TECHNICAL REPORT

ISO/TR 22299

First edition
2018-11

Document management — Digital file format recommendations for long-term storage

Gestion électronique — Recommandations de format de fichier numérique pour le stockage à long terme

Reference number
ISO/TR 22299:2018(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword ..... **iv**
Introduction ..... **v**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Basic selection criteria according to the content type description **1**
4.1 General 1
4.2 Selection methodology 2
4.2.1 File format description 2
4.2.2 Long-term availability of readers or players 2
4.2.3 File presentation stability 3
4.2.4 Software and/or operating system migrations 3
4.2.5 File format selection 3
5 File formats **4**
5.1 General 4
5.2 Coded text 4
5.3 Vector graphics 4
5.3.1 2D graphics 4
5.3.2 3D graphics 4
5.3.3 Technical drawings 5
5.4 Images 5
5.5 Sound 5
5.5.1 Linear formats for sound files 5
5.5.2 Lossless compression formats 6
5.5.3 Lossy compression formats 6
5.5.4 Container formats 6
5.6 Video 6
5.6.1 General 6
5.6.2 Coding 7
5.6.3 Digitalization 7
5.6.4 Compression 7
5.6.5 Video container formats 8
5.7 Office automation 8
5.8 Formats suitable for preservation 8
Bibliography **10**

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 171, *Document management applications*, Subcommittee SC 2, *Document file formats, EDMS systems and authenticity of information.*

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

The document management industry is heavily reliant on standardized file formats for both long-term storage and interoperability purposes.

Effective document management often requires the selection of an appropriate storage file format and eventually conversion between the native digital document format and the selected storage file format.

This document provides information and guidelines on file formats to assist in the selection of file formats.

Document management — Digital file format recommendations for long-term storage

1 Scope

This document gives guidelines for selecting the most appropriate file format(s) for the storage, usability, and exchange of data with a long-term management objective.

It is applicable to the selection of file formats to be used to store electronic documents. It provides guidance that takes into account:

— the durability of documents in a readable form;

— fidelity to the original and data integrity;

— interoperability, i.e. independence from creation applications, information systems and rendition platforms;

— compliance with relevant laws and regulations;

— compliance with format specifications;

— reducing costs by reducing the number of conversions/migrations over time.

This document is applicable to all office activities (e.g. text processing, spreadsheets, presentations), email and static web pages, as well as all types of electronic components, including images, video and sound.

It does not apply to database formats.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 12651-1, *Electronic document management — Vocabulary — Part 1: Electronic document imaging*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 12651-1 apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

4 Basic selection criteria according to the content type description

4.1 General

The following criteria can be considered when selecting a file format:

— the file format functionality, i.e. the type of content it is able to support (e.g. text only, enhanced text with images or style sheets, images, video, sound);

— the file format specifications that are made available as an open standard;

— the file format that can be used in the intended application;

— the metadata that can be incorporated into the file;

— the likelihood that a reader or a player will still be available on a long-term basis;

— whether the file format has widespread support by the industry and vendors.

4.2 Selection methodology

4.2.1 File format description

The unconstrained availability of a file format specification is essential for the development of software products, now and in the future, that are capable of correctly representing the content of files of this type.

End users should seek assurance of the openness (free availability) of a file format specification before using this format for long-term storage. If file format specifications are not freely available, the file format is not recommended for long-term retention and could be used only after a comprehensive risk analysis.

The file format should be available as an open standard, which has been developed and is maintained by an authoritative, neutral standardization body with no copyright restrictions or fees for use.

Electronic content can be stored in a document management environment so that software and/or users can use the content. There are standardized and non-standardized file formats that can be considered. Non-standardized formats should only be used with caution and only if the file format is fully documented. Examples of standardized file formats include JPG and PDF (and the PDF sub-sets). Non-standardized, but widespread (and commonly used) formats include TIFF, which is a proprietary format. The decision to use (select) standardized formats versus non-standardized formats should be considered by the end-user organization and is dependent on other aspects of the document management system. For example, a document may be received in PDF format, but then its pages may be extracted into TIFF or JPG for further processing, such as data extraction, etc.

4.2.2 Long-term availability of readers or players

From a long-term storage/archival perspective, the organization should always take into account the potential need to migrate and/or convert existing formats. As technology continues to mature and expand, file formats are being updated as required. For example, the PDF subsets that are now available. As a result, formats that are in use today may need to be updated to ensure the usability of the information they contain is retained in the future.

