SLOVENSKI STANDARD
oSIST ISO/DIS 24622-2:2019

01-september-2019

Upravljanje z jezikovnimi viri - Infrastruktura komponentnih metapodatkov (CMDI) - 2. del: Poseben jezik komponentnih metapodatkov

Language resource management -- Component metadata infrasctructure (CMDI) -- Part 2: The component metadata specific language

Gestion des ressources linguistiques -- Composante infrastructure de métadonnées (CMDI) -- Partie 2: Composante linguistique spécifique aux métadonnées

Ta slovenski standard je istoveten z: **ISO/DIS 24622-2**

ICS:

01.140.20	Informacijske vede	Information sciences
35.060	Jeziki, ki se uporabljajo v informacijski tehniki in tehnologiji	Languages used in information technology

oSIST ISO/DIS 24622-2:2019 **en,fr,de**

DRAFT INTERNATIONAL STANDARD

ISO/DIS 24622-2

ISO/TC **37**/SC **4**

Secretariat: **KATS**

Voting begins on:
2018-08-10

Voting terminates on:
2018-11-02

Language resource management — Component metadata infrasctructure (CMDI) —

Part 2:
The component metadata specific language

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

Partie 2: Composante linguistique spécifique aux métadonnées

ICS: 01.140.20

This document is circulated as received from the committee secretariat.

Reference number
ISO/DIS 24622-2:2018(E)

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

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 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.

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 metadata records 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 address 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 (Broeder *et al*, 2010 and Broeder *et al, 2012*).

In CMDI the metadata lifecycle starts with the need of a metadata modeller to create a dedicated metadata profile for a specific type of resources. 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. A component registry, e.g., the *CLARIN Component Registry*, 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 (Durco & Windhouwer, 2013). These semantic descriptions can be stored in a semantic registry, e.g., the *CLARIN Concept Registry*. 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 (Van Uytvanck et *al*, 2012).

History

CMDI has been developed in the context of the European CLARIN infrastructure 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 the CMDI toolkit 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 throughout CLARIN's construction phase. 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.

Language resource management — Component metadata infrastructure (CMDI) — Part 2: The component metadata specification language

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.

The scope of this document is to describe 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).

In ISO 24622-1:2015 the component metadata model has been standardized. This document is compliant with ISO 24622-1:2015, and also extends and constrains it at various places (see also the red parts in the UML class diagram below):

— 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.

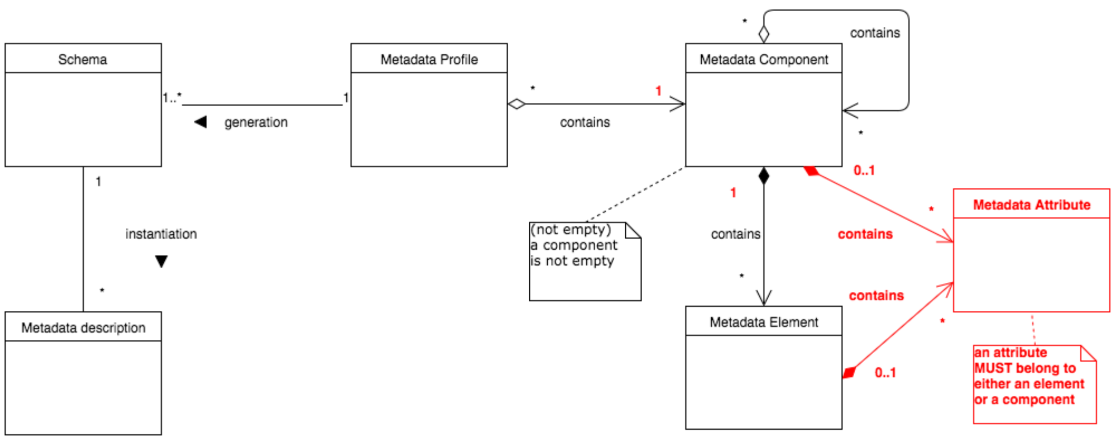

Figure 1 — Component metadata model and its extensions

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 24622-1:2015, *Language resource management— Component metadata infrastructure (CMDI) — Part 1: The component metadata model*

IETF BCP 47, *Tags for Identifying Languages, September 2009*, https://tools.ietf.org/rfc/bcp/bcp47.txt

