult to process without the use of HDF5 software. See the tools on http://www.hdfgroup.org/products/hdf5_tools/
In practice, Microsoft Access is widely used for creating databases. The Access MDB and ACCDB formats are, however, very poorly supported outside the commercial versions of Microsoft Access. Owing to the different versions of these formats, even Access itself will not always support the files very well.

DANS has enabled many databases created with Microsoft Access to be processed in a sustainable and accessible manner by storing the tables from the database as separate CSV text files. See the section on CSV files below for an explanation of how to deal with this format.
When storing the tables as CSV files, only the tabular data from a database are retained. Any parent documentation is described in a separate document. In Microsoft Access databases, the Database Documentation feature can be used to generate a document with column descriptions and table relationships: like formatted text, this document can be stored as a PDF/A file and supplied with the database tables.
It must be ensured that all codes and variables used can be explained. This may mean providing more detailed descriptions in a separate document or "code book".

CSV files
CSV (Comma Separated Values) provides a way to write tabular data as plain text. This format does not support data types and metadata beyond column titles. It is in fact based on the RFC4180 open standard, although there are different variants (dialects).
In a CSV file the values/cells from a table are separated by commas. CSV files can be imported into database applications, but they can also clearly and quickly be opened as a spreadsheet, for example in Microsoft Excel. These files can also be read as text files, for instance in Notepad.
Many applications will be able to open CSV files without problems. Depending on the computer's default settings for the use of separators, an application may not be able to automatically separate the columns. In the application, columns can be further distinguished on the basis of separation characters; also, the default on the computer can be adjusted. For Windows systems, this default setting can be found under List separator on the Region and Language screen. If the comma is selected as separator here, the CSV files will be correctly displayed in distinct columns in all applications.

2.7 Statistical data

There are several software packages with which statistical analyses can be performed. The most common are SPSS, STATA, R and SAS.

For long term archiving DANS uses the SPSS portable format. Although this is a proprietary format, it has been chosen for this purpose because information can be lost when dividing data and file information.
As the SPSS software is frequently used, it may be expected that the format will remain accessible in the future.
For software packages in which data cannot be converted to SPSS portable, DANS will archive data and setup information.
The SPSS portable file is suitable for permanent archiving, but in practice users are generally unfamiliar with this format. For the sake of accessibility, statistical data are therefore also made available in the standard formats of the most widely used statistical software packages:
-SPSS .sav
-STATA .dta

2.8 Raster images

For raster images, DANS recommends archiving them as uncompressed TIFF and also publishing them as JPEG files.

Devices such as computer monitors, printers and data projectors can process digital images. They translate the dots or pixels that make up a digital image to the device's specifications. The number and colours of the pixels determine the appearance of the digital photo. The pixels are the intersections of a fine grid, which is why these images are called raster images.

The quality of a raster image depends on the following factors, which are determined by the producer:

1. Resolution. The pixel dimensions of a raster image are the total number of pixels in the horizontal and the vertical dimension. The granularity or resolution is expressed as the number of pixels per inch (2.54 centimetres). The resolution must be tailored to the details of the object being digitized, i.e. not too coarse and not too dense.
2. Dynamics. To what extent does the raster image contain all the colours of the original and how are these colours coded; what colour space is applied? Accurate colour reproduction requires calibration of the recording equipment by an expert.
3. Compression. Because raster images may consist of millions of pixels, compression techniques can be used to reduce the file size.
4. Documentation. Descriptive and technical/administrative metadata. This can either be included in the raster image or created separately (or both). Many digital cameras support the EXIF standard, which contains descriptions such as the time of recording and the camera settings.
5. File format. The selected file format should efficiently and effectively support the above features.

With respect to the archiving and sustainability of raster images it is essential that they can be reproduced in the future in accordance with the intentions of the depositor. Using TIFF, JPEG and PNG it can be reasonably assumed that they will be displayed without problems and that standard image processing software will be available to render the images.