An organization may need to maintain the originals of documents that contain essential information for authenticity and integrity, such as digital signatures, seals or timestamps, recognizing that migration/conversion to another format could invalidate those elements. The organization should recognize the existence of different use cases for file formats and take this into account when selecting long-term file formats.

It is also important to take into consideration that a tool or application has to be available that can properly open and display the contents of the file. These “readers” should be kept up-to-date so that they are able to function in the current operating environment. In cases where non-standardized formats are used, it is important that the organization is able to maintain a reader to open/read the files. As technologies change and expand (e.g. a new sub-set of PDF), the organization should verify that the reader is not only able to open/display new files, but also legacy files.

There are three strategies for managing reader applications:

— porting the existing software to new operating systems;

— developing new software for new operating systems;

— emulation supporting the continuous usage of old software in new computing environments.

The first two options are suitable for widely adopted file formats. Porting is considered when the corresponding cost is relatively low. Developing new software allows the user to add new functionalities and to improve usability.

4.2.3 File presentation stability

Content retained for legal or records management purposes should be stored using tamper-resistant file formats.

Files should not depend on external resources that could be modified or become unavailable in the future.

Files should not contain embedded code (e.g. macros) or other features that could change the representation of the file content.

Enhanced text is characterized by the fact that:

— letters can be presented using different fonts;

— images can be represented using different file formats.

A reader may only support a reduced set of fonts. If there is a need to use one or more fonts in addition to those of that reduced set, the additional fonts should be embedded inside the file. Since this can increase the file size, it can be preferable to only use the fonts that are supported in the reduced set.

Where different fonts are supported by the reader, it is preferable to allow only embedded fonts, in order to avoid external dependencies.

A reader may only support a reduced set of image formats. It may support additional formats using external readers. However, the availability of these external readers should be demonstrated in the same way as those of the text readers.

4.2.4 Software and/or operating system migrations

Tests should be performed to provide assurance of the fidelity of the rendering when:

— porting the existing software to new operating systems;

— developing new software for new operating systems;

— emulation supporting continuous usage of old software in new computing environments.

4.2.5 File format selection

Different file formats may be considered where the content to be stored is coded text, enhanced text, 2D graphics, 3D graphics, images, sound or video. These formats are addressed in Clause 4.

Consideration should be given to the following criteria when selecting a file format:

— any intellectual property associated with the use of the format;

— available software tools for reading and writing the format;

— long-term access to the technical specification(s) defining the format;

— certification and/or compliance related to the format.

5 File formats

5.1 General

To reduce the volume of information processing, it is important to consider compressing the data (e.g. images, sound and video) while preserving the required quality and usability (e.g. evaluating the sound quality for the listener). For digitizing analogue materials or digital recordings for the purposes of long-term preservation, any lossy compression process should be avoided. Only a few of the numerous compression methods are identified below. It is important to understand that the same format name may be shared by a family of sub-formats with different compression characteristics.

5.2 Coded text

Plain text file contains only characters and special symbols. Different encodings can be used. See ISO/IEC 646, ISO 1073 (all parts), ISO/IEC 8859 (all parts) and ISO/IEC 10646.

NOTE Text documents requiring the advanced visual representation of the text to be preserved can be better served by using a different preservation format.

5.3 Vector graphics

5.3.1 2D graphics

Example file formats for two-dimensional (2D) graphics are:

— computer graphic metafile (CGM), see ISO/IEC 8859 (all parts), which is a standard for the exchange and retention of 2D graphic data;

— portable document format (PDF), see ISO 32000 (all parts);

— PDF/E, see ISO 24517-1, which is recommended for the creation of documents used in engineering workflows;

— PDF/A, see ISO 19005 (all parts), which is recommended where long-term storage is required;

— scalable vector graphics (SVG), which is a recommendation of the World Wide Web Consortium (W3C), based on the language XML for 2D graphics. SVG fidelity can vary significantly according to the viewer.

