INTERNATIONAL STANDARD

ISO 24622-1

First edition
2015-02-01

Language resource management — Component Metadata Infrastructure (CMDI) —

Part 1: The Component Metadata Model

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

Partie 1: Composant modèle de métadonnées

Reference number
ISO 24622-1:2015(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 ..... **iv**
Introduction ..... **v**
1 Scope **1**
2 Terms and definitions **1**
3 Metadata schema availability and reuse **5**
3.1 Overview 5
3.2 Metadata components and elements 5
4 Semantics in the component metadata model **7**
4.1 Overview 7
4.2 Concept registries 8
4.3 Relation registries 8
5 Metadata component and profile - compatibility and versioning **9**
6 Expressiveness of the component metadata model **9**
Annex A (informative) **Abbreviations** **10**
Bibliography **11**

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 meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT), see the following URL: Foreword — Supplementary information.

The committee responsible for this document is ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management.*

ISO 24622 consists of the following part, under the general title *Language resource management — Component metadata infrastructure (CMDI):*

— *Part 1: The component metadata model*

A future part will address the component metadata specific language.

Introduction

Component Metadata (CMD) is an approach to metadata modelling and metadata creation. It is being increasingly used these days to enable the metadata description of different types of Language Resources (LRs) with different metadata schemas, while still trying to maintain syntactic and semantic interoperability.

CMD[1)] is also the core of the Component Metadata Infrastructure (CMDI)[1]: this infrastructure contains not only the format specifications for this metadata modelling and creation approach, but also a set of registries and tools for metadata modelling and creation work.

The advantages of having such a unified approach to metadata descriptions for LRs, an approach that will be usable by many projects and initiatives, are obvious: firstly, there is a better chance of obtaining interoperability between metadata descriptions from different sources, and secondly, it will be possible to develop and share tools that work much more efficiently in this metadata framework.

The challenge of designing and organizing a comprehensive and unified approach to metadata description for the very varied set of LR types, and one that also can satisfy a sufficiently large section of the LR community, should not be underestimated. The landscape of metadata for LRs has been, and continues to be, fragmented. Until recently, it was the practice in creating the metadata descriptions for LRs to choose a specific metadata schema from a (small) existing set derived either from widespread traditions or from other disciplines; for example, OLAC[2] is an adapted version of DCMI,[3] which in turn originates in the library world. Additionally, there are, for the purposes of LR metadata description, specifically developed metadata schemas that can be limited in application to specific types of LR (e.g. IMDI[4]), or they can be of a proprietary nature (cf. the catalogues of the LR agencies such as LDC[2)] and ELRA[3)]). The result is a domain of LR metadata that is far from interoperable. Although some progress has been made in developing dedicated bridges for "translating" metadata from one specific schema to another and in providing a consolidated catalogue, this practice does not scale well since it depends on specific translations for each pair of different metadata schemas.

For some recent projects, founding principles have included the unification and consolidation of practices and the need to produce efficient and sufficiently specific metadata descriptions.

It follows that a number of international, European, and national projects and infrastructure initiatives such as CLARIN[5] and META-SHARE[6] now share the CMD approach to metadata for LRs. This International Standard will both standardize the fundamentals of this approach in order to achieve interoperability based on solid documentation, and foster cooperation between the various initiatives and projects that work on, and with, this International Standard.

The model description is the first part of an infrastructure that forms a complete package for the creation of metadata schemas. As stated in the Foreword, the complete infrastructure standard contains, in addition to this component metadata model specification (ISO 24622-1), one or more metadata component specification languages (planned), and a number of recommended metadata components and profiles (planned). Since this part of ISO 24622 specifies an abstract model, we will rely mainly on UML[7] to describe it.

