BS ISO 24612:2012

BSI Standards Publication

Language resource management — Linguistic annotation framework (LAF)

bsi.

...making excellence a habit.™

National foreword

This British Standard is the UK implementation of ISO 24612:2012.

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.

This publication does not purport to include all the necessary provisions of a contract. Users are responsible for its correct application.

ISBN 978 0 580 54235 0

ICS 01.020

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 30 November 2012.

Amendments issued since publication

Date	Text affected

BS ISO 24612:2012

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)

COPYRIGHT PROTECTED DOCUMENT

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.

INTERNATIONAL STANDARD **ISO 24612:2012(E)**

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).

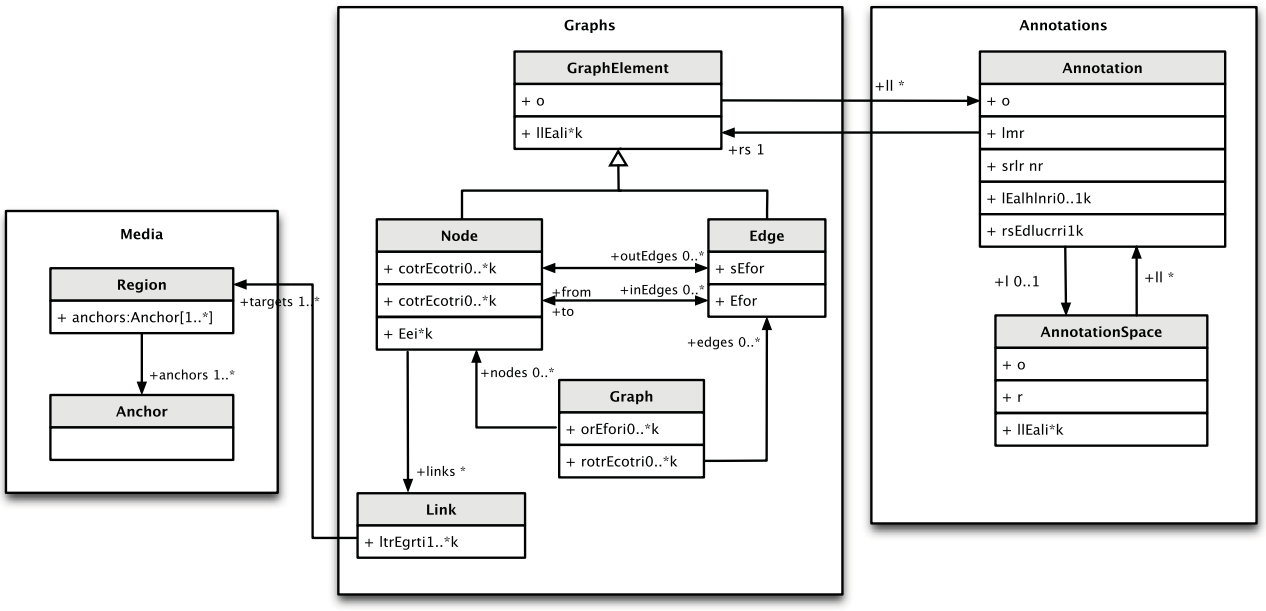

Figure 1 — LAF data model

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.

— *Annotation spaces* used to provide context for annotations and enable resolution of naming conflicts.

— *Annotation declarations* describing the annotations in the resource, including their names, creator, links to relevant documentation and, optionally, an associated annotation schema.

— *Media definitions* specifying the media types included in the corpus and file naming conventions for files containing data of that type.

— *Anchor types* associating anchor type definitions with media types.

— *Group definitions* providing the names, descriptions and members of user-defined groups of annotations.

3.3.5.3 Primary data document header

Each primary data document is associated with an XML header file containing information describing its contents. Because the primary data document is not an XML document, the LAF primary data header is obligatory and shall be provided as a standalone file.

The primary data document header provides information about the source and contents of the primary data, as well as specifying category definitions and medium type by reference to definitions in the resource header.

The primary data document header provides the PID for the primary data document and all associated annotation documents. The primary data document header provides all the information needed to process annotations associated with a given primary data document. It is presumed that this file is loaded first when a document and its annotations are to be processed.

The elements in the primary data document header are given in 3.6.

