SVENSK STANDARD
SS-ISO 24615-2:2019

Fastställd/Approved: 2019-02-08
Utgåva/Edition: 1
Språk/Language: engelska/English
ICS: 01.020

SWEDISH STANDARDS INSTITUTE

Hantering av språkliga resurser – Format för syntaktisk annotering – Del 2: XML serialisering (TigerXML) (ISO 24615-2:2018, IDT)

Language resource management – Syntactic annotation framework (SynAF) – Part 2: XML serialization (Tiger vocabulary) (ISO 24615-2:2018, IDT)

Standarder får världen att fungera

SIS (Swedish Standards Institute) är en fristående ideell förening med medlemmar från både privat och offentlig sektor. Vi är en del av det europeiska och globala nätverk som utarbetar internationella standarder. Standarder är dokumenterad kunskap utvecklad av framstående aktörer inom industri, näringsliv och samhälle och befrämjar handel över gränser, bidrar till att processer och produkter blir säkrare samt effektiviserar din verksamhet.

Delta och påverka
Som medlem i SIS har du möjlighet att påverka framtida standarder inom ditt område på nationell, europeisk och global nivå. Du får samtidigt tillgång till tidig information om utvecklingen inom din bransch.

Ta del av det färdiga arbetet
Vi erbjuder våra kunder allt som rör standarder och deras tillämpning. Hos oss kan du köpa alla publikationer du behöver – allt från enskilda standarder, tekniska rapporter och standard-paket till handböcker och onlinetjänster. Genom vår webbtjänst e-nav får du tillgång till ett lättnavigerat bibliotek där alla standarder som är aktuella för ditt företag finns tillgängliga. Standarder och handböcker är källor till kunskap. Vi säljer dem.

Utveckla din kompetens och lyckas bättre i ditt arbete
Hos SIS kan du gå öppna eller företagsinterna utbildningar kring innehåll och tillämpning av standarder. Genom vår närhet till den internationella utvecklingen och ISO får du rätt kunskap i rätt tid, direkt från källan. Med vår kunskap om standarders möjligheter hjälper vi våra kunder att skapa verklig nytta och lönsamhet i sina verksamheter.

Vill du veta mer om SIS eller hur standarder kan effektivisera din verksamhet är du välkommen in på www.sis.se eller ta kontakt med oss på tel 08-555 523 00.

Standards make the world go round

SIS (Swedish Standards Institute) is an independent non-profit organisation with members from both the private and public sectors. We are part of the European and global network that draws up international standards. Standards consist of documented knowledge developed by prominent actors within the industry, business world and society. They promote cross-border trade, they help to make processes and products safer and they streamline your organisation.

Take part and have influence
As a member of SIS you will have the possibility to participate in standardization activities on national, European and global level. The membership in SIS will give you the opportunity to influence future standards and gain access to early stage information about developments within your field.

Get to know the finished work
We offer our customers everything in connection with standards and their application. You can purchase all the publications you need from us - everything from individual standards, technical reports and standard packages through to manuals and online services. Our web service e-nav gives you access to an easy-to-navigate library where all standards that are relevant to your company are available. Standards and manuals are sources of knowledge. We sell them.

Increase understanding and improve perception
With SIS you can undergo either shared or in-house training in the content and application of standards. Thanks to our proximity to international development and ISO you receive the right knowledge at the right time, direct from the source. With our knowledge about the potential of standards, we assist our customers in creating tangible benefit and profitability in their organisations.

If you want to know more about SIS, or how standards can streamline your organisation, please visit www.sis.se or contact us on phone +46 (0)8-555 523 00

Den internationella standarden ISO 24615-2:2018 gäller som svensk standard. Detta dokument innehåller den officiella engelska versionen av ISO 24615-2:2018.

The International Standard ISO 24615-2:2018 has the status of a Swedish Standard. This document contains the official English version of ISO 24615-2:2018.

Upplysningar om sakinnehållet i standarden lämnas av SIS, Swedish Standards Institute, telefon 08-555 520 00. Standarder kan beställas hos SIS som även lämnar allmänna upplysningar om svensk och utländsk standard.

Information about the content of the standard is available from the Swedish Standards Institute (SIS), telephone +46 8 555 520 00. Standards may be ordered from SIS, who can also provide general information about Swedish and foreign standards.

Denna standard är framtagen av kommittén för Språk och terminologi, SIS/TK 115.