IETF RFC 2119, *Key words for use in RFCs to Indicate Requirement Levels*, March 1997, https://www.ietf.org/rfc/rfc2119.txt

IETF RFC 3023, *XML Media Types*, January 2001, https://tools.ietf.org/rfc/rfc3023.txt

IETF RFC 3986, *Uniform Resource Identifier (URI): Generic Syntax*, January 2005, https://tools.ietf.org/rfc/rfc3986.txt

IETF RFC 6838, *Media Type Specifications and Registration Procedures*, January 2013, https://tools.ietf.org/rfc/rfc6838.txt

W3C XML, *Extensible Markup Language (XML) 1.0*, (Fifth Edition), T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler and F. Yergeau (eds.), W3C Recommendation 26 November 2008, http://www.w3.org/TR/2008/REC-xml-20081126/

W3C XML *Namespaces, Namespaces in XML 1.0*, (Third Edition), T. Bray, D. Hollander, A. Layman, R. Tobin and H. S. Thompson (eds.), W3C Recommendation 8 December 2009, http://www.w3.org/TR/2009/REC-xml-names-20091208/

W3C XSD, *XML Schema Part 1: Structures*, (Second Edition), H. S. Thompson, D. Beech, M. Maloney and N. Mendelsohn (eds.), W3C Recommendation 28 October 2004, http://www.w3.org/TR/2004/REC-xmlschema-1-20041028/

W3C XSD Part 2: *Datatypes XML Schema Part 2: Datatypes*, (Second Edition), P.V. Biron and A. Malhotra (eds.), W3C Recommendation 02 May 2001, http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/

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 General terms

3.1.1
CLARIN infrastructure, CLARIN
infrastructure governed by the CLARIN ERIC

3.1.2
concept
abstract idea conceived in the mind or generalised from particular instances

Note 1 to entry: cf. Merriam-Webster **Dictionary and Thesaurus**, definition of *concept*.

3.1.3.
concept link
reference from a CMD profile, CMD component, CMD element, CMD attribute or a value in a controlled vocabulary to an entry in a semantic registry via a URI, typically a persistent identifier

3.1.4
concept registry
semantic registry maintaining concepts, e.g., the CLARIN Concept Registry as used in the CLARIN infrastructure

3.1.5
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

3.1.6
data category
result of the specification of a given data field

[SOURCE: ISO 12620:2009, 3.1.3]

3.1.7
language tag
textual code used to assist in identifying languages, whether spoken, written, signed, or otherwise signaled, for the purpose of communication

Note 1 to entry: This includes constructed and artificial languages but excludes languages not intended primarily for human communication, such as programming languages (IETF BCP 47).

3.1.8
media type, MIME type
type which specifies the nature of the data as described in IETF RFC 6838

3.1.9
metadata
resource 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 may contain information about the resource, aspects or parts of the resource and/or artefacts and actors connected to the resource.

3.1.10
persistent identifier, PID
Unique Uniform Resource Identifier that assures permanent access for a resource by providing access to it independently of its physical location or current ownership

3.1.11
resource
entity, possibly digitally accessible, that can be described in terms of its content and technical properties, referenced by a Uniform Resource Identifier

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

Note 1 to entry: These registries should also provide persistent identifier for their entries.

3.1.13
term
verbal designation of a general concept in a specific subject field

[SOURCE: ISO 1087-1:2000, 3.4.3]

3.1.14
Uniform Resource Identifier, URI
identifier for resource as described in IETF RFC 3986

3.2 CMDI

3.2.1
CCSL
CMDI Component Specification Language
XML based language for describing CMD **component** and CMD **profile** according to the CMD model

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

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

3.2.4
CMD component registry
component registry
service where a CMD specification can be registered and accessed

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

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** level follows the specific structure defined by the CMD profile it relates to

3.2.7
CMD instance envelope
section of a CMD instance which is structured uniformly for all instances, and contains the CMD instance header and the list of resource proxies which may be referenced from the CMD instance payload section

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

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

3.2.10
CMD model
Component Metadata model
component based metadata model according to ISO 24622-1

3.2.11
CMD profile
profile definition, profile
structured template for the description of a class of resource providing the complete structure for a CMD instance **payload** by means of a hierarchy of CMD components

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

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

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

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