Uncompressed TIFF is the preferred format for DANS to preserve raster images in maximum quality in the long term.
TIFF files can be very large, however, which may be detrimental to user-friendliness. It is therefore recommended to use TIFF as a preservation format and also make the images available for use in the widely supported JPEG format.
The PNG format can also be characterized as a suitable preservation format. It is smaller than TIFF, but beware: PNG offers limited options for storage of technical/administrative metadata in the file. For example, the format provides no support for the abovementioned EXIF standard. When using PNG one must therefore take care to ensure that any relevant metadata are preserved.
In recent years there has been much discussion about which format is suitable for archival purposes. The Dutch National Library now regards JPEG2000 as its archive format, but there are many experts who have a preference for other formats. One problem is that the formats (both JPEG and TIFF) support various compression methods which cannot all be regarded as sustainable. The JPEG2000 format, which supports core coding, is seen as the most durable file format.
The JPEG2000 images must comply with Part 1 of the JPEG2000 image compression standard (ISO/IEC 15444-1). A program called jpylyzer can be used to validate the format and extract the technical properties:
http://jpylyzer.openpreservation.org/

DANS has come across the DICOM format (Digital Imaging and Communications in Medicine) a few times. It is a standard which describes how medical image information is to be stored, shared and printed.
The standard also defines network and application protocols and is used in so-called PACS (Picture Archiving and Communication Systems): systems that crate and manage images, such as medical equipment.
In the standard, various types of still and moving images are supported for various medical applications. DICOM supports widely used compression standards such as JPEG and JPEG2000 or MPEG-2 video sequences.
The copyright of the standard is in the hands of the National Electrical Manufacturers Association (NEMA). DICOM viewing software can be subdivided into two groups: (1) proprietary viewers that are part of the (medical) recording systems and (2) DICOM viewing software for PCs. The most popular non-proprietary viewers are freely available: DicomWorks, Osiris and IrfanView (a commonly used all-format viewer).
Adobe has developed a Photoshop plug-in that enables viewing DICOM images and “header information” (metadata) as well as exporting to other formats.
IrfanView is also capable of capturing images and/or animations (sequences of images) from DICOM files.
These images can be stored in a preferred format as an image or a movie. In consultation with the depositor, we must determine whether the context information in the DICOM files is relevant to archiving.

2.9 Vector images

SVG stands for Scalable Vector Graphics. It is a robust, XML-based file format for statistical and dynamic vector images. SVG is an open standard and support for this format has greatly increased over time.

SVG vector images can be opened in web browsers like Firefox, Safari, Google Chrome and Explorer. For further editing, applications like Adobe Illustrator or Inkscape can be used. Inkscape is a free download from http://inkscape.org and works with Windows, Mac OS X and Linux.

All common vector image formats (EPS, AI, WMF, CDR) can be opened in Adobe Illustrator and Inkscape and converted to SVG.

2.10 Audio

The most important characteristics of audio files are:

1. The duration of the audio signal.
2. Bit depth: the number of bits with which the sampled signal is stored. The more bits there are used, the more accurate we can store - and reproduce - the original signal.
3. Sampling rate: the number of samples of the original signal that are made per second. This is usually expressed in Hz. According to the *nyquist frequency*, the original signal can be reproduced exactly when the sampling rate is twice the highest signal frequency.
4. Number of channels: a value indicating the number of unique signals in the audio object, for example 2 (stereo).

In order to save storage space and bandwidth, a number of “lossy” file formats have been conceived. These formats sacrifice some of the sound quality by removing frequencies from the audio track in a smart way so that fewer data need to be stored. This results in a smaller file size. For permanent archiving it is desirable to submit a “lossless” file, in a format with the best quality without data loss. For a usable file size, however, it can be much more user-friendly to offer a lossy data export. Best practice is to assess case by case whether it is desirable to submit lossy formats as well as the lossless original data.