Har du synpunkter på innehållet i den här standarden, vill du delta i ett kommande revideringsarbete eller vara med och ta fram andra standarder inom området? Gå in på www.sis.se - där hittar du mer information.

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.

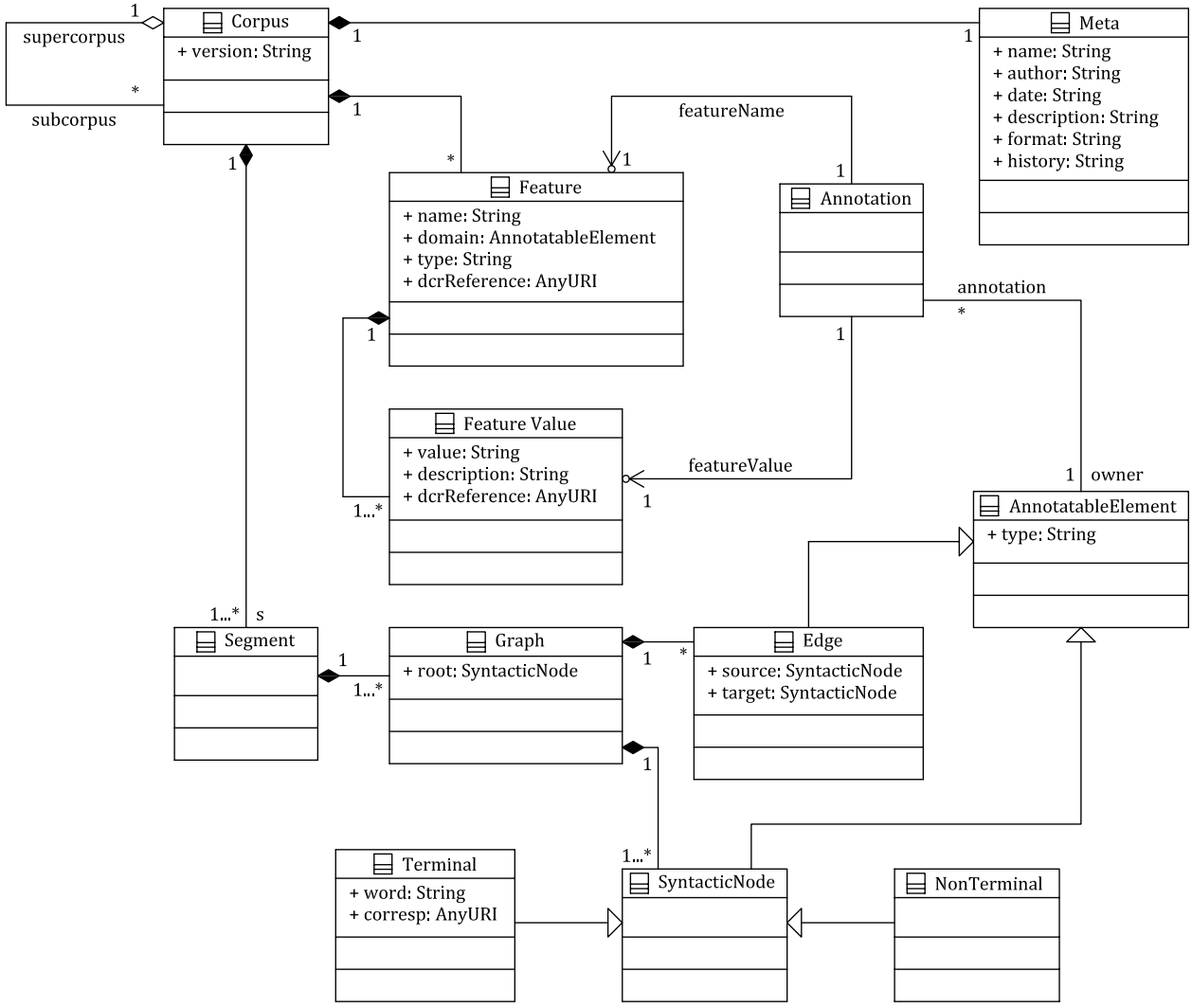

Figure 1 — Meta-model for the serialization format

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.

An edge is represented by the `<edge>` element and its source is given by the surrounding node element. An `<edge>` element is therefore always the child element of a `<t>` or an `<nt>` element. The target is given by the `@target` attribute and shall refer to a node within the same document.

The example shows the representation of the graph structure.

