BS ISO 24622-2:2019

BSI Standards Publication

Language resource management — Component metadata infrasctructure (CMDI)

Part 2: Component metadata specification language

Documents
Document Sharing Hub

Full document details: BS ISO 24622-2:2019

National foreword

This British Standard is the UK implementation of ISO 24622-2:2019.

The UK participation in its preparation was entrusted to Technical Committee TS/1, Terminology.

A list of organizations represented on this committee can be obtained on request to its secretary.

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

© The British Standards Institution 2019
Published by BSI Standards Limited 2019

ISBN 978 0 580 93087 4

ICS 01.140.20

Compliance with a British Standard cannot confer immunity from legal obligations.

This British Standard was published under the authority of the Standards Policy and Strategy Committee on 31 July 2019.

Amendments/corrigenda issued since publication

Date	Text affected

Documents
Document Sharing Hub

Full document details: BS ISO 24622-2:2019

INTERNATIONAL STANDARD

ISO 24622-2

First edition
2019-07

Language resource management — Component metadata infrasctructure (CMDI) —

Part 2: Component metadata specification language

Gestion des ressources linguistiques — Composante infrastructure de métadonnées (CMDI) —

Partie 2: Composante linguistique spécifique aux métadonnées

Reference number
ISO 24622-2:2019(E)

Documents
Document Sharing Hub

© ISO 2019

Full document details: BS ISO 24622-2:2019

COPYRIGHT PROTECTED DOCUMENT

© ISO 2019

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
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

© ISO 2019 – All rights reserved

Documents
Document Sharing Hub

Full document details: BS ISO 24622-2:2019

Contents

Page

Foreword ... **iv**

Introduction ... **v**

1 Scope ... **1**

2 Normative references ... **1**

3 Terms and definitions ... **1**
3.1 General terms ... 1
3.2 CMDI ... 3
3.3 XML ... 5

4 Notational and XML namespace conventions ... **7**

5 Structure of CMDI instances ... **8**
5.1 General structure ... 8
5.2 The main structure ... 9
5.3 The `<Header>` element ... 10
5.4 The `<Resources>` element ... 11
5.4.1 General structure of the <Resources> element ... 11
5.4.2 The list of resource proxies ... 11
5.4.3 The list of journal files ... 12
5.4.4 The list of relations between resource files ... 13
5.5 The `<IsPartOfList>` element ... 15
5.6 The CMD components ... 15

6 CCSL (CMDI Component Specification Language) ... **17**
6.1 General structure of the CCSL ... 17
6.2 CCSL header ... 19
6.3 CMD specification ... 20
6.4 Definition of CMD elements ... 21
6.5 CMD attribute definition ... 23
6.6 Value schemes for CMD elements and CMD attributes ... 24
6.7 Cue attributes ... 26

7 CMD ... **27**
7.1 Transformation of CCSL into a CMD profile schema definition ... 27
7.2 General properties of the CMD profile schema definition ... 27
7.3 Interpretation of CMD specifications in the CCSL ... 27
7.3.1 General structure of CMD specifications ... 27
7.3.2 Document structure prescribed by the CMD profile schema ... 28
7.4 Interpretation of CMD element definitions in the CCSL ... 28
7.5 Interpretation of CMD attribute definitions in the CCSL ... 29
7.6 Content model for CMD elements and CMD attributes in the schema definition ... 30

Bibliography ... **31**

© ISO 2019 – All rights reserved

Documents
Document Sharing Hub

Full document details: BS ISO 24622-2:2019

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 37, *Language and terminology*, Subcommittee SC 4, *Language resource management*.

A list of all parts in the ISO 24622 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.

© ISO 2019 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 24622-2:2019

Introduction

Many researchers, from the humanities and other domains, have a strong need to study resources in close detail. Nowadays more and more of these resources are available online. To be able to find these resources, they are described with metadata. These component metadata (CMD) instances are collected and made available via central catalogues. Often, resource providers want to include specific properties of a resource in their metadata to provide all relevant descriptions for a specific type of resource. The purpose of catalogues tends to be more generic and addresses a broader target audience. It is hard to strike the balance between these two ends of the spectrum with one metadata schema, and mismatches can negatively impact the quality of metadata provided. The goal of the component metadata infrastructure (CMDI) is to provide a flexible mechanism to build resource specific metadata schemas out of shared components and semantics[14][15].

