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).

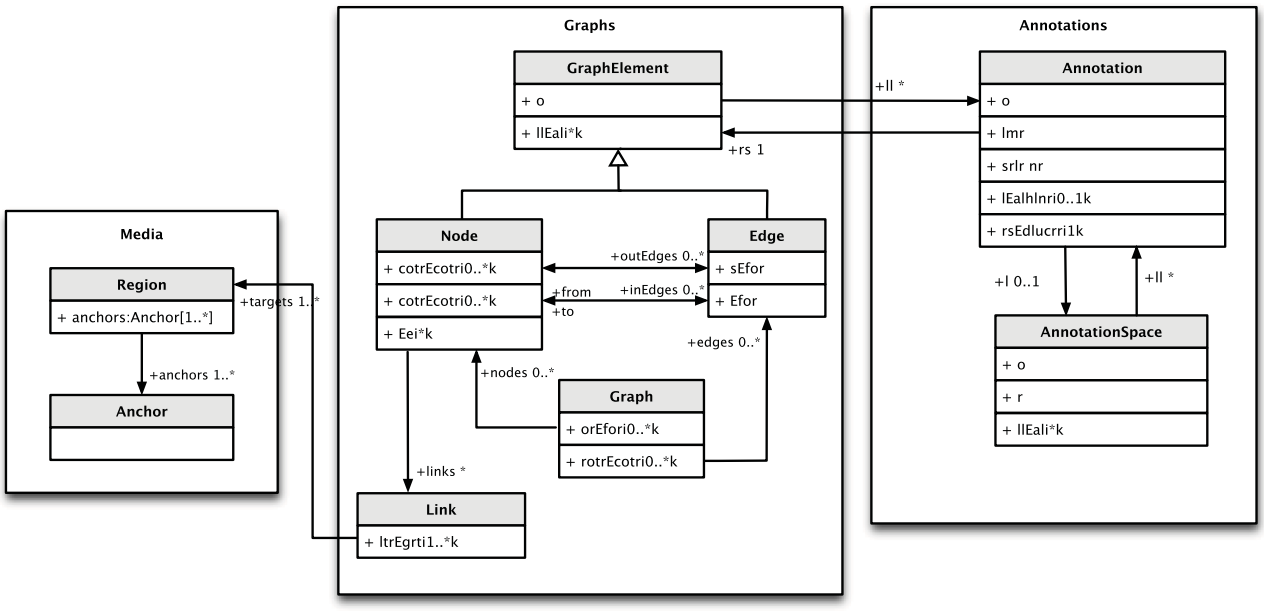

Figure 1 — LAF data model

3.3 LAF architecture

3.3.1 Overview

Language resources conforming to the LAF architecture consist of the following, described in more detail in 3.3.2 to 3.3.5.

— One or more primary data documents (see 3.3.2).

— Any number of annotation documents containing nodes, edges and feature structures associated with some or all of the nodes and/or edges in a directed graph. All nodes reference either a base segmentation document (in which case the node has no outgoing edges) or other nodes in the same or other annotation documents via edges. (See 3.3.3).

— One or more documents defining regions that reference each primary data document, which serve as the base segmentation for annotations (see 3.3.4.)

— A set of headers, including a resource header describing a collection of primary data documents and annotations, as well as headers for each primary data document and each annotation document in the collection (see 3.3.5).

It is recommended that whenever possible, each primary data document also be associated with an *original artefact* containing the source from which the primary data was adapted or extracted for annotation (e.g. the original text in the file format of a particular word processor or file viewer).

3.3.2 Primary data

Primary data consists of electronic data in any format, including character (text), image, audio and video. Primary data in a LAF-compliant resources are frozen as “read-only” to preserve the integrity of references to locations within the document or documents. Corrections and modifications to the primary data are treated as annotations and stored in a separate annotation document. Primary data documents containing textual data are encoded in UTF-8 (default) or UTF-16.

