SLOVENSKI STANDARD
SIST ISO 24612:2013

01-julij-2013

Upravljanje z jezikovnimi viri - Ogrodje za jezikoslovno označevanje (LAF)

Language resource management -- Linguistic annotation framework (LAF)

Gestion des ressources langagières -- Cadre d'annotation linguistique (LAF)

Ta slovenski standard je istoveten z: **ISO 24612:2012**

ICS:

01.020	Terminologija (načela in koordinacija)	Terminology (principles and coordination)

SIST ISO 24612:2013 **en,fr,de**

INTERNATIONAL STANDARD

ISO 24612

First edition
2012-06-15

Language resource management — Linguistic annotation framework (LAF)

Gestion des ressources langagières — Cadre d'annotation linguistique (LAF)

Reference number
ISO 24612:2012(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 24612 was prepared by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

Introduction

Effective creation, encoding, processing and management of language resources is facilitated by a single high-level data model that supports analysis and design of both annotation schemes and representation formats. This International Standard is designed to support the development and use of computer applications relying on linguistically annotated resources and the exchange of these resources among different applications.

Language resource management — Linguistic annotation framework (LAF)

1 Scope

This International Standard specifies a linguistic annotation framework (LAF) for representing linguistic annotations of language data such as corpora, speech signal and video. The framework includes an abstract data model and an XML serialization of that model for representing annotations of primary data. The serialization serves as a pivot format to allow annotations expressed in one representation format to be mapped onto another.

NOTE Standardization of linguistic data categories that provide annotation content is provided by ISO 12620 and other related International Standards.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

2.1
primary data
electronic representation of language data

EXAMPLE Text, image, speech signal.

Note to entry: Typically, primary data objects are addressed by “locations” in an electronic file, for example, the span of characters comprising a sentence or word, or a point at which a given temporal event begins or ends (as in speech annotation). More complex data objects may consist of a list or set of contiguous or non-contiguous locations in primary data.

2.2
annotate, verb
process of adding linguistic information to *primary data* (2.1)

2.3
annotation, noun
linguistic information added to *primary data* (2.1), independent of its representation

2.4
representation
format in which the *annotation* (2.3) is rendered, independent of its content

EXAMPLE XML, list or bracketed format, tab-delimited text.

2.5
segmentation annotation
annotation (2.3) that delimits linguistic elements that appear in the *primary data* (2.1)

Note to entry: These elements include (1) continuous segments (appearing contiguously in the primary data), (2) super- and sub-segments, where groups of segments will comprise the parts of a larger segment (e.g. contiguous word segment typically comprise a sentence segment), (3) discontinuous segments (linking continuous segments), and (4) landmarks (e.g. timestamp) that note a point in the primary data. In current practice, segmental information may or may not appear in the document containing the primary data itself.

2.6
linguistic annotation
annotation (2.3) that provides linguistic information about the segments in the *primary data* (2.1)

EXAMPLE Morphosyntactic annotation in which a part of speech and lemma are associated with each segment in the data.

Note to entry: The identification of a segment as a word, sentence, noun phrase, etc. also constitutes linguistic annotation. In current practice, when it is possible to do so, segmentation and identification of the linguistic role or properties of that segment are often combined (e.g. syntactic bracketing, or delimiting each word in the document with an XML element that identifies the segment as a word or sentence).

2.7
stand-off annotation
annotation (2.3) layered over *primary data* (2.1) and serialized in a document separate from that containing the primary data

Note to entry: Stand-off annotations refer to specific locations in the primary data, by addressing character offsets, elements, etc. to which the annotation applies. Multiple stand-off annotation documents for a given type of annotation can refer to the same primary document (e.g. two different part of speech annotations for a given text).

2.8
annotation document
XML document containing *annotations* (2.3)

2.9
anchor
fixed, immutable position in the *primary data* (2.1) being *annotated* (2.2)

Note to entry: The medium determines how an anchor is described. For example, text anchors may be character offsets, audio anchors may be time offsets, video anchors may be time offsets or frame indices, image anchors may be coordinates.