WAVE, developed by Microsoft and IBM, is the most widely used format for storing uncomplicated audio, therefore there are many tools available for playing or converting the format. There is a maximum file size of 4GB because of the 32- bit header. This corresponds to 6.8 hours of audio (CD quality 44.1 kHz, 16-bit stereo). For larger files, the European Broadcasting Union has created an extension to the WAVE format, the Broadcast Wave Format (BWF).

The AIFF format, developed by Apple, is also used to store uncompressed audio. It is similar to WAV but it is not as widely supported. Besides standard AIFF there is a variant called AIFF-C/sowt, which makes use of a different byte sequence resulting in smaller files. The two implementations of AIFF are identical in quality and both use the .aiff extension. The structural differences between the two implementations can cause problems when playing AIFF formats, especially with older applications that are not compatible with AIFF-C/sowt.

The open FLAC format applies lossless compression, but is less well known than other formats and thus has relatively little support.

The best known and most widely used lossy audio format is MPEG1 – Audio Layer 3 (MP3). It is very widely supported. A successor to MP3 is Advanced Audio Codec (AAC). Like MP3, it is a lossy format but it offers similar sound quality at a lower bitrate. As a result, the file size of AAC is smaller than MP3, without noticeable loss of quality.

2.11 Video

MPEG-2 is an open standard according to ISO/IEC 13818, developed for DVD and digital television. It contains a number of profiles (Simple, Main, 4:2:2) with various standard specifications for aspects such as rendition, size and data rate. MPEG-2 is not optimized for low bit rates (<1Mbit/s) but offers superior quality at higher bit rates (>3Mbit/s) as compared to MPEG-1.
MPEG-4 is an open standard developed from the MPEG video standard ISO/IEC 14496, based in part on Apple QuickTime (.mov). The enhancements were made primarily to facilitate the use of AV equipment over the Internet by means of better compression through the H.264 codec. MPEG-4 includes two main versions and contains a large number of profiles which are optimized for different purposes (Web streaming, voice, HD video, etc.).
As indicated, QuickTime (.mov) is used as a basis for the MPEG-4 standard. Although QuickTime offers the same functionality as MPEG-4 (or more), it is nevertheless recommended to use MPEG-4 whenever possible because MPEG4 is an established standard.
Matroska (.mkv) is an open source and very flexible alternative to existing container formats like AVI, ASF, MOV, RM, MP4, MPG, etc., and can contain almost all codecs. MKV, however, has the disadvantage of being supported by relatively little video software.

2.12 Computer Aided Design (CAD)

CAD (Computer Aided Design) is the use of computers to create digital drawings. Autodesk, with its major software application AutoCAD, is the absolute market leader in the field of CAD. This means that the most popular and widely used CAD formats are not open formats. Neither have open formats for exchanging CAD formats been developed.
AutoCAD's file formats are DWG and DXF. They are supported by virtually all other CAD applications. DXF is specifically designed to facilitate data interoperability between AutoCAD and other applications. DXF version R12 seems to be the best support for their correct and successful import into other applications.
A major problem with the use of DXF is the development of the DWG format. DWG now offers options of which not all properties can be saved in DXF. As yet, DXF R12 is the best option for CAD preservation in a relatively open and widely supported format. Always check, however, if the DWG export to DXF does not lead to loss of data. Otherwise, it is better to stick to DWG.
From AutoCAD it is easy to save CAD drawings as DXF R12: File => Save As => Files of type: AutoCAD R12/LT2 DXF. It is recommended to first clean up the CAD drawing in AutoCAD by deleting temporary information from the file using the purge (purge all) command.

CAD drawings can be formatted in a layout with an image for publication. Such formatted images can be printed directly from AutoCAD to PDF/A (File => Plot, use the Adobe PDF printer, set the properties to "PDF/A-1b: 2005 (RGB)"). This preserves the visual object of the image, which is an excellent solution. However, it will not be possible to re-import the digital drawing into CAD; the image loses its editable properties.