In CMDI the metadata lifecycle starts with the need of a metadata modeller to create a dedicated metadata profile for a specific type of resource. Modellers can browse and search a registry for components and profiles that are suitable or come close to meeting their requirements. A component groups together metadata elements that belong together and can potentially be reused in a different context. Components can also group other components. Existing component registries, e.g., the CLARIN (common language resources and technology infrastructure) Component Registry[16], might already contain any number of components. These can be reused as they are, or be adapted by modifying, adding or removing some metadata elements and/or components. Also completely new components can be created to model the unique aspects of the resources under consideration. All the needed components are combined into one profile specific for the type of resources. Any component, element and value in such a profile may be linked to a semantic description — a *concept* — to make their meaning explicit[21]. These semantic descriptions can be stored in a semantic registry, e.g., the CLARIN Concept Registry[17]. In the end metadata creators can create records for specific resources that comply with the profile relevant for the resource type, and these records can be provided to local and global catalogues[22].

CMDI has originally been developed in the context of the European CLARIN infrastructure initiative with input from other initiatives and experts. Already in its preparatory phase, which started in 2007, the infrastructure needed flexibility in the metadata domain as it was confronted with many types of resources that had to be accurately described. For Version 1.0 a toolkit[20] was created, consisting of the XML schemas and XSLT stylesheets to validate and transform components, profiles and records. Version 1.1 included some small changes and has seen small incremental backward compatible advances since 2011. This version has been in use, new developments and the development of this document resulted in Version 1.2[18]. Also CMDI has seen a growing number of tools and infrastructure systems that deal with its records and components and rely on its shared syntax and semantics.

In ISO 24622-1, the component metadata model has been standardized. This document is compliant with ISO 24622-1, and also extends and constrains it at various places (see also the red parts in the UML class diagram in Figure 1):

— support for attributes on both components and elements is added,

— a profile is limited to one root component, and

— an element always belongs to a specific component.

© ISO 2019 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 24622-2:2019

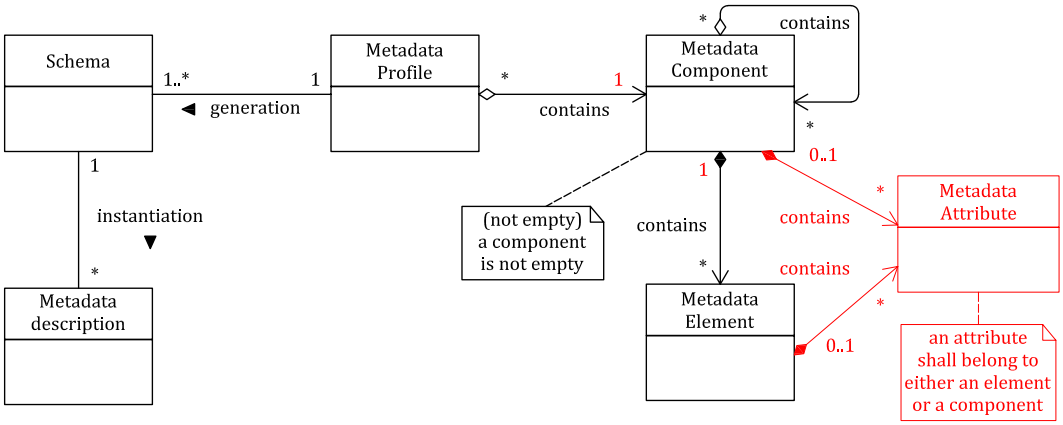

Figure 1 — Component metadata model and its extensions

© ISO 2019 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 24622-2:2019

Language resource management — Component metadata infrasctructure (CMDI) —

Part 2:
Component metadata specification language

IMPORTANT — The electronic file of this document contains colours which are considered to be useful for the correct understanding of the document. Users should therefore consider printing this document using a colour printer.

1 Scope

The component metadata lifecycle needs a comprehensive infrastructure with systems that cooperate well together. To enable this level of cooperation this document provides in depth descriptions and definitions of what CMDI records, components and their representations in XML look like.

This document describes these XML representations, which enable the flexible construction of interoperable metadata schemas suitable for, but not limited to, describing language resources. The metadata schemas based on these representations can be used to describe resources at different levels of granularity (e.g. descriptions on the collection level or on the level of individual resources).

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 terminological databases for use in standardization at the following addresses:

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

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

3.1 General terms

3.1.1
concept
unit of knowledge created by a unique combination of characteristics

[SOURCE: ISO 1087:—[1)], 3.2.3, modified — Note 1 to entry and Note 2 to entry have been deleted.]