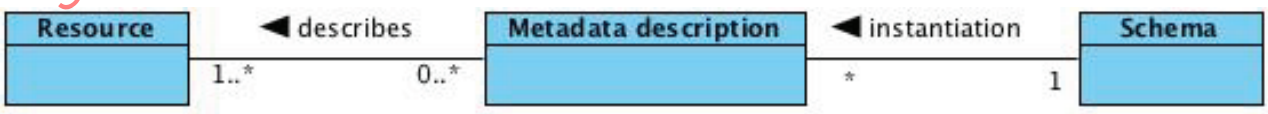

Figure 1 — Describing resources with metadata

1) Abbreviations are explained in Annex A.

2) Linguistic Data Consortium, http://www.ldc.upenn.edu/

3) European Language Resources Association, http://www.elra.info/

This part of ISO 24622 addresses the basic need to provide a model that makes it easy for metadata modellers (e.g. researchers and resource description experts) to create new metadata schemas, which can in turn be used either to describe new types of resources or to enable a more appropriate description for resources in specific circumstances. The metadata schema is instantiated into metadata records [i.e. the metadata descriptions that describe the actual resource(s)] (see Figure 1).

The context of this desire for flexible metadata modelling is that for scientific work there are usually various requirements for the proper description of LRs, and these requirements can derive from the specific needs of a project or from the facility or repository that will be used to store the resource for future use. This variation requires a flexible framework that enables the easy creation of new metadata schemas for different purposes, but is also a framework (i) in which the instantiations have a strictly defined format so that at least syntactic correctness can be checked, and (ii) which provides explicit semantics for the metadata schema elements for interpretation of the metadata record content.

The metadata descriptions generated by schemas compliant with this model will also be compliant with other TC 37 International Standards, for example, those requiring that references to the described resources and resource parts use ISO 24619:2011 PISA-compatible persistent identifiers (PIDs)[9].

The definition of a resource in this context is very broad. This part of ISO 24622 takes a pragmatic view: for example, an image can be a resource in itself when it is associated with a PID and can be referenced as such, or it can be part of a document where it lacks an identity of its own. In addition, a reference can point to a part of this image. An individual resource can stand alone in one environment and be treated as part of a collection in another environment. Also, metadata descriptions describe resources, but they, too, are a resource in different contexts. This part of ISO 24622 needs to support all such cases, and the model needs to provide descriptions at all levels of granularity.

This part of ISO 24622 takes two types of collections into account:

a) A complex resource may have been created as a collection originally and, versioning aside, it will exist as such in a rather static published form. Its specification will be treated as an independent entity by the responsible archiving institution that also provides a PID for such a collection. In the context of this part of ISO 24622, the metadata for the collection is the collection specification. The archiving institution is responsible for maintaining the metadata representing the collection.

b) In contrast, a different type of collection is one that was not planned and designed as a collection by its creators or by the holding archive, but achieves its status as a federated resource based on research that needs to be verifiable. Such collections, although purposefully constructed by the researcher, may not have any significance outside the context of the research for which they were created. Referring from the research documents to the collection may also become tedious if the collection contains hundreds of individual resources. It follows that there is a need to capture these types of collection with a metadata record that is associated with all its constituent resources and appropriate metadata, but only as the incarnation of this collection. There is no natural responsible party to maintain this metadata record. It is unlikely that the researcher who created the "virtual" collection (VC) has any way of consistently maintaining and curating this metadata record in the long term. There may be special registries maintained by digital archives or publishers where researchers can register such virtual collections.

Both types of collection are identified with the PID that refers to the collection metadata.

Language resource management — Component Metadata Infrastructure (CMDI) —

Part 1: The Component Metadata Model

1 Scope

The scope of this part of ISO 24622 is to describe a model that enables the flexible construction of interoperable metadata schemas for Language Resources (LRs). The metadata schemas based on this model can be used to describe resources at different levels of granularity (e.g. descriptions both on the collection level and on the level of individual resources).