2.13 Geographical information (GIS)

A GIS (Geographical Information System) is used to create digital maps and images, usually vector graphics based on an underlying data table. This table can be accessed in the GIS application as tabular data.

The most important GIS applications are ESRI ArcGIS and Pitney Bowes MapInfo Professional. ArcGIS stores data mainly as Shapefiles: a .shp with at least two associated .shx and .dbf files, and optionally up to 12 additional files (.prj, .shp, .xml, ...).
MapInfo uses TAB files. Like the shapefiles, TAB files are collections of files belonging together. The main file is a .tab file with a tabular data file in .dat, .dbf or .xls format, plus optional associated files (.map, .id, .ind).
MapInfo TAB and ESRI Shapefiles are widely used and can be submitted for use if required. These formats are not suitable for long-term preservation, however. Both formats usually consist of binary data for which no assurance can be given that they are seamlessly accessible for other applications than the one in which they were created.

With a view to the long term it is advisable to store GIS data in an open, well-supported and robust text file. Two formats are suitable and have been selected as preferred formats for GIS:
- GML is an XML ISO standard for geographical data. Prior to becoming an ISO standard, support for GML was limited, but it has since grown and is expected to continue to improve.
- The MapInfo Interchange Format (.mif), usually associated with a .mid file, is the MapInfo export format, designed for GIS interoperability. It is a clear, well-documented, well-supported and stable ASCII text file format.

By default, GIS applications contain import options for GML and MIF as well as storage and export options to GML and/or MIF. For better export and import capabilities with ArcGIS, a "Data Interoperability extension" is available, which allows easy execution of bulk conversions.

2.14 Georeference images

Georeference images are raster images (TIFF, JPEG) with the capability of being read into Geographical Information Systems (GIS). The images are projected and scaled to a coordinate system (grid).
GeoTIFF is a metadata standard for adding georeferenced to a TIFF image. The metadata are included in the TIFF file itself. It is an open and well-supported format.

2.15 Raster GIS

Geographical Information Systems (GIS) are mainly used for the production of digital vector images (maps) with an underlying data table. GIS can, however, also be used for creating raster images. Based on GIS input, for example, acontour elevation map can be made. A raster image generated in GIS can be enhanced with a colour scheme.
Such a GIS image is often simply referred to as a *grid*.

Grid files associated directly with commercial packages will enjoy a low degree of openness, interoperability and robustness. An ESRI ArcInfo Grid (aka ArcGrid) can use various subdirectories with mostly binary files: .adf .nit, .dir, log, ...

It is recommended to convert grid files to ASCII text as much as possible. GIS applications may be expected to correctly import ASCII grid files.
The ESRI ArcGIS ArcCatalog provides "convert GRID to ASCII" capabilities, while Surfer has a Grid => Convert => Save to GS ASCII option.
Please note that the conversion possibilities are neither unlimited nor faultless.

2.16 3D

For the storage and display of 3D images/models no file formats have been developed that can easily be characterized as preferred formats. It is a recognized problem in the world of digital archiving: all kinds of 3D programs use their own formats, interoperability is limited and conversion to other formats quickly leads to loss of functionality or certain file properties.
3D data is best preserved in their original format. In addition, the possibility of exporting to an open format may be considered. The export format of primary preference is X3D. If X3D does not store the 3D model as desired, the COLLADA .dae format will be the next best choice. Check the export format to see if the desired properties have been saved; describe any elements that are missing.

For geometric objects only, without further aspects such as animation or interactivity, WaveFront OBJ is the preferred format. OBJ is a very widely supported open format for the display of 3D geometry. The spatial positions of each point of the object as well as its texture coordinates are written to a clear and simple structure.

Additionally, the possibilities of transferring parts of the data in an alternative way may be considered. Can videos (screencasts) or static images be used to display certain information?
Although there is no preferred format for an interactive, dynamic 3D model, there may be preferred formats for certain elements of it.

