Technical Report

ISO/IEC TR 5812

Information technology — Semantic metadata support in office documents

Technologies de l'information — Support de métadonnées sémantiques dans les documents bureautiques

First edition
2026-10

Reference number
ISO/IEC TR 5812:2026(en)

© ISO/IEC 2026



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2026

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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 ..... **iv**
Introduction ..... **v**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Document semantic metadata **1**
5 Office document communities **2**
6 Typical use cases **2**
6.1 Themes 2
6.1.1 Information discovery and retrieval 2
6.1.2 Document organization and presentation 3
6.1.3 Document content analysis and processing 3
6.1.4 Long-term document preservation 4
6.2 General questions to consider 5
7 Representation of semantic metadata **5**
8 Metadata used in office documents **6**
9 Storing metadata in OOXML/ODF **8**
10 Associating semantic metadata with content in OOXML/ODF **10**
11 Application examples **11**
12 Future work on document format standards **14**
Bibliography **15**

© ISO/IEC 2026 – All rights reserved

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

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 or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take 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 and IEC 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 and https://patents.iec.ch. ISO and IEC 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. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1 *Information technology*, Subcommittee SC 34, *Document description and processing languages*.

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 and www.iec.ch/national-committees.

© ISO/IEC 2026 – All rights reserved

Introduction

Office document refers to the electronic document created, edited or processed by office applications (e.g. word processors, spreadsheets, presentation software), which adheres to structured markup standards for interoperability [see ISO/IEC 29500 (series)]. OOXML [see ISO/IEC 29500 (series)] and ODF (see ISO/IEC 26300) are widely used office document formats.

NOTE 1 There are other document types, such as PDF, EPUB and HTML, that incorporate semantic information to which the methods proposed in this document apply; however, they fall outside the scope of this document.

Documents are becoming more intelligent, which requires that their content be understood by both humans and machines. Adding tags to markup the semantic content of a document, or incorporating semantic information into a document through tags, is an effective way to achieve human- and machine-readable documents. In this document, the document semantic information expressed through tags is referred to as semantic metadata (abbreviated as metadata). The goal of semantic metadata is to provide a clear and accurate description of the content and context of a document, in an interoperable and accurate way. These semantic metadata can be defined in terms of industry vocabularies, resulting in normative metadata that can be shared across multiple application scenarios.

Traditionally, metadata were primarily used in web pages and fixed-format documents. This usage was due to the static nature of these documents, which were rarely modified and were suitable for long-term preservation, making metadata essential for locating key content within these documents. In contrast, support for metadata in editable documents is limited. It has been generally believed that since editable documents serve as intermediaries for editing, incorporating metadata is unnecessary.

Furthermore, there is a belief that metadata become unstable as documents undergo editing. However, recent findings indicate that there are billions of editable documents worldwide today, which are more numerous and contain richer semantic information compared to fixed-layout documents. These editable documents are valuable knowledge resources, making it sensible to add metadata to them. Additionally, the metadata annotated by the author during the document editing process can accurately reflect the semantics of the document.

To this end, this document proposes semantic support methods for office documents.

NOTE 2 Common office document components include word processing documents, spreadsheets and presentations. The methods presented in this document generally apply to spreadsheets and presentations as well. Due to space limitations, the examples provided in this document focus primarily on word processing documents.

NOTE 3 The internal objects in the composite documents based on OOXML and ODF sometimes contain their own metadata and support for such metadata is also beyond the scope of this document.

There are hundreds of industry vocabularies that exist today, each a potential source of metadata. Typically, it is up to different user communities to decide which vocabulary to use metadata from.

© ISO/IEC 2026 – All rights reserved

Information technology — Semantic metadata support in office documents

1 Scope

This document outlines:

— descriptions of use cases for identified application domains that benefit from additional semantic metadata support in office documents;

— descriptions of semantic metadata requirements from experts of identified application domains;

— specifications of possible technical strategies for associating domain-specific semantic metadata with office documents entities.

This document applies to users of office documents, for example, publishers, librarians, archivists, data scientists and other document users, as well as office software developers and service system integrators.

This document focuses on the methods of supporting various kinds of metadata, rather than on metadata itself.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

No terms and definitions are listed in this document.

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/

4 Document semantic metadata

ISO/IEC 11179-1:2023, 5.1 describes metadata as “data that define and describe other data.” ISO 19115-1:2014, 4.10 describes metadata as “information about a resource.” As for semantic metadata, W3C describes metadata as enriching data with formal semantics, using ontologies or vocabularies. [5] Dublin Core Metadata Initiative (DCMI) describes metadata as being grounded in a formal knowledge model, where terms are linked to defined concepts in a shared ontology.[6] In the document, which parts belong to semantic metadata and which do not depends on the requirements of the application. That is, if the application requires machines to interpret only certain parts of the document, then those parts can be represented as semantic metadata.

Metadata selection differs among communities, each with specific information requirements to be conveyed by means of metadata. Before the advent of the information society, metadata were widely used for retrieving books in libraries and identifying publications among publishers, with most of this metadata being manually created. Currently, most of the metadata attached to various data sources come from diverse origins, some of which are manually created or annotated, some inherited or translated from other documents and some

© ISO/IEC 2026 – All rights reserved