2 Terms and definitions

2.1
archive
digital archive
repository (2.26) dedicated to the long-term preservation of the associated data

Note 1 to entry: The data in digital archives are also often available on-line. This highlights the need for reliable *PIDs* (2.22)

2.2
cardinality
metadata component cardinality
metadata element cardinality
specification of the number of occurrences of a *metadata component* (2.14) or *metadata element* (2.12) in an instantiation

2.3
citation
object containing information that directs a textual resource reader's or user's attention from one resource to another

2.4
closed vocabulary
limited set of items that forms the mandatory value domain of a *metadata element* (2.12)

2.5
concept reference
concept link
reference to the definition of a concept in a *concept registry* (2.6)

2.6
concept registry
registry (2.25) for registering concepts enabling their identification with a unique identifier

2.7
collection
resource collection
grouping of multiple, different constituting elements, each of which is independent of the others and may be accessed individually

Note 1 to entry: A collection can be a virtual collection if its constituent elements come from other different (virtual) collections, and possibly if the elements are distributed over different repositories.

2.8
fragment identifier
identifier (2.9) used to reference a *resource part* (2.28) in a web context

[SOURCE: ISO 12619:2011]

2.9
identifier
digital identifier
compact sequence of characters associated with digital, non-digital, or abstract entities

[SOURCE: Adapted from ISO 12619:2011]

Note 1 to entry: Identifiers can apply to entities such as books, images, reports, metadata records, and events.

2.10
metadata record
metadata description
metadata
record (2.23) containing a description of a *resource* (2.27)

2.11
metadata schema
schema
specification of a format and structure for a *metadata record* (2.10)

Note 1 to entry: In the context of this part of ISO 24622, a machine-readable and verifiable format specification usually defined by an XML schema language.

2.12
metadata element
resource property name that can be used in metadata and that can be given a value

Note 1 to entry: A metadata element is referred to as metadata attribute in other communities.

EXAMPLE The DCMI elements.[3]

2.13
metadata set
metadata element set
collection of *metadata elements* (2.12) used within a particular discipline, tradition, or practice to describe *resources* (2.27)

Note 1 to entry: A metadata set is more general than a metadata schema in that it does not additionally specify the syntax (e.g. the DCMI elements[3]).

2.14
metadata component
grouping of *metadata elements* (2.12) and *metadata components* (2.14) that can be used to describe a specific aspect of a *resource* (2.27)

EXAMPLE The biographical data of a person or the contact information for an organization.

2.15
metadata component registry
component registry
registry (2.25) of *metadata components* (2.14) and *metadata profiles* (2.16) for their sharing

2.16
metadata profile
set of *metadata components* (2.14) that can be used together to describe a *resource* (2.27) and be transformed into a *metadata schema* (2.11)

Note 1 to entry: A metadata profile can be transformed into different metadata schemas that are still logically equivalent (i.e. they give logically equivalent resource descriptions).

2.17
metadata editor
actor that creates *metadata records* (2.10) to describe specific *resources* (2.27) or the tool that is used to edit the metadata record

2.18
metadata modeler
actor that creates new *metadata schemas* (2.11) for new types of *resources* (2.27) or new applications

Note 1 to entry: In this part of ISO 24622, metadata schemas are created by producing *metadata profiles* (2.16), which in turn form specifications for a metadata schema.

2.19
metadata provider
organization or software service that makes *metadata* (2.10) available

2.20
open vocabulary
set of items forming part of the value domain of a *metadata element* (2.12) on the recommendation of the *metadata modeler* (2.18)

2.22
Persistent Identifier
PID
unique *identifier* (2.9) that ensures permanent access to a digital object by providing access to it independently of its physical location or current ownership

[SOURCE: 24619:2011]

Note 1 to entry: In this context, "unique" means that the same PID will not be subsequently issued for other *resources* (2.27). However, the same PID may, at the *resource-provider's* (2.30) discretion, reference different representations or incarnations of the *resource* (2.27)

2.23
record
structured information that can be read by software services

2.24
reference
object that links to data stored elsewhere

Note 1 to entry: The words "citation" and "reference" are commonly used as quasi-synonyms. However, for the purposes of this part of ISO 24622, "citations" provide information for human readers and users, while "references" include the precise location where the referenced resource can be found. References can be machine-readable, and can be configured as actionable given the required criteria.