2.17 RDF

RDF (Resource Description Framework) is a data model in which knowledge is expressed in graphs and organized with labels. Several RDF standards are supported by the World Wide Web Consortium (W3C). It is expected that RDF applications will always be able to deal with the following W3C standards:

- RDF/XML (.rdf)
- Trig (.trig)

- Turtle (.ttl)
- NTriples (.nt)
- JSON-LD

2.18 Computer Assisted Qualitative Data Analysis (CAQDAS)

Unlike quantitative research, which is based on numerical data with which calculations are performed, qualitative research uses non-numerical qualitative data, such as texts, images and films. Qualitative data can be stored in a number of file types for which in many cases preferred and acceptable formats are available (such as TIFF or MP3).
CAQDAS (Computer Assisted Qualitative Data AnalysiS) applications are software programs that facilitate various research and analysis methods for the enrichment and analysis of qualitative data. Examples of CAQDAS applications are ATLAS.TI, HyperRESEARCH, MAXqda, QDA Miner, NVivo, Qualrus and Transana.
Over the course of time various releases of these applications have appeared, and there are versions for different operating systems (e.g. Windows, MacOS). Usually the files from these applications cannot be interchanged. CAQDAS applications use closed, proprietary formats and do not support open import or export formats. This hampers the sustainability of the data files.

A few years ago, an open standard was developed to represent “richly encoded qualitative data” and to serve as an archive format for CAQDAS files. The standard is known as QuDEX (Qualitative Data Exchange Format) and available, inter alia, in the form of an XML schema (see http://data-archive.ac.uk/create-manage/projects/qudex). The project has so far produced a proof of concept, including a QuDEX connection to the closed format of ATLAS.TI, but no practicable archival services have been developed based on the format. The standard is not yet supported by the manufacturers of the various CAQDAS applications.

In recent years, DANS has had experience with research data from two types of CAQDAS applications: ATLAS.TI and NVivo. When archiving files created by these applications, their export functions should be used:
ATLAS.TI has a “copy bundle” function. It is intended to create one coherent bundle of all files belonging together. This bundle serves as a backup for the files and enables file migration between computers.
NVivo (version 10) has a “save project” option with which all parts of a project (sources, nodes, matrices, casebook) can be copied. These files constitute the archive.
The application and operating system versions should be included in the documentation of the export files.

Abbreviations and acronyms

.7bdat = SAS version 7b dataset (file extension used in SAS)
AAC = Advanced Audio Coding
.accdb = Access Database (file extension used in Microsoft Access 2007 and later)
.adf = ESRI ArcInfo Data File
.ai = Adobe Illustrator (file extension used in Adobe Illustrator)
AIFF = Audio Interchange File Format
ASCII = American Standard Code for Information Interchange
AVC = Advanced Video Coding
AVI = Audio Video Interleaved
BWF = Broadcast Wave Format
CAD = Computer-Aided Design
CAQDAS = Computer Assisted Qualitative Data Analysis
CDR = CorelDRAW file format
COLLADA = Collaborative Design Activity
CSS = Cascading Style Sheets
CSV = Comma Separated Values
.dae = Digital Asset Exchange (file extension used in COLLADA)
.dbf = dBase file (file extension used in dBase)
DDI = Data Documentation Initiative (metadata standard for statistical data, see http://ww.ddialliance.org/)
DICOM = Digital Imaging and Communications in Medicine
.docx = Office Open XML document (file extension used in Microsoft Office)
.dta = data file (file extension used in STATA)
DTD = Document Type Definition
.dwg = Drawing (file extension used in AutoCAD)
.dxf = Drawing Interchange Format (file extension used in AutoCAD)
EPS = Encapsulated PostScript
ESRI = Environmental Systems Research Institute
ES = ECMAscript
.fbx = Filmbox (file extension used in Autodesk software)
FLAC = Free Lossless Audio Codec
GIS = Geographic Information System
GML = Geography Markup Language
H.264 = video coding standard, defined by the International Telecommunication Union (ITU). Also known as MPEG-4 Part 10 or AVC.
HDF = Hierarchical Data Format
HTML = HyperText Markup Language
IEC = International Electrotechnical Commission
ISO = International Organization for Standardization
JPEG = Joint Photographic Experts Group
JS = JavaScript
KML = Keyhole Markup Language
M4A = MPEG-4 Audio
MathML = Mathematical Markup Language
.mdb = Microsoft Access Database (file extension used in Microsoft Access 2003 and earlier)
MIF/MID = MapInfo Interchange Format / MapInfo Data File
MKV = Matroska Video
.mov = Apple QuickTime movie (file extension used in QuickTime)

