BS ISO 24622-1:2015

BSI Standards Publication

# Language resource management — Component Metadata Infrastructure (CMDI)

## Part 1: The Component Metadata Model

bsi.

...making excellence a habit.™

This British Standard is the UK implementation of ISO 24622-1:2015.

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.


Published by BSI Standards Limited 2015

ISBN 978 0 580 81379 5
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 January 2015.

**Amendments/corrigenda issued since publication**

| Date | Text affected |
|---|---|

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

# 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.

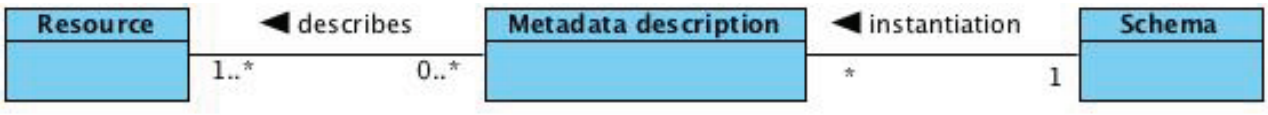


**Figure 1 — Describing resources with metadata**

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.

INTERNATIONAL STANDARD ISO 24622-1:2015(E)



# 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