INTERNATIONAL STANDARD

ISO 22938

Second edition
2017-06

Document management — Electronic content/document management (CDM) data interchange format

Gestion de documents — Format d'échange de données pour la gestion de documents/du contenu électronique

Reference number
ISO 22938:2017(E)

COPYRIGHT PROTECTED DOCUMENT

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

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 on 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 the following URL: 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.*

This second edition cancels and replaces the first edition (ISO 22938:2008), which has been technically revised.

Introduction

This document specifies a consistent interchange format for data contained in electronic content/document management (CDM) systems, including documents, their associated resources, and retrieval index values that are stored in, or managed by, these technologies. Such a standard should facilitate the *exact* interchange of CDM data, i.e. the standard should not require that the data be irreversibly modified or packaged within a format that does not allow the reconstruction of the original data. Therefore, this document avoids choosing one particular data format and anointing it as the interchange standard for CDM. Rather, this document specifies a common markup format, based on the XML (eXtensible Markup Language), which encapsulates all forms of CDM data. A DTD (document type definition) describes the XML markup used for CDM data transfer. The XML format is a W3C (World Wide Web Consortium) standard, adopted in February 1998. XML is extensible, so that additional CDM formats may be easily specified by appropriately updating the DTD.

The purpose of this document is to define standards for information interchange in a way that benefits both the consumers and vendors of content/document management systems. Some possible benefits are as follows:

a) document information can be exported from one standard's compliant CDM system and afterwards imported to another standard's compliant CDM system;

b) disparate CDM systems within an enterprise (due to autonomous selection, replacement, or merger/acquisition) will be able to exchange or consolidate CDM information.

To this end, the standards are defined with the goal of striking a balance between being either too restrictive or too general. They should be broad enough to encompass all common CDM information types and all common uses of CDM systems, as well as ones that might be expected in the future. On the other hand, the standards should be restrictive enough so that CDM vendors do not have inordinate difficulty complying with the standards.

Document management — Electronic content/document management (CDM) data interchange format

1 Scope

This document defines the interchange format for content/document management (CDM) data and all associated resources.

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/IEC 29500-2:2012, *Information technology — Document description and processing languages — Office Open XML file formats — Part 2: Open Packaging Conventions*

Berners-Lee T., Fielding R. and Masinter L. RFC 3986: *Uniform Resource Identifier (URI): Generic Syntax*. The Internet Society, 2005 [viewed 2017-05-15]. Available from: http://www.ietf.org/rfc/rfc3986.txt

3 Terms and definitions

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

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

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

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

3.1
document
discreet unit or collection of content

3.2
rendition
electronic encoding of a *document* (3.1)

3.3
packages
collection containing *rendition(s)* (3.2) and related metadata

4 Abbreviated terms

CDM	content/document management
DTD	document type definition
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

5 XML-based data interchange format with OPC-based packaging

5.1 General

The document interchange format for electronic documents is an application of the XML. XML is an extensible, flexible, platform-independent format, and has been adopted by the W3C as a standard (officially a "recommendation" in W3C terminology).

The primary use of this document is to exchange data between diverse document management systems that do not already have an exchange methodology in place. This document is considered to be the foundational platform from which other XML-based exchange standards are developed, ensuring a common framework throughout the document management industry. The use of the ZIP-based Open Packaging Convention (OPC) to group the document interchange format XML, the content it describes, and related resources into a single standardized archive file allows the interchange of documents among CDM systems without the risk of the related parts becoming separated or out of sync.

5.2 Use of XML and OPC for content/document management data

5.2.1 Overview of OPC structure

The document interchange format for electronic documents utilizes the packaging format described in ISO/IEC 29500-2 ("OPC"). This is a ZIP-based format containing data files ("Parts") and metadata describing relationships between these parts.

5.2.2 Content/document management (CDM) — Specific OPC structure

A document of the format specified in this document which implements OPC packaging shall be an OPC package, as specified in ISO/IEC 29500-2. In addition to the requirements specified in ISO/IEC 29500-2, the package shall contain the OPC parts shown in Table 1.