2.25
registry
central directory designed for the persistent provision of negotiated information that can be reliably accessed

Note 1 to entry: In this part of ISO 24622, a registry is a software service that allows registering and for the registry to be queried for information.

2.26
repository
digital repository
facility that provides reliable access to managed digital *resources* (2.27)

2.27
resource
object on the web with a specific identity

Note 1 to entry: For the purpose of this part of ISO 24622, the identity is one that can be expressed using a *URI* (2.32)

Note 2 to entry: In the context of this part of ISO 24622, a resource can also be a language resource that has an off-line representation

2.28
resource part
identifiable, accessible entity embedded in an independent *resource* (2.27) or in a larger part thereof

Note 1 to entry: Resource parts can be part of larger resource parts. In dynamic web environments, subsetting into parts is subject to change and interpretation, and this in turn requires a certain level of user decision-making to designate and identify such sub-entities.

2.29
resource part identifier
string of characters that refers to a *resource part* (2.28), and which can be identified by some means within a given resource type

EXAMPLE Such means are time for a media file, area for an image or record in a data stream.

2.30
resource provider
organization that makes a *resource* (2.27) available on-line

2.31
Unified Modeling Language
UML
language for specifying, visualizing, constructing, and documenting the artifacts of software systems and abstract models in general

2.32
Uniform Resource Identifier
URI
compact string of characters used to identify or name a *resource* (2.27) with a syntax defined in IETF RFC 3986

2.33
value scheme
metadata element value scheme
specification of the value domain of a *metadata element* (2.12)

3 Metadata schema availability and reuse

3.1 Overview

There is heavy demand in science and industry for a flexible description of digital language resources, resource parts, and collections of language resources. Not only is it desirable to be able to store and manage resources with high quality metadata for later use, but it is also becoming increasingly useful to use the metadata itself as a scientific resource. This makes it essential for there to be well-designed metadata schemas that are able to capture the maximum relevant information.

Although LRs are marked by considerable variety, and resource types accordingly warrant the use of many different metadata schemas, there is often a considerable information content overlap between these schemas. To minimize the time spent on creating new schemas, this part of ISO 24622 specifies a model with a flexible component-based approach that allows metadata modellers to create new schemas by reusing existing metadata modelling work captured in so-called metadata components. Metadata component specifications are specified in a component specification language and can be saved for future use, for instance in metadata component registries. To describe a specific resource type, a metadata modeller selects a number of appropriate metadata components and bundles these in a metadata profile. The metadata profile is specified in the metadata component specification language and is also stored for future use or extension, for example, in a metadata component registry.

To be functional, a metadata profile needs to be transformed into a metadata schema before it can be used to generate metadata descriptions. This approach resembles that taken by the metadata application profile initiative[8]. However, the use of metadata components turns the creation of a new schema by metadata modellers into a conceptual exercise rather than one of (XML) schema building.