MP2 = MPEG-1 Audio Layer II **MP3** =
MPEG-1/-2 Audio Layer III **MPEG** =
Moving Picture Experts Group
MPEG-2 = Moving Pictures Experts Group 2 (video format)
MPEG-4 = Moving Pictures Experts Group 4
MS = Microsoft (with software: MS Word; MS Excel, …)
.obj = WaveFront Object
ODS = OpenDocument Spreadsheet
ODT = OpenDocument Text.doc = document (file extension used in Microsoft Word)
OOXML = Office Open XML
PDF = Portable Document Format
PNG = Portable Network Graphics
.por = SPSS Portable (file extension of the SPSS exchange format)
R = software bundle and programming language, a modern implementation of S (Statistical programming language)
RDF = Resource Description Framework
RTF = Rich Text File
SAS = Statistical Analysis System (statistical data analysis software)
.sav = SPSS data file (file extension used in SPSS, the extension name is derived from 'save')
.sd2 = SAS dataset (file extension used in SAS)
SGML = Standardized General Markup Language
.shp = Shapefile (file extension used in ESRI software)
SIARD = Software Independent Archiving of Relational Databases
SPSS = Statistical Package for the Social Sciences (which is now being used outside the social sciences as well)
SQL = Structured Query Language
SVG = Scalable Vector Graphics
.tab = Table File (file extension used in MapInfo)
TEI = Text Encoding Initiative
.tfw = TIFF World File
TIFF = Tagged Image File Format
.tpt = SAS Transport file (file extension used in SAS)
TXT = text
UTF = Unicode Transformation Formats
W3C = World Wide Web Consortium
WAVE = Waveform Audio File Format
WMF = Windows Metafile
X3D = XML file format for rendering 3D computer graphics
XHTML = Extensible HyperText Markup Language
.xls = Excel Spreadsheet (file extension used in Microsoft Excel)
.xlsx = Office Open XML spreadsheet (file extension used in Microsoft Office)
XML = Extensible Markup Language
XSLT = Extensible Stylesheet Language Transformations

<<<

This is the Preferred Formats document. This document describes the file formats of which DANS is confident that they will offer the best long-term guarantees in terms of usability, accessibility and sustainability.
For each data type, a brief overview is given of the preferred format chosen, the use of the data and any conversion possibilities. This document is intended as a guide for data depositors.
For more information, please contact DANS.

Data Archiving and Networked Services (DANS)

DANS promotes sustained access to digital research data. For this, DANS encourages scientific researchers to archive and reuse data in a sustained form, for instance via the online archiving system EASY (easy.dans.knaw.nl) and DataverseNL (dataverse.nl). With NARCIS (narcis.nl), DANS also provides access to thousands of scientific datasets, publications and other research information in the Netherlands. The institute furthermore provides training and consultancy and carries out research on sustained access to digital information. Driven by data, DANS ensures the further improvement of access to digital research data with its services and participation in (inter)national projects and networks. For more information and contact options, please visit dans.knaw.nl.

Data Archiving and Networked Services (DANS)
P.O. Box 93067 | 2509 AB Den Haag
Anna van Saksenlaan 51 | 2593 HW Den Haag
The Netherlands
+31 70 349 44 50 | info@dans.knaw.nl | dans.knaw.nl

DANS is an institute of KNAW and NWO

KNAW NWO

Driven by data