EXAMPLE Graph structure of a syntactic annotation.

```
<s>
    <graph xml:id="s1_g1">
        <terminals>
            <t xml:id="s1_t1"/>
            <t xml:id="s1_t2"/>
        </terminals>
        <nonterminals
            <nt xml:id="s1_nt1">
                <edge xml:id="s1_e1" target="#s1_t1" />
              <edge xml:id="s1_e2" target="#s1_t2" />
            </nt>
        </nonterminals>
    </graph>
</s>
```

6 Primary data and terminal representation

6.1 General

Terminal nodes and the primary data to which they refer can be defined in two different ways. The primary data can either be included within the XML document (sequential representation) or they can be specified in another file and referred to externally (standoff representation). The sequential representation makes it possible to have all data in just one XML document, whereas the standoff representation makes it possible to refer to other formats, for instance, a MAF file containing tokens and wordForms.

Both options are possible and the decision is largely made based on the needs of the corpus project.

6.2 Sequential representation

In a sequential representation, primary data is represented sequentially directly within the `<t>` elements of each `<s>` element in the document. As a result, all the primary data of a document is grouped together in one single XML document. This improves human readability and basic machine processing, but reduces representational flexibility. Tokens of the primary data are encoded as values of the `@word` attribute of the `<t>` element as illustrated in Example 1.

EXAMPLE 1 Simple sequential representation of word forms.

```
<t xml:id="s1_t1" word="two"/>
<t xml:id="s1_t2" word="words"/>
```

Example 1 shows two terminal nodes, the first containing the word "`two`" and the second containing the word "`words`". The order of the tokens in the primary data is mirrored by the order of the XML elements.

However, it should be pointed out that this representation does not need to fully preserve primary data, since only textual material that has a corresponding `<t>` element is preserved. This means that parts of an utterance that are not considered to be tokens (untokenized material), such as whitespaces between words or enumerations, non-linguistic characters, etc., will be lost. Example 2 and Example 3 illustrate this.

EXAMPLE 2 Primary data.

```
This is a sample.
```

EXAMPLE 3 Sequential representation.

```
<t xml:id="s1_t1" word="This"/>
<t xml:id="s1_t2" word="is"/>
```

```
<t xml:id="s1_t3" word="sample"/>
<t xml:id="s1_t4" word="."/>
```

In these examples, it is not possible to resolve whether there was originally a whitespace character between the tokens corresponding to the `@word` attributes of the terminals or not. In the case of the terminals "`s1_t1`" and "`s1_t2`" there was one, whereas between "`s1_t3`" and "`s1_t4`" there was no intervening whitespace. This can be handled by means of the `@join` attribute as provided by ISO 24611. It is also possible to omit textual material on purpose, leading to the text span corresponding to the word "`a`" becoming lost in the XML document.

Because of the restrictions of XML attributes, reserved XML signs such as angle brackets ("<") cannot be represented inside an XML attribute value. These are therefore escaped using the sequence "`&lt`"; (see Example 4), which, in some cases, can cause confusion in character-based offset calculations.

EXAMPLE 4 Sequential representation with character escaping.

```
<t xml:id="s1_t4" word="&lt; "/>
```

Cases where the primary data contains additional mark-up can be difficult to handle on a sequential encoding strategy.

6.3 Standoff representation

Problems concerning special characters or mark-up embedded in the primary data, described in 6.2, can be avoided by using a standoff representation. In a standoff representation, the `@corresp` attribute of the `<t>` element is used to point to a sequence of tokens or word forms in the primary data. The `@corresp` attribute replaces the `@word` attribute and contains a URI reference to another resource, for example, a MAF file. If both attributes are provided, the `@word` attribute takes precedence and the `@corresp` attribute can usually be ignored. The following examples show the use of the standoff representation with the `@corresp` attribute. For these examples, the token and wordForm elements from ISO 24611 data model are applied and the references into the primary data utilize anchors to refer to locations in between the base units of the data representation as specified in ISO 24612.

EXAMPLE 1 Locations in the primary data document.

```
<s xml:id="s1">We can…</s>
```

This sentence is associated with the following segmentation, with an inter-character numbering:

```
|W|e| |c|a|n|…
0 1 2 3 4 5 6
```

EXAMPLE 2 MAF standoff representation, in TEI flavour.

```
<w xml:id="t1" corresp="#string-range(s1,0,2)">
<w xml:id="t2" corresp="#string-range(s1,3,6)">

…

<span type="wordForm" xml:id="wf_1" lemma="we" corresp="#t1"/>
<span type="wordForm" xml:id="wf_2" lemma="can" corresp="#t2"/>
```