SLOVENSKI STANDARD

SIST ISO 24615-2:2018

01-september-2018

Upravljanje z jezikovnimi viri - Ogrodje za skladenjsko označevanje (SynAF) - 2. del: Serializacija XML (nabor Tiger)

Language resource management -- Syntactic annotation framework (SynAF) -- Part 2: XML serialization (ISOTiger)

Gestion de ressources linguistiques -- Cadre d'annotation syntaxique (SynAF) -- Partie 2: Sérialisation XML (ISOTiger)

Ta slovenski standard je istoveten z: ISO 24615-2:2018

ICS:

01.020	Terminologija (načela in koordinacija)	Terminology (principles and coordination)
01.140.20	Informacijske vede	Information sciences
35.240.30	Uporabniške rešitve IT v informatiki, dokumentiranju in založništvu	IT applications in information, documentation and publishing

SIST ISO 24615-2:2018 **en,fr,de**

INTERNATIONAL STANDARD

ISO 24615-2

First edition
2018-02

Language resource management — Syntactic annotation framework (SynAF) —

Part 2:
XML serialization (Tiger vocabulary)

Gestion de ressources linguistiques — Cadre d'annotation syntaxique (SynAF) —

Partie 2: Sérialisation XML (vocabulaire Tiger)

Reference number
ISO 24615-2:2018(E)

COPYRIGHT PROTECTED DOCUMENT

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, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

A list of all parts in the ISO 24615 series can be found on the ISO website.

Introduction

The need for standardization of syntactic annotation was recognized and addressed in detail with the publication of ISO 24615-1:2014. As a result of the work on ISO 24615-1:2014, it was anticipated that such a reference model for syntactic annotation should be associated with a concrete XML serialization in order to meet the specific needs of such applications as syntactic parsers or syntactic treebanks, where representations have to be exchanged and reused. Furthermore, such a serialization should be independent from the theoretical orientation and specific details of any specific annotation scheme.

This document answers this need on the basis of the seminal work carried out on the TigerXML format[3]. This starting point was chosen as a reference because it is widely used as a de facto standard for unrelated XML treebanks, with the advantages in terms of interoperability offered by its XML-based representations, as opposed to other frequently used formats, in particular, the Penn Treebank bracketing format[5] or the CoNLL format for dependency structures (see Reference[4]).

The document is designed to complement ISO 24615-1:2014 and to coordinate closely with ISO 24610, ISO 24611, ISO 24612 and ISO 12620.

This document therefore extends ISO 24615-1:2014 with an XML model based upon the Tiger XML vocabulary for the interchange of syntactically annotated data which is both standardized as well as language- and theory-independent. The proposed format directly instantiates all features of the meta-model defined in ISO 24615-1 and defines concrete serialized interfaces to the complementary ISO 24611 and ISO 12620, which provides the background for the DatCatInfo data category registry.

Language resource management — Syntactic annotation framework (SynAF) —

Part 2: XML serialization (Tiger vocabulary)

1 Scope

This document describes an XML-conformant serialization of the ISO 24615-1 meta-model, with the objective of supporting interoperability across language resources or language processing components in the domain of syntactic annotations. As an extension of ISO 24615-1, this document is also coordinated with ISO 24612.

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 12620, *Terminology and other language and content resources — Data category specifications*

ISO 24610 (all parts), *Language resource management — Feature structures*

ISO 24611, *Language resource management — Morpho-syntactic annotation framework (MAF)*

ISO 24612, *Language resource management — Linguistic annotation framework (LAF)*

ISO 24615-1, *Language resource management — Syntactic annotation framework (SynAF) — Part 1: Syntactic model*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 12620, ISO 24610 (all parts), ISO 24611, ISO 24612 and ISO 24615-1 and the following 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
domain
class of elements to which a certain set of *labels* (3.2) can be assigned

Note 1 to entry: Domains can refer generally to the set of all edges, terminal nodes or non-terminal nodes.

3.2
label
unit of annotation consisting of the name of a feature and a value, which together can be applied to appropriate model elements and add arbitrary feature-value annotations to such elements

3.3
primary data
initial raw linguistic content that is being encoded

3.4
sequential representation
representation of annotation content where the XML element structure mirrors the sequence of linguistic objects in the primary source

4 Graph structure and meta-model

In the XML Tiger format, annotations are represented in a graph structure. The graph structure can be described as G = (V, E, A) with

- a set of nodes V,
- a set of edges E with e = (v ∈ V, v ∈ V) ∈ E,
- a set of annotations A, where an annotation a is defined by a feature-value pair, and
- a function *annot*: E ∪ V → A.

A graph represents a bundle of interrelated nodes and edges. It is not specified which parts of a primary text are covered by a single graph, e.g. a sentence, a sub-sentence, a chapter or a whole text. Linguistic annotations represented by labels can be attached to nodes as well as edges.

The meta-model (see Figure 1) consists of three parts:

a) the structural organization of corpora and associated meta-data;

b) an annotation tagset definition;

c) the linguistic annotation graph.

The structural organization of corpora is represented by a recursively defined corpus element (Corpus) and its corresponding metadata (Meta). A corpus can contain subcorpora.

The annotation tagset definition is represented by a list of categories (Feature) containing the name of a category (Feature.name) and a list of category values (FeatureValue). Each FeatureValue contains a string representation of the value (FeatureValue.value). Together, both elements declare a tagset which is part of a specific corpus object. Such a tagset declaration is derivable, which means that all categories defined in a supercorpus object can also be used by its subcorpus objects. Further attributes are used to declare to which types of nodes and edges a category is applicable. Both elements allow reference to DatCatInfo entries in compliance with ISO 12620 via Unified Resource Identifiers (URIs) in the attributes Feature.dcrReference and FeatureValue.dcrReference.

