INTERNATIONAL STANDARD

ISO 22938

First edition
2008-12-01

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 22938:2008(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web 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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

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.

ISO 22938 was prepared by Technical Committee ISO/TC 171, *Document management applications*, Subcommittee SC 2, *Application issues*.

Introduction

This International Standard 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 International Standard avoids choosing one particular data format and anointing it as the interchange standard for CDM. Rather, this International Standard 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 International Standard 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 both exported from and imported to one standards-compliant CDM system to another;

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 International Standard defines the interchange of content/document management (CDM) data and all associated resources.

2 Normative references

The following referenced documents are indispensable for the application of this International Standard. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 12651, *Electronic imaging — Vocabulary*

3 Terms and definitions

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

3.1
document
XML stream containing information content and related metadata

3.2
rendition
electronic encoding of a page of content

4 Symbols and abbreviated terms

CDM content/document management

DTD document type definition

W3C World Wide Web Consortium

WWW World Wide Web

XML eXtensible Markup Language

5 XML-based data interchange format

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 International Standard is to exchange data between diverse document management systems that do not already have an exchange methodology in place. This International Standard 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.

5.2 Use of XML for content/document management data

5.2.1 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. Data is 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 International Standard 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 DTD. The precise DTD is the essential content of this International Standard.

5.2.2 Content/document management (CDM) – specific XML structure – the DTD

Figures 1 and 2 describe the high-level model to create the DTD. The elements and their meanings are:

a) cdm_interchange
This is the name of the XML application, or DTD.

b) cdm_collection
This is the collection of documents contained in the XML. It consists of a name, a set of index values for the collection and a set of documents.

c) 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.

d) index_record
This element organizes multiple index_field entries into a logical group.

e) 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.

f) rendition
This element defines the renditions, if any, and their attributes. Rendition includes the content and 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 access_method include Base64, URI, and MIME.

g) 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.

h) 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.

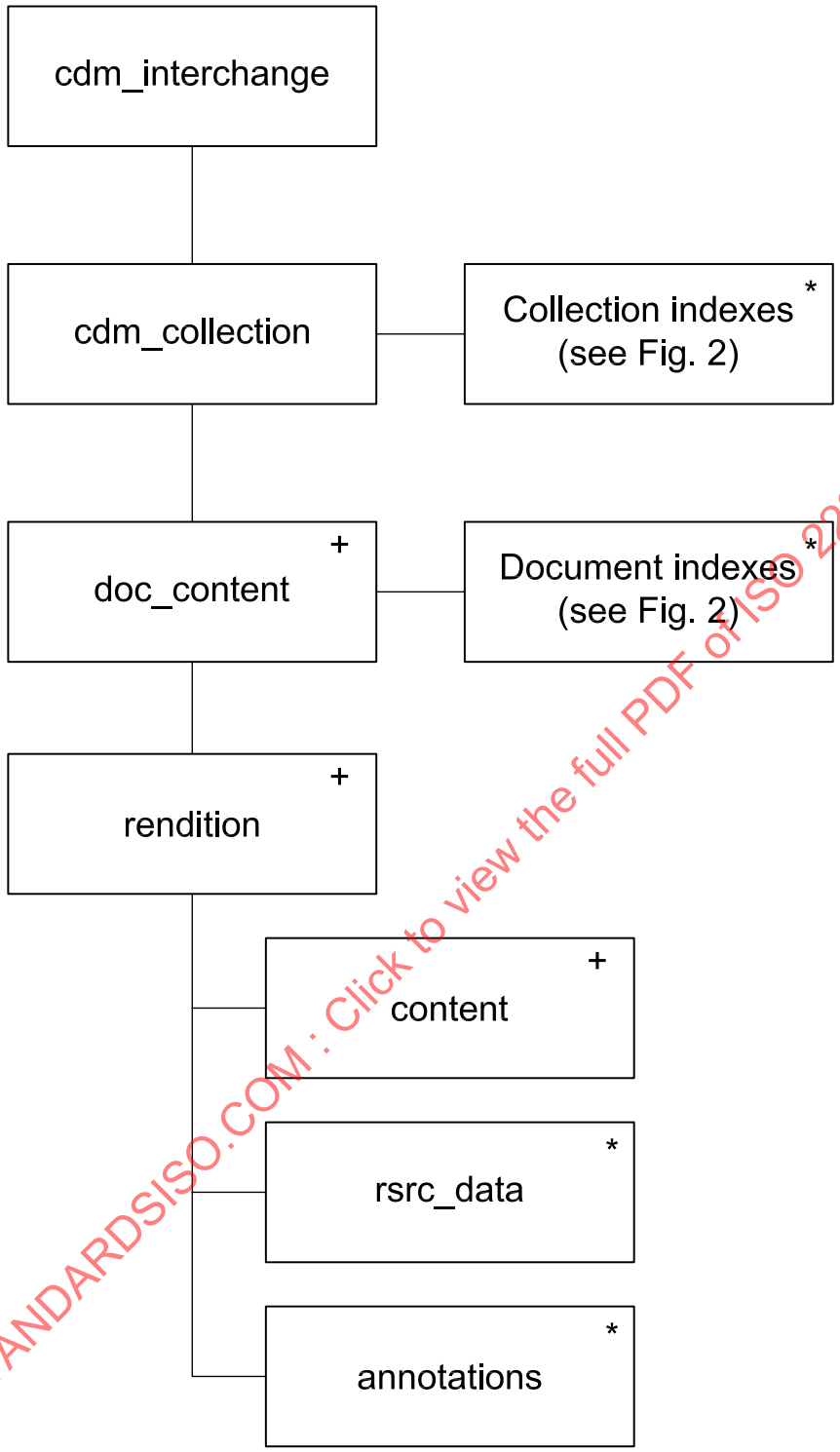
i) content
This element provides information related to the required access_method, form of data encoding and compression technique.