3.1.2
concept link
reference from a *CMD profile* (3.2.11), *CMD component* (3.2.3), *CMD element* (3.2.5), *CMD attribute* (3.2.2) or a value in a *controlled_vocabulary* (3.1.4) to an entry in a *semantic registry* (3.1.11) via a *Uniform Resource Identifier* (3.1.13)

Note 1 to entry: Typically a concept link is provided as a *persistent identifier* (3.1.9).

1) Revision of ISO 1087:2000 under preparation. Stage at the time of publication: ISO/FDIS 1087:2019.

© ISO 2019 – All rights reserved

AVA Documents Document Sharing Hub
Full document details: BS ISO 24622-2:2019

3.1.3
concept registry
semantic registry (3.1.11) maintaining *concepts* (3.1.1)

EXAMPLE The CLARIN Concept Registry[17] as used in the CLARIN infrastructure.

3.1.4
controlled vocabulary
closed/open vocabulary
set of values that can be used either to constrain the set of permissible values or to provide suggestions for applicable values in a given context

3.1.5
data category
class of data items that are closely related from a formal or semantic point of view

EXAMPLE /part of speech/, /subject field/, /definition/.

Note 1 to entry: A data category can be viewed as a generalization of the notion of a field in a database.

[SOURCE: ISO 30042:2019, 3.8, modified — Note 2 to entry has been deleted.]

3.1.6
language tag
textual code used to assist in identifying languages in every mode of communication

Note 1 to entry: This includes constructed and artificial languages but excludes languages not intended primarily for human communication, for example in spoken, written, signed, or otherwise signaled, communication (see IETF BCP 47[6]).

Note 2 to entry: Language tags may be used to assist in the identification of a language in every mode of communication, for example in spoken, written, signed, or otherwise signaled, communication.

3.1.7
media type
MIME type
media type specification used originally for textual, non-textual, multi-part message bodies of emails and which provides technical format information on data

Note 1 to entry: For the purposes of this document, it is as described in IETF RFC 6838[8].

3.1.8
metadata
resource (3.1.10) that is a description of another resource, usually given as a set of properties in the form of attribute-value pairs

Note 1 to entry: This description can contain information about the resource, aspects or parts of the resource and/or artefacts and actors connected to the resource.

3.1.9
persistent identifier
PID
unique identifier that ensures permanent access for a digital object by providing access to it independently of its physical location or current ownership

Note 1 to entry: Unique in this context means that the PID will not be issued again for other *resources* (3.1.10). However, the same PID can reference different representations or incarnations of the resource at the discretion of the resource provider.

[SOURCE: ISO 24619:2011, 3.2.4]

© ISO 2019 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 24622-2:2019

3.1.10
resource
entity, possibly digitally accessible, that can be described in terms of its content and technical properties, referenced by a *Uniform Resource Identifier* (3.1.13)

3.1.11
semantic registry
directory of (authoritative) definitions of *term* (3.1.12), *concept* (3.1.1) or *data category* (3.1.5), or the system maintaining it

Note 1 to entry: These registries generally also provide *persistent identifiers* (3.1.9) for their entries.

3.1.12
term
designation that represents a general *concept* (3.1.1) in a specific domain or subject

EXAMPLE "planet", "tower", "pen", "numeral", "number", "square root", "logarithm", "unit of measurement", "base of a logarithm", "chemical element", "chemical compound", "HP Laserjet 1100", "Nobel Prize in Physics".

Note 1 to entry: Terms may be partly or wholly verbal.

Note 2 to entry: Terms can include letters and letter symbols, numerals, mathematical symbols, typographical signs and syntactic signs (e.g. punctuation marks, such as hyphens, parentheses, square brackets and other connectors or delimiters), sometimes in character styles (i.e. fonts and bold, italic, bold italic, or other style conventions) governed by domain-, subject-, or language-specific conventions.

[SOURCE: ISO 1087:—, 3.4.2]

3.1.13
Uniform Resource Identifier
URI
sequence of characters that identifies a *resource* (3.1.10)

Note 1 to entry: IETC RFC 3986[7] defines the generic URI syntax and a process for resolving URI references that might be in relative form, along with guidelines and security considerations for the use of URIs on the Internet.