In the general case, primary data does not contain markup of any kind. If markup does exist in primary data (e.g. HTML or XML tags), it is treated as a part of the data stream by referring annotations; no distinction is made between markup and other characters in the data when referring to locations in the document.

3.3.3 Annotation documents

Annotation documents contain linguistic information describing primary data. Annotations are always associated with a node in a graph that directly references regions defined over primary data, either directly or via a path through reachable nodes. In the latter case, the annotations are said to be *layered* over the primary data. LAF recommends representing each of the linguistic layers defined in language resource management, in a separate annotation document for the purposes of exchange.

The granularity of the annotation — i.e. the smallest information unit to which the annotation applies — is dependent on the application. For example, a single annotation over text may cover a phoneme, word, sentence, paragraph, document, or an entire corpus; for audio it may cover any temporal interval, including a temporal “instant” (timeslot, timestamp, etc.).

3.3.4 References to primary data

Direct reference to locations in primary data is accomplished using *anchors*. In most cases, these nodes are located between the base units of the primary data representation.

Anchors are medium-dependent. Regions of a resource may be defined by specifying the anchors that bound the region. Regions in artefacts such as an image map or video may be defined in terms of anchors specifying

one or more coordinates, frame indexes, etc. Regions in audio data may be referenced in terms of anchors that refer to one or more points in the medium (e.g. an “instant” or “timestamp”). Anchors are represented by n-tuples consisting of sets of spatial and temporal offsets. For example, consider the text “My dog has fleas”:

```
                          1
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6
|M|y| |d|o|g| |h|a|s| |f|l|e|a|s|
```

The anchors for each word are the following:

```
My: start=0, end=2
dog: start=3, end=6
has: start=7, end=10
fleas: start=11, end=16
```

A set of regions defined over a document containing primary data need not be contiguous (i.e. there may be portions of the primary data not included in any region), but they should not, in general, overlap. Overlapping regions should be treated as composed of finer-grained sub-components. For example, two spans, <5, 9> and <7, 15>, can be reconstrued as three spans, a = <5, 7>, b = <7, 9, and c = <9, 15>. Two graph nodes can then be created that reference nodes <a, b> and <b, c>, thereby providing the coverage of regions <5, 9> and <7, 15>. Discontiguous regions are referenced by creating nodes referencing each component region and adding a node that is in turn linked to them.

The media types included in the resource are defined in the resource header. Each medium is associated with one or more anchor types. The header for each primary data document identifies the medium for that document, which in turn indicates the type of anchors used.

In the general case, primary data does not contain markup of any kind. If markup appears in primary data (e.g. HTML or XML tags), it is treated as a part of the data stream by referring annotations; no distinction is made between markup and other characters in the data when referring to locations in the document. For primary data comprising a valid XML document, anchors may reference XML elements using the W3C XPath 2.0 Language (www.w3.org/TR/xpath20/), in which case the associated anchor type is defined in the resource header as an XPath expression. References to locations within these XML elements (i.e. XML element content) can be made using standard offsets, which will be computed by including the markup as part of the data stream; in this case, two media types would be associated with the primary document’s file type. See 3.3.5.2 for a full description of anchor and media type definitions in the resource header.

3.3.5 Headers

3.3.5.1 Overview

LAF defines a header for a resource consisting of a collection of primary data documents and annotations, as well as headers for primary data and annotation documents themselves. This set of headers provides all metadata describing the provenance and encoding conventions for the data and its annotations, information required for processing such as anchor types or relations among primary data and annotation documents in the corpus.

3.3.5.2 Resource header

The resource header describes the resource as a whole, including its contents, file structure and encoding, and establishes definitions that are used in the primary data document and annotation document headers. Among these are the following.

— *Categories* used to describe primary data documents, typically the domain/subject area of general text.

— *File types* providing their naming conventions, media, annotation type, and dependencies (i.e. other file types that are referenced and therefore required). The specification of file types enables automatic validation that all required elements of the resource are present.