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

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

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

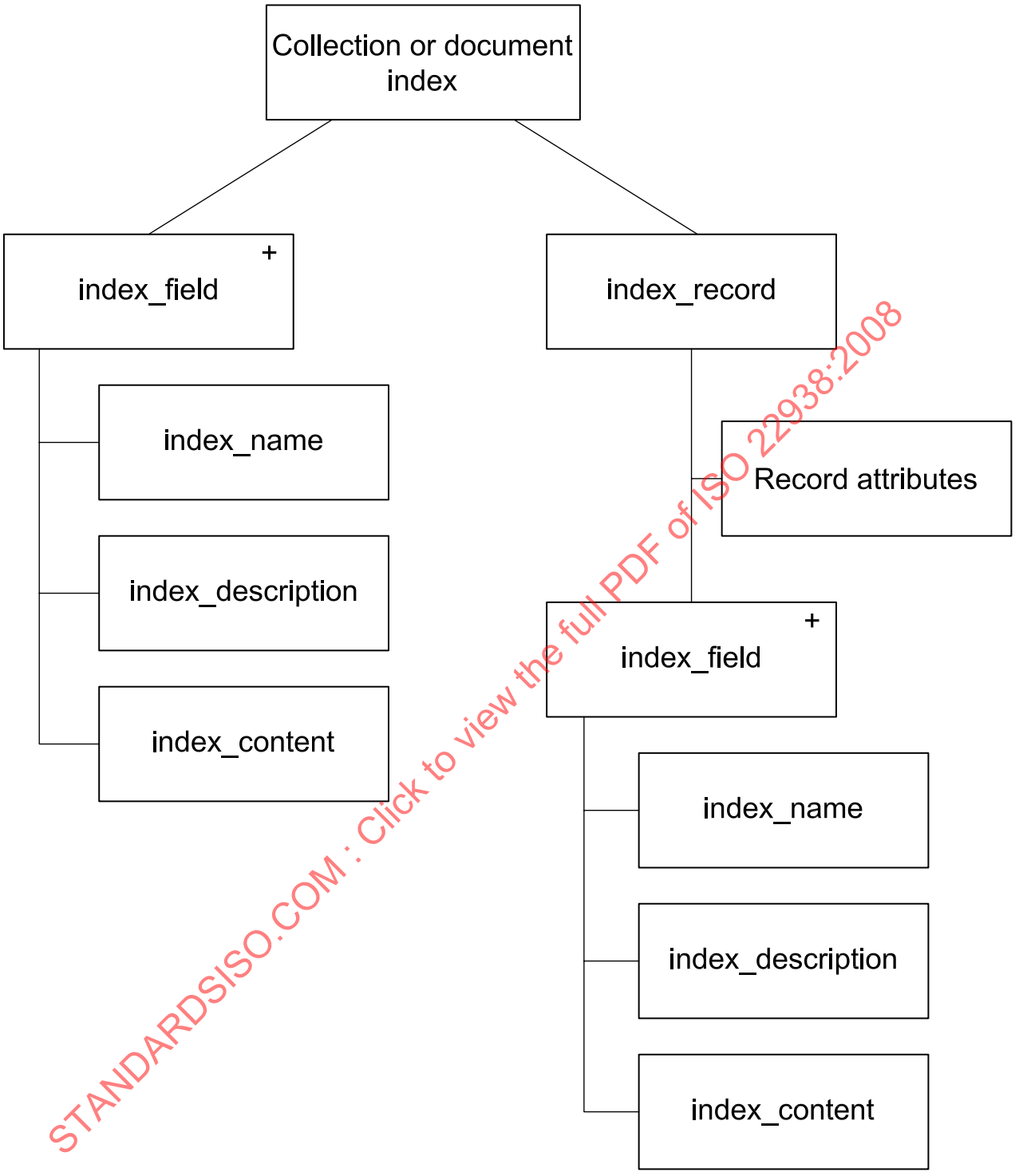
m) index_content
This element contains the value for the index.