5.3.2 3D graphics

Example file formats for three-dimensional (3D) graphics are:

— JT, see ISO 14306, which is used primarily in industrial use cases as the means for capturing and repurposing lightweight 3D product definition data;

— PDF 2.0, see ISO 32000-2;

— PDF/E, see ISO 24517-1;

— product representation compact (PRC), see ISO 14739-1;

— STEP, see ISO 10303 (all parts), which can represent 3D objects and related information;

— extensible 3D (X3D), see ISO/IEC 19775 (all parts), which is a 3D-directed graphic and multimedia file format. It was created by the consortium Web3D with the aim of succeeding VRML 2.0.

5.3.3 Technical drawings

Example file formats for technical drawings are:

— PDF, see ISO 32000 (all parts), which is for technical drawings exchanges in unrevisable mode;

— PDF/E, see ISO 24517-1.

5.4 Images

Example file formats for images (raster graphics) are:

— bitmap (BMP), which is a very widely used proprietary format adapted to large-sized images; it is not generally compressed, however, this format accepts a lossless compression;

— digital negative (DNG), which is a format derived of TIFF that records the raw signals of cameras;

— graphic interchange format (GIF), which is suitable for images in 256 colours or fewer;

— Joint Photographic Experts Group (JPEG), which is widely used in digital photography and is defined in ISO/IEC 10918 (all parts);

— JPEG 2000, see ISO/IEC 15444-1 and ISO/IEC 15444-2, which is an evolution of JPEG with greater compression, lossless support and greater bit depths;

— portable network graphics (PNG), which supports 16 million colours and is defined in ISO/IEC 15948;

— tagged image file format (TIFF), which is a BMP graphic file format widely supported by image-manipulation applications and which can support multi-page images;

— tagged image file format/electronic photography (TIFF/EP), which is a digital image file format standard, see ISO 12234-2;

— PDF, see ISO 32000 (all parts);

— PDF/A, see ISO 19005 (all parts);

— PDF/E, see ISO 24517-1.

Many formats support image compression to reduce the amount of space required for storage. Using lossy or lossless compression should be considered according to the context. Lossy parameters can impact image quality depending on the compression–decompression processes used.

5.5 Sound

5.5.1 Linear formats for sound files

Example linear formats for sound files are:

— pulse code modulation (PCM), which is a generic name indicating a process of digitalization of the audio data without compression; the resolution is defined by the frequency of sampling (kHz) and the length of the digital word describing the sample (number of bits), it is generally admitted that the minimal quality of the restoration of the music corresponds to the format of the audio digital compact disk (see CD audio format below);

— linear pulse code modulation (PCML) format, which proposes a type of multi-channel encoding (up to six ways);

— CD audio (CDA) format, which contains digital characteristics of conversion strictly dedicated to a physical media (a compact disk); the CDA quality (44,1 kHz on 16 bits) establishes a reference point.

5.5.2 Lossless compression formats

An example of a lossless compression format for sound files is:

— free lossless audio CODEC (FLAC), which is a format of compression–decompression (CODEC) without loss, specific for audio; reduction of size can be from 30 % to 70 % according to the characteristics of the source; it can process any audio data PCM, at any depth and frequency of sampling, from one to eight channels.

5.5.3 Lossy compression formats

Examples of lossy compression formats for sound files are:

— MP3, which is a process of coding with compression developed by the Moving Picture Experts Group (MPEG); it is an abbreviation of MPEG 1 (and MPEG 2, which establishes an extension) layer 3; the compression ratio MP3 varies 1/4,4 in 1/44 (320 kb/sec to 32 kb/sec);

— advanced audio coding (AAC), which is an extension of the coding MPEG 2 stemming from several firms, chosen by Apple that developed its own rights management system (FairPlay);

— Vorbis, which is a free compression process with 10 levels of compression.

5.5.4 Container formats

As the name implies, container formats are formats containing various types of audio data (with or without compression).

Example container formats for sound files are:

— wave (WAV), which is a very widespread proprietary format that can welcome numerous file types whether compressed or not;

— broadcast wave format file (BWF), which is a format audio container defined by the European Union of broadcasting (UER) and the European Broadcast (EBU); it integrates only coded WAV formats in PCM or MPEG and allows the insertion of documentary data;

— OGG, which is a free format and generally embeds the open format Vorbis;

— audio interchange file format (AIFF), which is a proprietary format for compressed data only.

Lossy compression processing should be avoided during the digital conversion of audio documents intended for long-term preservation. Audio formats for archiving (such as WAV or BWF) embed a linear format (PCM) of the highest possible resolution.

