International Standard

ISO 20271-2

Document management — Reference model for long-term preservation of textual documents —

Part 2: Fundamentals

First edition
2026-05

Reference number
ISO 20271-2:2026(en)

COPYRIGHT PROTECTED DOCUMENT

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

Published in Switzerland

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 document 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).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

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*.

A list of all parts in the ISO 20271 series can be found on the ISO website.

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

Over time, numerous file formats have been created and subsequently become obsolete, resulting in digital files that are no longer accessible.

This situation typically arises when the technologies, software environments, or underlying specifications and standards – whether international, industry-based, or proprietary – are no longer maintained, and when insufficient information is available to interpret the file structure. Consequently, digital documents created decades ago can become unreadable or unanalyzable, even though the data itself still exists. This challenge has prompted sustained discussion among governments and organizations regarding the long-term preservation of digital documents and has established digital preservation as a critical issue in electronic document management.

The primary objective of this document is to support the long-term preservation of textual documents by ensuring that they remain technically interpretable and understandable despite potential format obsolescence. To achieve this, this document defines a reference model that enables systematic technical analysis and quantitative evaluation of textual document formats, while accommodating different preservation requirements and levels of available information.

This document defines multiple abstraction layers for textual documents and specifies the categories of properties associated with each layer. It establishes technical criteria for recording and assessing these properties within specific file formats, in order to identify risks related to long-term accessibility and interpretability.

The reference model defined in this document serves as a practical resource for professionals involved in document management, including institutional archivists and records managers, by providing a common basis for evaluating the long-term preservation readiness of textual document formats, in order to support consistent structure, interoperability, and long-term interpretability across different technologies and systems. In addition, the reference markup presented in this document can be used as a reference when developing new textual document formats or when enhancing the long-term preservation capabilities of existing formats.

Accordingly, this document supports the following activities:

— format analysis for selection and evaluation of textual document formats for long-term preservation;

— technical design activities related to the development of new textual document format specifications;

— activities aimed at improving existing textual document standards through the addition of properties or structural refinements;

— classification and comparative analysis of textual document formats.

The ISO 20271 series currently consists of the following parts:

— Part 1 (ISO 20271-1[1)]) provides an overview and contextual background for this document;

— Part 2 (this document) defines the fundamental concepts of the reference model;

— Part 3 (ISO 20271-3[2)]) defines a taxonomy and XML-based reference markup for digital preservation.

1) Under preparation. Stage at the time of publication: ISO/DIS 20271-1:2026.

2) Under preparation. Stage at the time of publication: ISO/WD 20271-3:2026.

Document management — Reference model for long-term preservation of textual documents —

Part 2: Fundamentals

1 Scope

This document specifies fundamental concepts of the reference model for textual documents and provides guidance to support long-term preservation from the perspectives of its five layers.

It defines:

— the layers that constitute the reference model for textual documents;

— the types of elements incorporated within textual documents;

— property types associated with textual documents;

— classifications of properties by type; and

— properties inherent to textual documents relevant to long-term preservation.

This document does not cover:

— specific technical methods for checking whether the properties exist within a specific textual document;

— specific technical methods for analysing particular textual document format (e.g. DOC, DOCX, ODT, TXT, PDF);

— specific metadata items for the long-term preservation of textual documents;

— processes, procedures, or management practices related to long-term preservation or records management.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

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

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

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

3.1
Common file formats
documents formats include plain-text format (TXT), Office Open XML (DOCX), Open Document Text (ODT), Portable Document Format (PDF), Open Word Processor Markup Language (OWPML), TeX and Hypertext Markup Language (HTML)

3.2
element
component included in a *textual document* (3.6)

3.3
property
attribute, *element* (3.2), and other components found in *textual documents* (3.6), which are subject to long-term preservation

3.4
rendering engine
software component responsible for converting document content (such as text, images, and formatting instructions) into a visual or printable output on various devices, like screens or printers

Note 1 to entry: It interprets the document's code or format and displays it in a way that users can view or interact with.

3.5
semantic information
properties of *textual documents* (3.6) that encompass all semantic content

Note 1 to entry: This includes both the substantive information and the structural aspects that convey meaning within the document.

3.6
textual document
document that conveys its core message primarily through the use of human language characters, regardless of the encoding or rendering method used

Note 1 to entry: A textual document may also include structured layouts, stylesheets, images, audio and other embedded content elements.

Note 2 to entry: Common file formats for textual documents include plain text (TXT), Office Open XML (DOCX), OpenDocument Text (ODT), Portable Document Format (PDF), Hangul Word Processor XML (HWPX), Hypertext Markup Language (HTML) and TeX.

3.7
unicode
character encoding standard maintained by the Unicode Consortium designed to support the use of text written in all of the world’s major writing systems

3.8
vector image
form of computer graphics in which visual images are created directly from geometric shapes defined on a Cartesian plane, such as points, lines, curves, and polygons

4 Textual documents

Textual documents can be represented in a variety of file formats.

Common file formats for textual documents include TXT, DOCX, ODT, PDF, HWPX, HTML and TeX.

Textual documents can encompass a wide range of content, from simple text to multimedia elements like images, videos and audio. They additionally support rich expressions through various styles, complex layouts and integration with external elements such as fonts.

The structural content types in textual documents range from simple text-only formats to those incorporating multimedia elements like images and videos. Furthermore, these documents contain various properties enabling rich expression – such as diverse styles and complex layouts – through integration with external elements like fonts.

The reference model defined in this document aims to provide a layered abstraction of technical information, which helps break down the structure of textual documents. This breakdown facilitates the establishment of evaluation criteria for long-term preservation and the categorization of documents. In practical applications, this reference model can require additional layers beyond the five foundational abstract layers initially identified for textual documents. These foundational layers typically include aspects such as content, structure, presentation, interaction and metadata. Additional layers can be necessary for documents with non-textual primary content, such as spreadsheets (for numerical data) and presentation files (with dynamic features like animations). This document primarily focuses on text-centric documents that store, preserve and deliver information conveyed through text content, ranging from simple structured formats to those with complex layouts.

Figure 1 — Examples of textual documents

Figure 1 illustrates various types of textual documents. These documents can have the following layout characteristics:

— a header, footer, and body aligned to fit a specific paper size;

— a body composed of several paragraphs, each of which can be represented by one or more sections;

— paragraphs constructed from characters encoded in various standards [such as Unicode, American Standard Code for Information Interchange (ASCII), Shift-JIS, Extended Unix Code for Korean (EUC-KR), Big5];

— images, and tables, including various styles to decorate them.