* There may be 0 or more occurrences of the item.

\+ There may be 1 or more occurrences of the item.

Figure 1 — High-level XML design for CDM data interchange



\+ There may be 1 or more occurrences of the item.

Figure 2 — High-level XML design continued

5.3 The explicit DTD for CDM data interchange

Table 1 presents the data interchange structure presented as a DTD and comments to allow non-readers of XML to understand the model.

The following DTD for ISO 22938 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 DTD shall use this model for specific information exchange between diverse document management systems.

To create an application-specific instance of this DTD, users shall use the ISO 22938 DTD as the framework, or model, ensuring the appropriate level of information exchange between diverse document management systems.

Table 1 — CDM interchange DTD in XML

XML	Comments
<?xml version="1.0" encoding="UTF-8"?>	These are XML notations and comments.
<!-- The ISO 22938 CDM Data Interchange DTD -->	
<!ELEMENT cdm_interchange (cdm_collection)+>	The cdm_interchange element shall contain CDM data, and may also contain descriptions of index fields and resource data.
<!ATTLIST cdm_interchange	
cdm_action (store | get | verify | delete) #IMPLIED interchange_id ID #REQUIRED	The XML document may contain "cdm_action" (action to be taken on the document), "creator" (who wrote the document), "vendor" (whose CDM system) and/or creation date/time information in ISO 8601 format.
creator CDATA "creator"	
vendor CDATA "vendor"	"Vendor" is the manufacturer of the system that generated this XML stream.
creation_date CDATA #IMPLIED	Metadata that specifies the date of creation of the content that is described in this XML stream.
creation_time CDATA #IMPLIED>	Time of creation, if known.
<!ELEMENT cdm_collection (index_set, cdm_doc+)> <!ATTLIST cdm_collection coll_id CDATA #REQUIRED coll_name CDATA #REQUIRED>	
<!ELEMENT cdm_doc (index_set, doc_content)>	Each cdm_doc element shall be given a unique ID for reference. The cdm_doc element may contain resources. 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.
<!ATTLIST cdm_doc	
doc_id CDATA #REQUIRED	"doc_id" is a unique identifier within or traceable to the document originator.
type CDATA #REQUIRED title CDATA #REQUIRED>	"Type" is the classification of the document. "Title" is the name by which the document is identified.