3.3.5.4 Annotation document header

The annotation document header includes a relevant subset of elements from the primary data header (i.e. those that describe the file contents rather than the provenance of an original text, etc.), together with additional elements that provide or point to information concerning the annotation content categories and dependencies between the annotation document and other documents. The annotation document header is not a separate document, but rather is included at the beginning of the annotation document. The elements in the annotation document header are given in 3.4.3.

3.4 XML pivot format

3.4.1 Overview

The LAF provides an XML serialization of the data model that is designated as the *pivot format*. A pivot format is intended to serve as an “interlingua” for translation among multiple other formats, by providing a common target into and out of which other formats can be transduced. Although the LAF pivot format may be used in any context, it is assumed that users will represent annotations using their own formats, which can then be transduced to the LAF pivot format for the purposes of exchange, merging and comparison.

The graph annotation format (GrAF) specifies the XML serialization of the LAF pivot format.

In GrAF:

— The fundamental data structure is a directed graph consisting of a set of nodes and a set of edges.

— An annotation is a label and (optionally) a feature structure associated with a node or an edge in the graph.

— A feature structure is an attribute value graph (AVG). The value of a feature may be an atomic value or another feature structure.

— An atomic feature value is a mapping from one string (the feature name) to another string (the atomic value). GrAF makes no attempt to do typing of feature values.

— Nodes may be associated with regions in the primary document, or connected to other nodes in the same or another annotation document. Nodes are associated with regions by <link> elements. Edges are used to connect (associate) nodes to other nodes.

— An edge represents a relationship between nodes. By default, the set of *out edges* from a node represent an ordered set of constituents of the annotation associated with the node. Other relationships may be specified by associating an annotation with the edge.

For all GrAF documents the namespace http://www.xces.org/ns/GrAF/1.0/ is used.

3.4.2 XML elements for annotation documents

The overall content model of a GrAF standoff annotation file is as follows:

```
graph = graphHeader (node | edge | a | anchor)*
node  = link*
a     = fs?
fs    = f+
f     = atomic | fs
start = graph
```

The root element is defined in Table 1. Required attributes are given in bold.

Table 1 — Root element for GrAF annotation documents

<graph>	Root node of the graph.
Attribute	***@xmlns* [URL]**: namespace declaration for the GrAF schema
Example	`<graph xmlns="http://www.xces.org/ns/GrAF/1.0/">`

3.4.3 The annotation document header

The overall content model for the annotation document header is as follows:

```
graphHeader      = labelsDecl, dependencies, annotationSpaces
labelsDecl       = labelUsage+
dependencies     = dependsOn+
annotationSpaces = annotationSpace+
start            = graphHeader
```

Elements of the annotation document header are given in Table 2, and elements to define graphs and annotations are given in Table 3. Required attributes are given in bold.

Table 2 — Elements of the annotation document header

<graphHeader>	Bracketing tag for elements of the annotation document header.
<labelsDecl>	List of the annotation labels used in the document and their frequencies.
<labelUsage>	Information for individual annotation labels.
Attributes	***@label*** **[string]**: Element name.
	@occurs **[integer]**: Number of occurrences in the document.
<dependencies>	Documents required to process the annotations in this document, which will include a segmentation document and/or any annotation documents directly referenced in this document.
<dependsOn>	File required to process this annotation.
Attributes	***@ann.id*** **[IDREF]**: The ID of the document as given in the associated primary data document.
<annotationSpaces>	Annotation spaces referenced in this document.
<annotationSpace>	Annotation space used in this document.
Attributes	***@as.id*** **[IDREF]**: The ID of the annotation space as defined in the resource header.
	@*default* [yes | no]: Indicates whether or not this annotation space is the default in this document. If the attribute is not present, *no* is assumed.