3.2 CMDI

3.2.1
CCSL
CMDI component specification language
XML (3.3.4) based language for describing a *CMD component* (3.2.3) and a *CMD profile* (3.2.11) according to the *CMD model* (3.2.10)

3.2.2
CMD attribute
unit within a *CMD element* (3.2.5) that describes the level at which properties of a CMD element can be provided by means of *value scheme* (3.2.20) constrained atomic values

3.2.3
CMD component
component
reusable, structured template for the description of (an aspect of) a *resource* (3.1.10), defined by means of a *CMD specification* (3.2.14) document with the potential of including other CMD components, either through reference or inline definition

3.2.4
CMD component registry
component registry
service where a *CMD specification* (3.2.14) can be registered and accessed

© ISO 2019 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 24622-2:2019

3.2.5
CMD element
element definition
unit within a *CMD component* (3.2.3) that describes the level of the *CMD instance* (3.2.6) that can carry atomic values governed by a *value scheme* (3.2.20), and does not contain further levels except for that of the *CMD attribute* (3.2.2)

3.2.6
CMD instance
metadata instance
CMDI file
CMDI instance
metadata record
CMD record
file that conforms to the general CMD instance structure as described in this document and, at the *CMD instance payload* (3.2.9) level, follows the specific structure defined by the *CMD profile* (3.2.11) it relates to

3.2.7
CMD instance envelope
section of a *CMD instance* (3.2.6) which is structured uniformly for all instances and contains the *CMD instance header* (3.2.8) and the list of *resource proxies* (3.2.18) which may be referenced from the *CMD instance payload* (3.2.9) section

3.2.8
CMD instance header
section of a *CMD instance* (3.2.6) marked as 'header', providing information on that *CMD instance* as such, not the *resource* (3.1.10) that is described by the metadata file

3.2.9
CMD instance payload
section of a *CMD instance* (3.2.6) that follows the structure defined by the *CMD profile* (3.2.11) it references and contains the description of the *resource* (3.1.10) to which that CMD instance relates

3.2.10
CMD model
component metadata model
metadata model that is based on *CMD components* (3.2.3)

Note 1 to entry: For the purposes of this document, it is as specified in ISO 24622-1.

3.2.11
CMD profile
profile
structured template for the description of a class of *resources* (3.1.10) providing the complete structure for a *CMD instance payload* (3.2.9) by means of a hierarchy of *CMD components* (3.2.3)

3.2.12
CMD profile schema
schema definition by which the correctness of a *CMD instance* (3.2.6) with respect to the *CMD profile* (3.2.11) it pertains to can be evaluated

Note 1 to entry: The CMD profile schema may be expressed as *XML Schema* (3.3.11) but also in other XML schema languages.

3.2.13
CMD root component
CMD component (3.2.3) that is defined at the highest level within a *CMD profile* (3.2.11) that may have one or more child *CMD components* (3.2.3) but no siblings

Note 1 to entry: In the *CMD instance payload* (3.2.9), it is instantiated exactly once.

© ISO 2019 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 24622-2:2019

3.2.14
CMD specification
component specification
component definition
profile specification
profile definition
representation of a *CMD component* (3.2.3) or *CMD profile* (3.2.11), expressed using the constructs of the *CCSL* (3.2.1)

3.2.15
CMD specification header
component header
profile header
section of a *CMD specification* (3.2.14) marked as 'header', providing information on that CMD specification as such that is not part of the defined structure

3.2.16
CMDI
component metadata infrastructure
metadata description framework consisting of the *CMD model* (3.2.10) and infrastructure to process instances of parts of the model

3.2.17
inline CMD component
CMD component (3.2.3) that is created and stored within another CMD component and cannot be addressed from other CMD components

3.2.18
resource proxy
CMD resource reference
representation of a *resource* (3.1.10) within a *CMD instance* (3.2.6) containing a *Uniform Resource Identifier* (3.1.13) as a reference to the resource itself and an indication of its nature

3.2.19
resource proxy reference
reference from any point within the *CMD instance payload* (3.2.9) to any of the *resource proxy* (3.2.18) elements

3.2.20
value scheme
set of constraints governing the range of values allowed for a specific *CMD element* (3.2.5) or *CMD attribute* (3.2.2) in a *CMD instance* (3.2.6), expressed in terms of an *XML Schema datatype* (3.3.12), *controlled vocabulary* (3.1.4), or *regular expression* (3.3.3)

3.3 XML

3.3.1
foreign attribute
XML attribute (3.3.5) defined in a *namespace* (3.3.2) other than those declared in *CMDI* (3.2.16), to be included in a *CMD instance* (3.2.6) as additional information targeted to specific receivers or applications

3.3.2
XML namespace
namespace
method for qualifying element and attribute names used in XML

Note 1 to entry: For the purposes of this document, it is as described in W3C XML Namespaces[10].

© ISO 2019 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 24622-2:2019