Table 1 — OPC parts

Logical Name	Description	Content type
/metadata.xml	XML metadata content/document management structure (as specified in 5.2.4)	application/vnd.documentmanagement-metadata+xml
/_rels/.rels	XML representation of relationships between Parts included in the package as specified in 5.2.3.	application/vnd.openxmlformats-package.relationships+xml
Other parts	Renditions of content as specified in 5.2.5, f).	Appropriate to content

The content types of OPC Parts contained in the package shall be mapped to package data as defined in ISO/IEC 29500-2:2012, 10.1.2, which includes mapping of the content type of most types of data stored in the package to the data in a *Content Types stream* with the logical name *[Content_Type].xml* included in the package as specified in ISO/IEC 29500-2:2012, 10.2.6.

5.2.3 Content/document management (CDM) — Specific relationships

A document of the format specified in this document which implements the OPC packaging described in 5.2 shall include a *Relationships part* as specified in ISO/IEC 29500-2:2012, 9.3.1. The Relationships part shall include, at a minimum, a Relationship identifying the document interchange format XML, with the relationship type identified as http://placeholder_uri/documentmanagement-metadata.

5.2.4 Overview of XML structure

XML consists of markup and data. The markup consists of (usually paired) tags called elements, which may contain descriptive data called attributes. The data are the non-markup content residing between

element pairs. The elements can be nested, so that one element may contain sub-elements, which can in turn contain sub-sub-elements, etc.

This document defines the elements, element structure, and element attributes suitably, so that the various forms of CDM data, resources, index values, etc., can be clearly and unambiguously described and included as data. The model which describes this is an XML Schema. The precise schema is the essential content of this document.

5.2.5 Content/document management (CDM) — Specific XML structure

The XML structure of a CDM is described in an XML Schema Definition (XSD) below. The elements used in that XSD and their meanings are the following.

a) cdm_interchange

This is the root node of the interchange XML. It consists of an identifier to uniquely identify the interchange operation (interchange_id), the action that a CDM system should execute when processing the interchange XML (cdm_action), information about the creation of the interchange package (creator, vendor, creation_date, creation_time), and a set of document collections (cdm_collection). Creation_time should be a string in ISO 8601 format.

b) cdm_collection

This is the collection of documents contained in the package. It consists of a collection identifier (coll_id), a name (coll_name), a set of index values for the collection (index_set), and a set of documents (cdm_doc).

c) cdm_doc

This is the element representing a document contained in a document collection. It consists of a unique document identifier (doc_id), a document type (type), a document title (title), a set of index values for the document (index_set), and the content that comprises the actual document data (doc_content). It shall contain an index_set of metadata and a doc_content element, which contains the method used to encode or provide explicit external reference to the data.

d) index_set

This element contains metadata related to a document or document collection. It consists of a set of fields (index_field) or a record (index_record). Index_set shall contain at least one index_field for each cdm_doc, with the attributes of index_name, index_description and index_content.

e) index_field

This element references index_name, index_description, and index_content elements. Any index_set element shall contain at least one index_field element.

f) index_record

This element organizes multiple index_field entries into a logical group.

g) doc_content

This element defines the document contents being transmitted as part of the cdm_interchange operation. Each doc_content shall contain one or more renditions.

h) rendition

This element defines the renditions, if any, and their attributes. Rendition includes the document content (content) and resources needed to use the content (rsrc_data) elements. These elements are used to provide a mechanism to define the access_method, encoding and compression for each rendition. The access_method is required, and the encoding and compression attributes are optional. Supported values of access_method include Base64, URI, and MIME.