5.6 Video

5.6.1 General

The specific objectives and needs of the numerous processes involved in capture and handling of video materials (e.g. the requirements and constraints related to acquisition, recording, transfer, use and storage of the information) should be taken into account. It is important to use a capture process that produces the best possible quality images within the acceptable economic conditions.

Generally, there are three steps of processing:

— coding of the primary components (components – composite PAL/SECAM/NTSC);

— digitalization (Betacam, DV and DVCPRO);

— compression.

5.6.2 Coding

Analogue coding of primary signals can be realized in two ways:

— system of coding in components: RVB, YUV, YC;

— system of coding in composites, i.e. standards assembly of the components in a single signal following rival standards: PAL, SECAM and NTSC.

5.6.3 Digitalization

Coded analogue signals are digitalized with variable resolutions:

— in cases involving the conversion of signal components, the selection and distribution of samples of luminance and chrominance on a line (a line or two maybe ignored), e.g. 4:2:2 (digital Betacam) or 4:1:1 (DV and DVCPRO);

— in cases involving composite signals, the sampling directly concerns the expanding signals of luminance and chrominance.

5.6.4 Compression

The quantity of captured information (e.g. the file size) and bandwidth constraints frequently requires compression. Compression techniques are based on the redundancy of an image (degree of uniformity), the redundancy from one image to another image (degree of change), the desired quality (imperfections tolerance), and, finally, on the economy constraints relative to the digital description of information.

Standards relative to the coding, conversion, compression and management of content constitute "toolboxes" corresponding to various applications, e.g. operation in a given context, viewing, backup, preservation of information.

The MPEG standards are as follows.

— MPEG 1: reduction of the flows at 1,5 Mbits/sec.

— MPEG 2: reduction applied to the signal components up to 10 Mbits/sec.

— MPEG 3: aims at flows from 18 to 25 Mbits/sec.

— MPEG 4: elaborated by means of specifically developed signal processing, this standard, see ISO/IEC 14496 (all parts), has the authority to accommodate typical multimedia programs, from the most narrowband (64 kb/sec) to the broadest (1,2 Gb/sec). It covers several applications (creation, scientific imaging, sound synthesis, games, digital TV, etc.) with high degree of interactivity.

  MPEG 4 Part 10, see ISO/IEC 14496-10, proposes a very wide range of flows answering, in a flexible way, every type of application for transmission and recording up to high definition (HD).

  In cases of high-definition video, the Blu-ray Disc format supports a useful maximum flow of 48 Mb/sec. This format supports several average and high definitions and the main codecs (MPEG-2, MPEG-4).

  Sound with HD is the object of specific coding, due to the required quality and the multi-broadcasting by six or eight channels.

— MPEG 7 (interface of the description of multimedia contents): this standard is not connected to the ones described above. It concerns the representation of multimedia contents (digital or analogue). All the elements of the document can be taken into account: creation (authors, interpreters, producers), use [rights of access, distribution (broadcasting)], physical media and digital format (compression), description of contents, abstract aspects (display, criticism of the contents).

5.6.5 Video container formats

There are several container formats of video data. The most widespread are:

— audio video interleave (AVI), which is proprietary, developed by Microsoft, and which supports all types of video or audio compressed files;

— MPEG; certain MPEG standards (MPEG 2/4, MPEG 21) support interoperability of various multimedia contents;

— MOV Quick Time, which is proprietary, developed by Apple, and which manages numerous video formats, even the standard HD;

— MXF, which is developed by the Society of Motion Picture and Television Engineers (SMPTE), and which is an object-based file format that wraps video, audio and other bitstreams optimized for content interchange or archiving by creators and/or distributors;

— OGG (OGG media video), which is practically the only free format and supports formats free of video compression "Théora", as well as audio free formats;

— real media (RM), which is proprietary, developed by RealNetwork, and which associates the formats of video compression (RealVideo) and audio (RealAudio) to allow the broadcasting of continuous streaming;

— transport stream (TS), which formats the communication of audio, video programmes with technical data as specified in MPEG 2 Part 1;

— video object file (Vob), which is a format based on the standard MPEG 2 developed for DVD-Video;

— Windows video media (WMV), which is a proprietary format, frequently used for the communication of audio-video files.