Table 1 (*continued*)

XML	Comments
<!ELEMENT index_set (index_field | 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.
<!ELEMENT index_field (index_name, index_description, index_content)>	
<!ATTLIST index_field	
scheme CDATA #IMPLIED datatype CDATA #IMPLIED	
language CDATA #IMPLIED>	
<!ELEMENT index_name (#PCDATA)>	Name of the index for the cdm_doc.
<!ELEMENT index_description (#PCDATA)>	Unconstrained information describing the index field.
<!ELEMENT index_content (#PCDATA)>	Content of the index for the cdm_doc.
<!ELEMENT index_record (record_name, record_description, index_field+)> <!ATTLIST index_record record CDATA #REQUIRED> <!ELEMENT record_name (#PCDATA)> <!ELEMENT record_description (#PCDATA)>	
<!ELEMENT doc_content (rendition)+>	Each doc_content may contain 1 or more rendition elements.
<!ELEMENT rendition (content+, rsrc_data*, annotations*)>	Each rendition element shall contain 1 or more content elements and 0 or more rsrc_data elements. There shall be a 1 to 1 relationship between the content and rsrc_data elements for each rendition.
<!ELEMENT content (#PCDATA)>	
<!ATTLIST content	This attribute provides information related to the required access_method, form of data encoding and compression technique.
access_method (URI|base64|MIME) #REQUIRED filetype CDATA #REQUIRED	Examples of filetype could be “TIFF”, “JPEG”, “JPEG2000” and “RTF”. It is recommended to use only IANA registered mimetypes.
encoding CDATA #REQUIRED	Encoding is the base64 representation of the document data.
compression CDATA #IMPLIED>	Compression technique applied to the encoding data. Choices are “NONE”, “ZIP”.
<!ELEMENT rsrc_data (#PCDATA)>	
<!ATTLIST rsrc_data	
type NMTOKEN #REQUIRED	
location CDATA #IMPLIED> <!ELEMENT annotations (#PCDATA)> <!ATTLIST annotations access_method (URI|base64|MIME) #REQUIRED filetype CDATA #REQUIRED encoding CDATA #IMPLIED compression CDATA "None">	

5.4 Representing CDM data using the DTD – an example

The following example shows how the CDM interchange DTD could be used to exchange multiple CDM documents.

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE cdm_interchange SYSTEM "22938.dtd">
<cdm_interchange creation_time="18:00:00" vendor="ECM Vendor A" cdm_action="store"
interchange_id="ID001"
      creator="System Admin" creation_date="2007-01-01">
      <cdm_collection coll_id="Feb-2007" coll_name="Testing Supplies">
            <index_set>
                  <index_field datatype="Alphanumeric" language="en">
                        <index_name>Supplier</index_name>
                        <index_description>Supplier Name</index_description>
                        <index_content>Testing Supplies</index_content>
                  </index_field>
            </index_set>
            <cdm_doc doc_id="DOCID000000" type="Vendor Invoices"
        title="Invoice 2007-101 from Testing Supplies">
                  <index_set>
                        <index_field datatype="alphanumeric" language="En">
                              <index_name>Invoice Number</index_name>
                              <index_description>Vendor Invoice Number</index_description>
                              <index_content>2007-101</index_content>
                        </index_field>
                        <index_field datatype="currency" language="En">
                              <index_name>Invoice amount</index_name>
                              <index_description>Total amount of Invoice</index_description>
                              <index_content>$1021.55</index_content>
                        </index_field>
                  </index_set>
                  <doc_content>
                        <rendition>
                              <content filetype="RTF"
            encoding="http://invoices/data/invoice101.rtf"
            access_method="URI"
                                    compression="None" />
                        </rendition>
                        <rendition>
                              <content filetype="TIFF"
          encoding="http://invoices/data/invoice101.tif"
              access_method="URI"
                                    compression="None" />
                              <annotations filetype="TIFF"
            encoding="http://invoices/data/invoice101_note.tif"
            access_method="URI" compression="None" />
                        </rendition>
                  </doc_content>
            </cdm_doc>
            <cdm_doc doc_id="DOCID000001" type="Vendor Invoices"
          title="Invoice 2007-102 from Testing Supplies">
                  <index_set>
                        <index_field datatype="alphanumeric" language="En">
                              <index_name>Invoice Number</index_name>
                              <index_description>Vendor Invoice Number</index_description>
                              <index_content>2007-102</index_content>
                        </index_field>
                        <index_field datatype="currency" language="En">
                              <index_name>Invoice amount</index_name>
```