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

© ISO 2018

iTeh Standards
(https://standards.iteh.ai)
Document Preview

ISO/TR 22299:2018
https://standards.iteh.ai/catalog/standards/iso/b3641d24-789f-4f6c-b4b7-98bb07f93616/iso-tr-22299-2018

Contents

Page

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.

iTeh Standards
(https://standards.iteh.ai)
Document Preview

ISO/TR 22299:2018
https://standards.iteh.ai/catalog/standards/iso/b3641d24-789f-4f6c-b4b7-98bb07f93616/iso-tr-22299-2018

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.