When using OPC to package the CDM data XML, content, and rsrc_data, the access_method for renditions included in the OPC package shall be URI, and the encoding shall be set to the relative URI (as specified in RFC 3986:2005, 4.2) of the content or rsrc_data within the OPC package as specified in ISO/IEC 29500-2:2012, A.3. For such renditions, the compression attribute should not be included by producing applications, and may be ignored by consuming applications.

i) rsrc_data
This element encloses CDM resource data within each rendition. Examples of resource data are bitmaps and fonts that are needed to render the contained document. It provides information defining the method to be used to access the resource (access_method), the type of the resource file (filetype), the encoding used to store the resource (encoding), and any method used to compress the resource in the package (compression). Examples of filetype could be TIFF, PDF, PDF/A, JPEG, JPEG2000 and RTF. It is recommended to use only IANA-registered mimetypes.

j) annotations
This element encloses the annotation-related information for a rendition. The annotation is expressed as a stream of knowledge that would be defined by the vendor. Some vendors have highlight information, while others might have blobs, bitmaps or data files. The knowledge content of the annotation would be vendor-specific. It provides information defining the method to be used to access the annotations (access_method), the type of the annotation file (filetype), the encoding used to store the annotation (encoding), and any method used to compress the annotation in the package (compression).

k) content
This element provides information defining the method to be used to access the content (access_method), the type of the content file (filetype), the encoding used to store the content (encoding), and any method used to compress the content in the package (compression). Encoding is the base64 representation of the document rendition data based on the value of the access_method attribute.

l) index_name

This element provides for a name to be associated with the index element record attributes.

m) record attributes

This element provides a name and description for the index record.

n) index_description
This element allows a description containing unconstrained text to be associated with the index for documentation of information purposes.

o) index_content
This element contains the value for the index.
The schema used for CDM data interchange is below. Schemas for other XML parts included in CDM packages using OPC packaging are specified in ISO\IEC 29500-2:2012, Annex D.
This schema is intended to provide the framework/mechanism to exchange data between diverse systems in the absence of a specific schema. Organizations that do not have an implementation-specific model of this schema shall use this model for specific information exchange between diverse document management systems.
To create an application-specific instance of this schema, users shall use this schema as the framework, or model, ensuring the appropriate level of information exchange between diverse document management systems.


```
<?xml version="1.0" encoding="UTF-8"?>
<!-- The ISO 22938 CDM Data Interchange DTD -->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="cdm_interchange">
    <xs:complexType>
      <xs:sequence>
       <xs:element maxOccurs="unbounded" ref="cdm_collection"/>
      </xs:sequence>
      <xs:attribute name="cdm_action">
        <xs:simpleType>
          <xs:restriction base="xs:token">
             <xs:enumeration value="store"/>
            <xs:enumeration value="get"/>
            <xs:enumeration value="verify"/>
            <xs:enumeration value="update"/>
            <xs:enumeration value="delete"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:attribute>
      <xs:attribute name="interchange_id" use="required" type="xs:ID"/>
      <xs:attribute name="creator" default="creator"/>
      <xs:attribute name="vendor" default="vendor"/>
      <xs:attribute name="creation_date"/>
      <xs:attribute name="creation_time"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="cdm_collection">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="index_set"/>
        <xs:element maxOccurs="unbounded" ref="cdm_doc"/>
      </xs:sequence>
      <xs:attribute name="coll_id" use="required"/>
      <xs:attribute name="coll_name" use="required"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="cdm_doc">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="index_set"/>
        <xs:element ref="doc_content"/>
      </xs:sequence>
      <xs:attribute name="doc_id" use="required"/>
      <xs:attribute name="type" use="required"/>
      <xs:attribute name="title" use="required"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="index_set">
    <xs:complexType>
      <xs:choice minOccurs="1" maxOccurs="unbounded">
        <xs:element minOccurs="1" ref="index_field"/>
        <xs:element ref="index_record"/>
      </xs:choice>
    </xs:complexType>
  </xs:element>
  <xs:element name="index_field" >
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="index_name"/>
        <xs:element ref="index_description"/>
        <xs:element ref="index_content"/>
      </xs:sequence>
      <xs:attribute name="scheme"/>
      <xs:attribute name="datatype"/>
      <xs:attribute name="language"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="index_name" type="xs:string"/>
  <xs:element name="index_description" type="xs:string"/>
  <xs:element name="index_content" type="xs:string"/>
  <xs:element name="index_record">
```