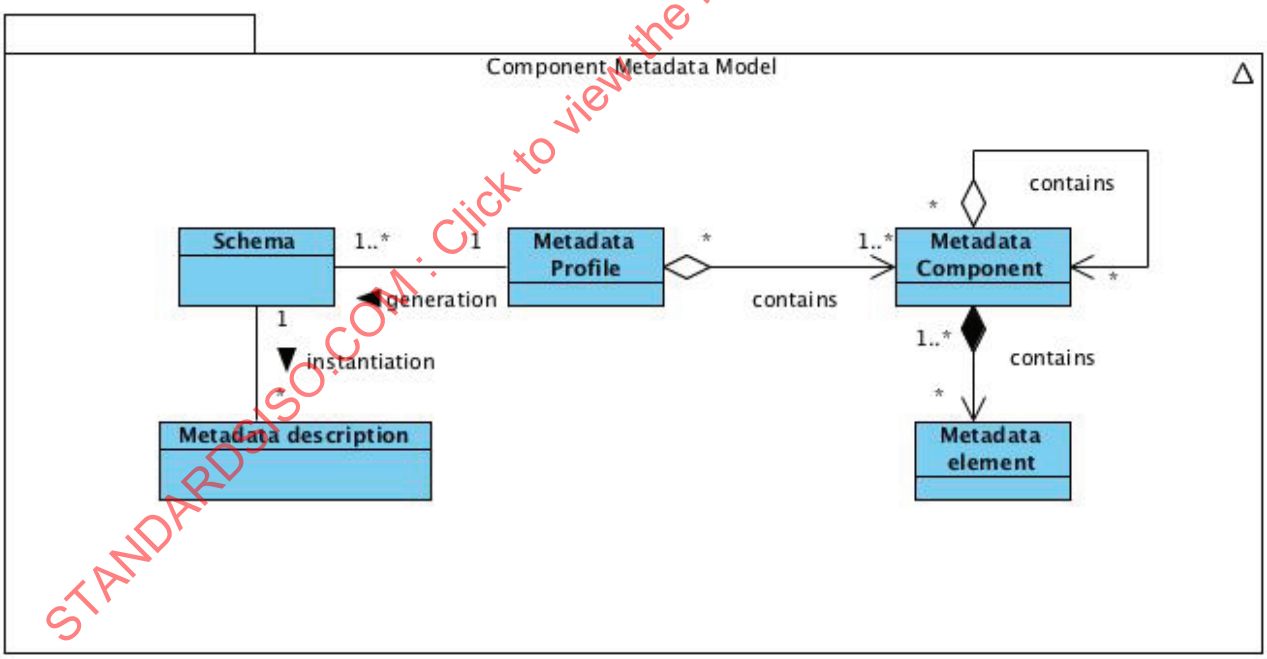

Figure 2 — The component metadata model

Figure 2 shows the relations between the metadata component, the metadata elements, and the metadata profiles. A metadata profile can be transformed into a schema specification from which the actual metadata descriptions can be instantiated.

3.2 Metadata components and elements

Metadata components can contain metadata elements and other metadata components, and the component model is therefore recursive (see Figure 2). Metadata components have a name and an

optional concept reference, and can have cardinality that specifies how often the metadata component can occur in a metadata record. The name of the metadata component is the name that the metadata component creator thinks is appropriate for the component.

A metadata element describes a single atomic aspect of a resource; it has a name, a concept reference, a value range specification, and cardinality (see Figure 3). The metadata element name is the name of the resource aspect that the metadata component creator believes to be appropriate. The concept reference is a PID and refers to the semantic definition of the metadata element (see the Semantics section in Clause 4) Semantics in the component metadata model). The value range or value scheme should provide an appropriate restriction of the values that the metadata element can take. The metadata element will, in the resultant instantiated metadata record, provide a real value for the resource aspect in the metadata description, for example:

element name: Country

element concept reference: (URI for wiki def.[4])

element value scheme: ISO 3166-2

element cardinality: 0..1

The value scheme of a metadata element can take several forms:

a) **data types** specific to the implementation language (to be defined in CMDI Part-2);

b) **closed vocabulary**;

c) **open vocabulary**;

d) **regular expression**.

If the value scheme is an open or closed vocabulary, the items in this vocabulary need to have their own concept references (see Clause 4). The metadata component shall provide such a list of items in the metadata component itself or else a reference to a remote source for the vocabulary.

A metadata component can also be optionally associated with a concept reference; if so, a PID pointing to a concept in a concept registry is required.

The metadata element can have a cardinality specified; if a CMDI implementation supports cardinalities, it shall support at least the following options:

a) **cardinality (number of occurrences): *** (short for 0..*) is the default and stands for zero or more;

b) **cardinality: 0..1** stands for zero or exactly one;

c) **cardinality: 1..*** stands for minimal one;

d) **cardinality: 1** stands for exactly one.

The last two cases are examples of obligatory elements, that is to say the metadata component or element shall occur at least once in every instantiation.

4) Some may not see a wiki as an acceptable concept registry.