The final part of the meta-model, the linguistic annotation graph, defines a set of elements containing the primary data and the annotation structure covering the primary data. It consists of the graph element itself (Graph), two classes of syntactic nodes (Terminal and NonTerminal), an edge element (Edge) and an annotation element (Annotation) realizing the *annot*-function and therefore referring to a feature name and its value. Graph is contained within Segment, which is a grouping mechanism to aggregate a set of syntactic nodes together. Such a group can have linguistic structural semantics, corresponding usually to a sentence, but possibly also to a line in a manuscript or other meaningful segments depending on the application and annotation scheme used by a specific project.

A terminal node (in ISO 24615-1 referred to as T_Node) constitutes the point of reference to the primary data. This can be a direct reference to a text span within the XML document or an indirect reference to an object outside the model, e.g. an element contained by a MAF file in compliance with ISO 24611, the Morpho-syntactic annotation framework. A non-terminal node is an inner node, referring directly or indirectly to a terminal node within the XML document.

An edge shall always have a source and a target node. Both of them can be either a terminal or a non-terminal node.

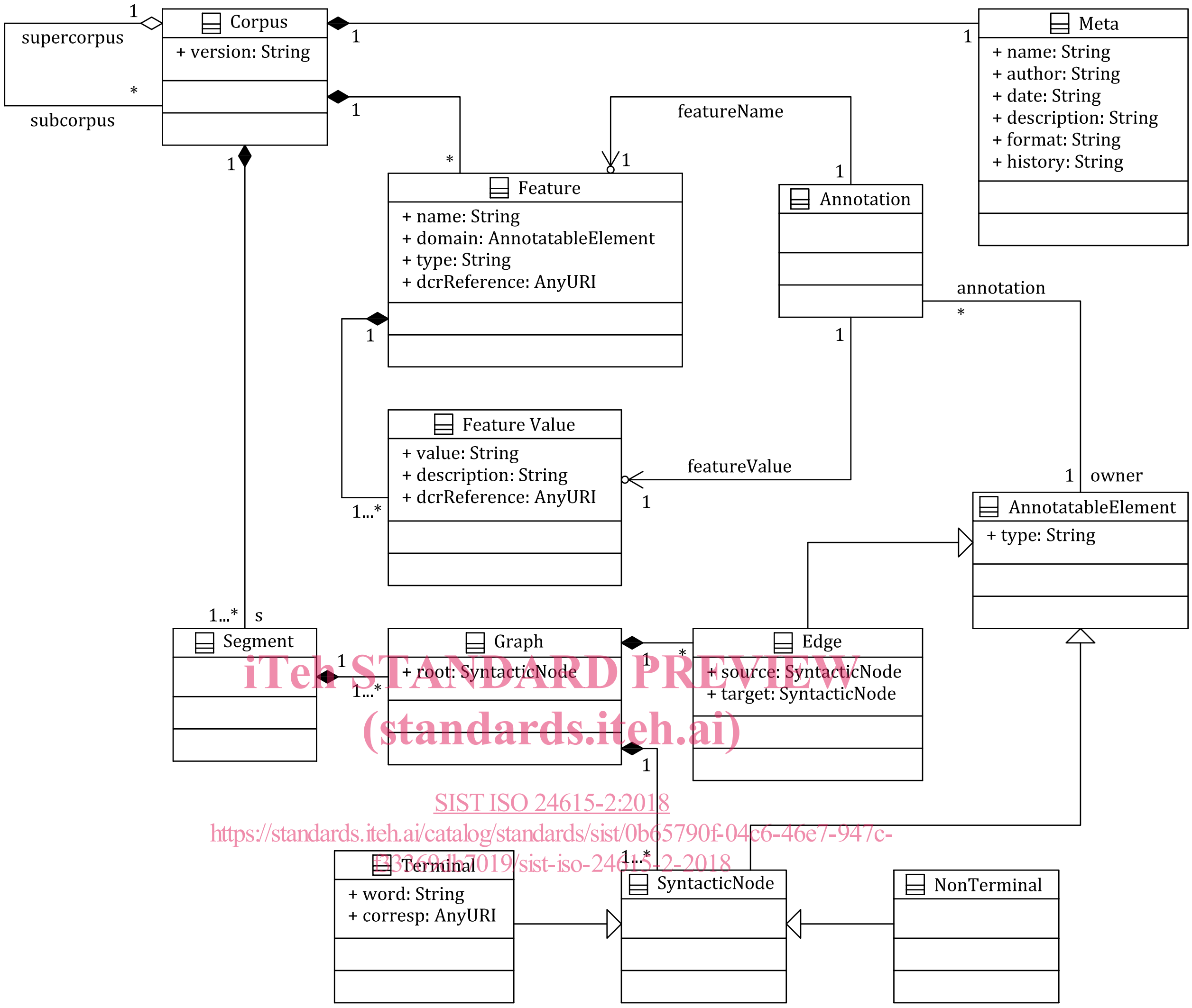

Figure 1 — Meta-model for the serialization format

5 Meta-model objects in XML serialization

The names of the XML elements and attributes follow those of the corresponding meta-model elements and attributes. All XML elements belong to the following namespace: http://www.clarin.eu/standards/ns/synaf. In this document, unless specified otherwise, all XML elements will be assumed to belong to this namespace.

Terminal nodes in the XML serialization are represented by the `<t>` element and are nested together in a `<terminals>` element.

A segment node is represented by the `<s>` element. The `<s>` element shall contain one or more `<graph>` elements, which may be used to express possible multiple annotation graphs alternating within a single segment or to represent a sequence of subgraphs.

A non-terminal node is represented by the `<nt>` element and is nested in the `<nonterminals>` element.