5.7 Office automation

When considering office documents, the recommended file format choices include:

— for revisable documents:

  — ODF, see ISO/IEC 26300 (all parts), which is an office automation format based on the language XML;

  — Open Office XML, see ISO/IEC 29500 (all parts);

— for unrevisable documents:

  — PDF 1.7, see ISO 32000-1, which was standardized by ISO in 2008 and its specifications made publicly available; before this standardization, the PDF format had become a de facto standard due to its widespread adoption;

  — PDF 2.0, see ISO 32000-2, which was published in 2017 and is the first PDF specification developed within the ISO process; it enhances the PDF format in several ways, including clarifications, new and improved features, various deprecations, etc.

5.8 Formats suitable for preservation

Example formats for unrevisable documents include the following.

— PDF/A-1, see ISO 19005-1, which answers the problems of long-term archiving. It is important to note that there are two conformance levels for PDF/A-1.

— PDF/A-2, which is based on a PDF 1.7, see ISO 32000-1, and which offers a number of new features:

  — JPEG 2000 image compression;

  — support for transparency effects and layers;

  — embedding of OpenType fonts;

  — provisions for digital signatures in accordance with the PDF advanced electronic signatures (PAdES) standard,

  — the possibility to embed PDF/A files in PDF/A-2 for archiving sets of documents as individual documents in a single file.

  Part 2 defines three conformance levels: PDF/A-2a, PDF/A-2b and a new conformance level PDF/A-2u. PDF/A-2u represents Level B conformance (PDF/A-2b), with the additional requirement that all text in the document have Unicode mapping.

— PDF/A-3, which allows embedding of arbitrary file formats (such as XML, CSV, CAD, word-processing documents, spreadsheet documents and others) into PDF/A files.

PDF/A-3 should be used with care because it allows arbitrary resources to be packaged alongside the document.

Bibliography

Cited documents

[1] ISO/IEC 646, *Information technology — ISO 7-bit coded character set for information interchange*

[2] ISO 1073 (all parts), *Alphanumeric character sets for optical recognition*

[3] ISO/IEC 8859 (all parts), *Information technology — 8-bit single-byte coded graphic character sets*

[4] ISO/IEC 10646, *Information technology — Universal Coded Character Set (UCS)*

[5] ISO 10303 (all parts), *Industrial automation systems and integration — Product data representation and exchange*

[6] ISO/IEC 10918 (all parts), *Information technology — Digital compression and coding of continuous-tone still images: JPEG File Interchange Format (JFIF)*

[7] ISO 12234-2, *Electronic still-picture imaging — Removable memory — Part 2: TIFF/EP image data format*

[8] ISO 14306, *Industrial automation systems and integration — JT file format specification for 3D visualization*

[9] ISO/IEC 14496 (all parts), *Information technology — Coding of audio-visual objects*

[10] ISO 14739-1, *Document management — 3D use of Product Representation Compact (PRC) format — Part 1: PRC 10001*

[11] ISO/IEC 15444-1, *Information technology — JPEG 2000 image coding system: Core coding system — Part 1*

[12] ISO/IEC 15444-2, *Information technology — JPEG 2000 image coding system: Extensions — Part 2*

[13] ISO/IEC 15948, *Information technology — Computer graphics and image processing — Portable Network Graphics (PNG): Functional specification*

[14] ISO 19005 (all parts), *Document management — Electronic document file format for long-term preservation*

[15] ISO/IEC 19775 (all parts), *Information technology — Computer graphics, image processing and environmental data representation — Extensible 3D (X3D)*

[16] ISO 24517-1, *Document management — Engineering document format using PDF — Part 1: Use of PDF 1.6 (PDF/E-1)*

[17] ISO/IEC 26300 (all parts), *Information technology — Open Document Format for Office Applications (OpenDocument) v1.2*

[18] ISO/IEC 29500 (all parts), *Information technology — Document description and processing languages — Office Open XML File Formats*

[19] ISO 32000 (all parts), *Document management — Portable document format*

Information available on public sites

[20] The Library of Congress. Washington, DC, USA: https://www.loc.gov/

[21] The National Archives. Richmond, UK: http://www.nationalarchives.gov.uk/

[22] Bibliothèque Nationale de France. Paris, France: http://www.bnf.fr/