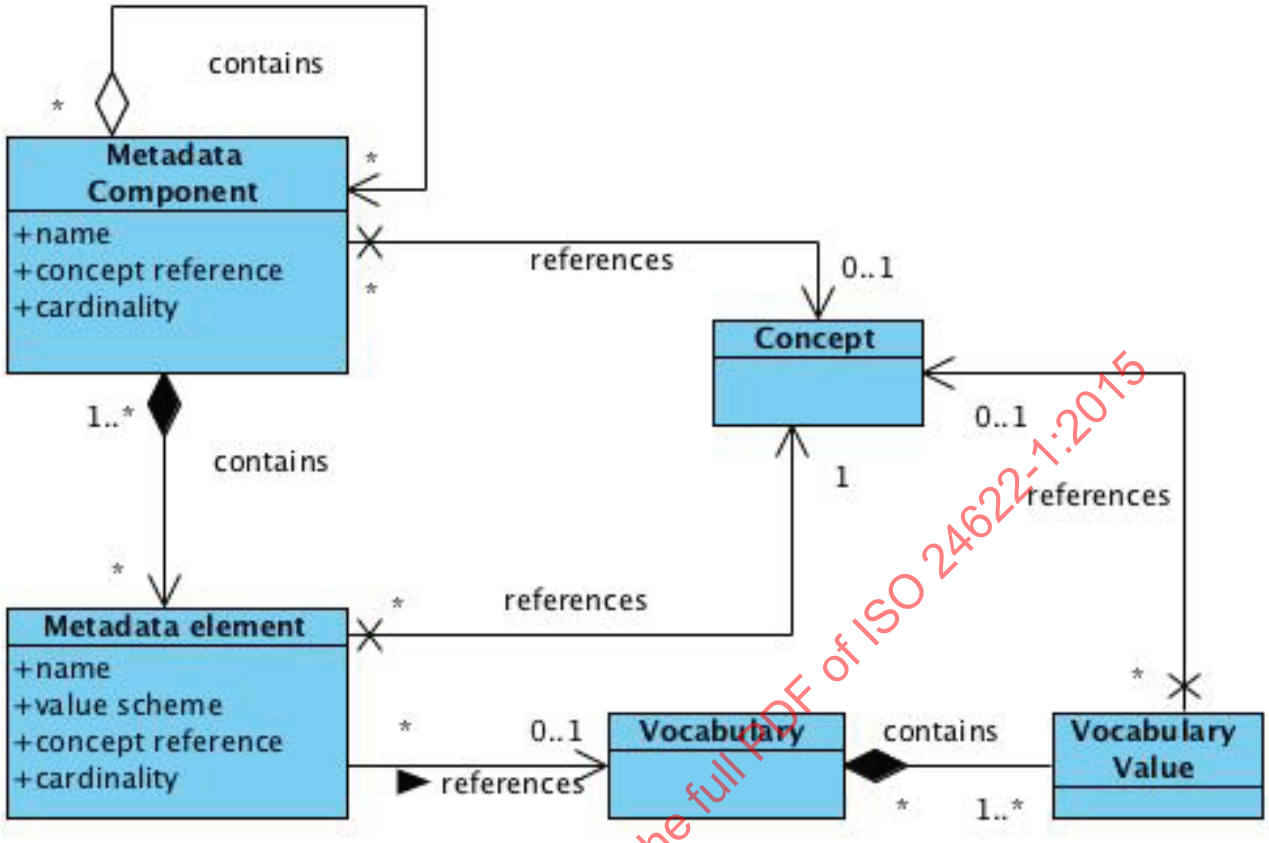

Figure 3 — Metadata component and element detail

4 Semantics in the component metadata model

4.1 Overview

The metadata component framework encourages the creation of new metadata components and profiles to facilitate accurate descriptions of resources.

The terminology used within metadata profiles and metadata components is determined by the metadata modeller; this enables communities and even individual modellers to use familiar names that match their traditions and which they feel comfortable using. However, to make semantic interoperability between metadata components possible, every metadata element must refer to a concept definition in a recognized concept registry. Metadata components may also be associated with a concept definition, although this is not required (see Figure 4). Using the concept references, semantic relations can be computed between metadata records originating from different metadata schemas. This can be used by metadata service providers that collect metadata records from communities using different metadata schemas and wish to show all the metadata in one consolidated metadata catalogue.