2.10
region
area in the *primary data* (2.1) defined by a non-empty, ordered list of *anchors* (2.9)

2.11
original artefact
artefact or *annotation* (2.3) from which the *primary data* (2.1) is derived

2.12
graph
set of nodes (vertices) V(G) and a set of edges E(G)

2.13
node
vertex
terminal point in a graph G, or the intersection of edges in G

Note to entry: The terms *node* and *vertex* are used interchangeably in this document.

2.14
edge
ordered pair of nodes [u,v] from V(G)

Note to entry: The order of the nodes determines the direction of the edge.

3 LAF specification

3.1 Overview

LAF consists of the following.

— A data model for linguistic annotations and the data to which they apply.

— An architecture for representing language data and its annotations.

— An XML serialization of the data model, which describes the referential structure of annotations associated with language data, consisting of a directed graph or graphs. Nodes in the graph may be linked to regions of primary data. Nodes and edges may be associated with feature structures describing linguistic properties of regions of primary data linked to reachable nodes.

3.2 LAF data model

The LAF data model consists of

a) a structure for describing media, consisting of *anchors* that reference locations in primary data and *regions* defined in terms of these anchors,

b) a *graph structure*, consisting of nodes, edges and links to regions, and

c) an *annotation structure* for representing annotation content with feature structures.

The data model for annotations thus comprises a directed graph referencing n-dimensional regions of primary data as well as other annotations, in which nodes are associated with feature structures providing the annotation content. LAF conformance requires that an annotation scheme shall be (or be rendered via the mapping) isomorphic to the LAF data model.

NOTE LAF does not include specifications for annotation content categories (i.e. the contents of the associated linguistic phenomena).

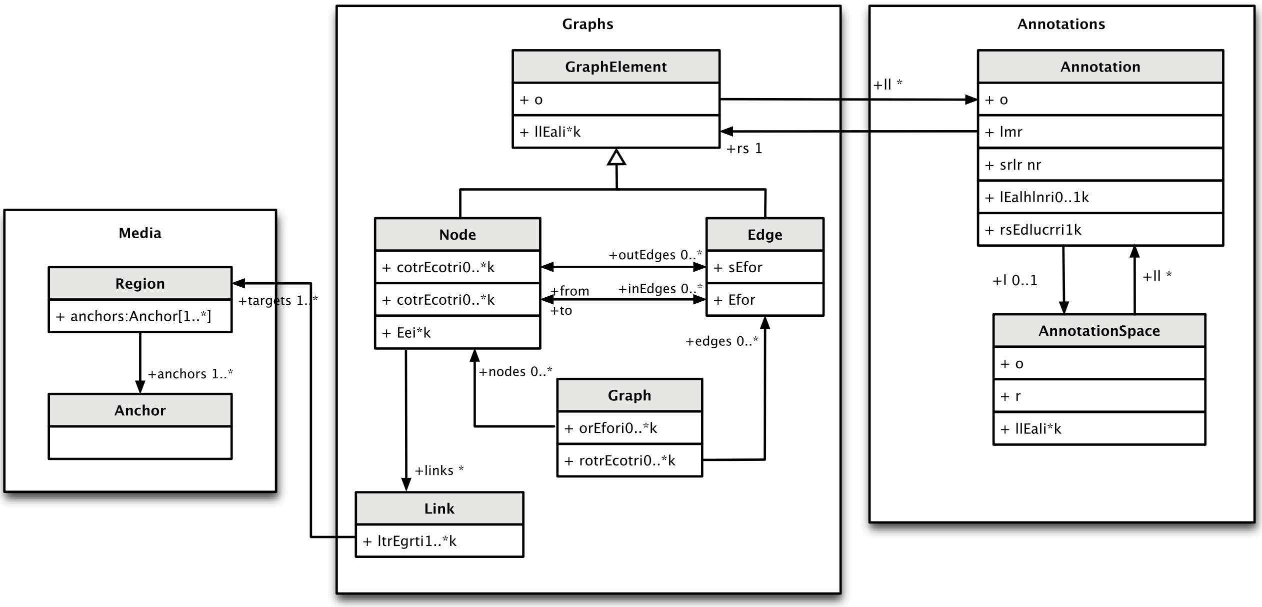

Figure 1 — LAF data model