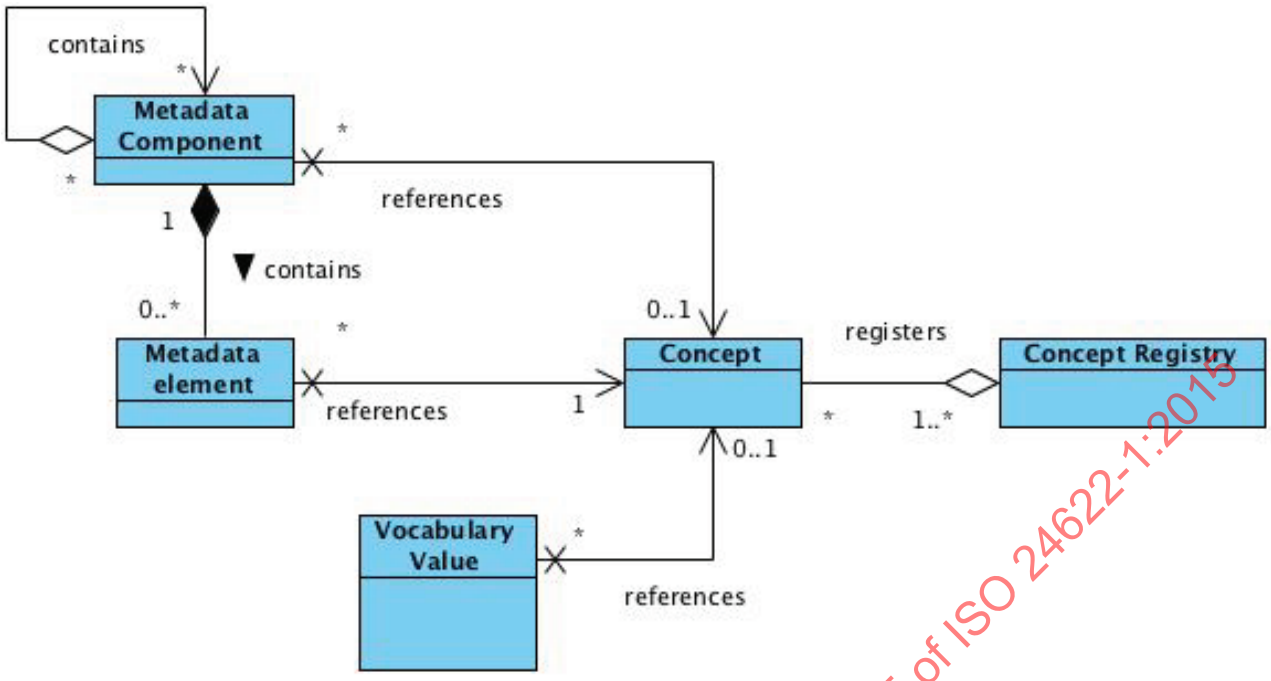

Figure 4 — Concept references in CMDI

4.2 Concept registries

A concept registry is a directory that stores definitions of concepts together with a unique identifier for each concept. The identifier is used to compute semantic relations between different metadata records.

The minimum requirements for an "accepted" concept registry are the following:

a) stable and persistent accessibility;

b) unique identifiers for registered concepts;

c) at least an English language definition section;

d) a transparent versioning and governance policy.

The ISOcat DCR, an implementation of ISO 12620:2009[10] under the control of TC 37, can be used as one such concept registry; DCMI is another.

4.3 Relation registries

A relation registry is a directory that stores semantic relations between concepts. Some concept registries are capable of registering relations between the registered concepts to varying degrees. If, in the semantic processing of the metadata records, the concept registry does not provide sufficient expressiveness, an additional relation registry can be introduced that is able to correct this deficiency. The relation registry is a directory that stores a relation type and the (concept registry) concept identifiers involved.