Table 3 — Graph and annotation elements in GrAF annotation documents

<roots>	One or more root elements that identify root nodes in the graph. This element is used when the graph contains either a graph that is a tree or a forest, i.e. more than one graph that is a well-formed tree.
<root>	The node ID of a root node in the graph. Not all graphs will form a tree, but those that do can use the root element to identify the root node of the tree.
Attribute	***@node.id*** **[IDREF]**: The ID of the root node.
<region>	Region in the artefact being annotated, defined as the area bounded by a non-empty, ordered list of anchors. The number of anchors required to bound a region depends on the medium being annotated.
Attributes	***@xml:id*** **[ID]**: Unique ID for reference from nodes in the graph.
	@*anchors* [string] (alternative to @*refs*): The anchors that bound this region. The anchors attribute contains a whitespace-delimited list of values that represent the anchor values. Applications are expected to know how to parse the string representation of an anchor into a location in the artefact being annotated. The <region> element shall have either an @*anchors* attribute or an @*ref* attribute.
	@*refs* [IDREFS] (alternative to @*anchors*): ID references to the anchors that bound the regions. The <region> element shall have either an @*anchors* attribute or an @*refs* attribute.
	@*anchor.id* [IDREF]: The anchor type of the anchors referenced in the @*anchors* attribute. This is the @*xml:id* of one of the anchorTypes defined in the resource header. If no @*anchor.id* is specified for the region, the default anchor type for the document (indicated on the <anchorType> element in the resource header) is assumed. If the @*refs* attribute is used to refer to <anchor> elements, the @*anchor.id* attribute will be specified on the <anchor> elements and should not be given on <region>.

Table 3 (*continued*)

Example	`<region xml:id="r1" anchor.id="time-slot"`
	`anchors="980 983"/>`
	`<region xml:id="r2" anchor.id="image-point"`
	`anchors="10,59 10,173 149,173 149,59"/>`
	`<region xml:id="r4" anchor.id="text-anchor"`
	`anchors="34 42"/>`

<anchor>	Location in the artefact being annotated. How the location is represented is medium-dependent. Applications are required to be able to serialize and de-serialize location values to and from strings appearing as attributes on the @value attribute as well as the @anchors attribute on the <region> element.
Attributes	***@xml*:*id* [ID]**: Unique ID for reference from nodes in the graph
	***@value* [string]**: The offset value of the anchor. How the attribute value is interpreted as a location in the artefact being annotated is medium-dependent.
	@anchor.id [string]: The *@xml:id* of an anchor type defined in the resource header.
<node>	Node in the graph. The element is empty when connected by an <edge> element to another node in the graph (i.e. when the node is a non-terminal node). A child <link> element is used when the node refers to a region or regions of primary data (i.e. when the node is a terminal/leaf node).
Attribute	***@xml*:*id* [ID]**: Unique ID for reference from edges and annotations.
<link>	Identifies region(s) in a base segmentation document referred to by this node
Attribute	***@targets* [IDREFS]**: Identifiers of referenced region(s)
Example	<node xml:id="penn-n10"> <link targets="seg-r20"/> </node>
<edge>	Edge in the graph.
Attributes	***@xml*:*id* [ID]**: Unique ID for reference from nodes and annotations.
	***@from* [IDREF]**: ID of the start node of the edge.
	***@ to* [IDREF]**: ID of the end node of the edge.
Example	`<edge xml:id="e595" from="fn-n361" to="fn-n348"/>`
<a>	Annotation information associated with a node or edge. This tag may be empty if the annotation consists of a label only.
Attributes	***@label* [string]**: The label of the annotation. This may be the string used to identify the annotation as described by the annotation documentation, a category identifier from a data category registry, an identifier from a feature structure library, or any reference to an external annotation specification.
	***@ref* [IDREF]**: The ID of the node or edge with which the annotation is associated.
	@as [string]: The ID of the annotation space of which this annotation is a part, as defined in the resource header; if no *@as* attribute is specified, the annotation space designated as the default in the annotation document header is assumed.
<fs>	Feature structure providing additional annotation information. An <a> element may not contain more than one <fs> element. The <fs> element may contain one or more <f> elements.
<f>	Attribute/value pair. In the concise form (given here), the <f> element is empty and includes attributes providing simple name/value pairs. More complex feature structures may be represented according to the specification in ISO 24610-1, which should be consulted for details.

Table 3 (*continued*)

Attributes	***@name* [string]**: The name of the attribute, as specified in the documentation of the annotation scheme.
	***@value* [string]**: The value for the attribute, taken from possible values provided in the documentation or schema for the annotation scheme.

Example	`<a xml:id="de6" label="vchunk" ref="vc-n10" as="xces">`
	`<fs>`
	`<f name="voice">active</f>`
	`<f name="tense">SimPre</f>`
	`<f name="type">FVG</f>`
	`</fs>`
	`< </a>`

3.4.4 Examples

```
<!-- Fragment of an annotation document header
     and an annotation over text                -->

<graph xmlns="http://www.xces.org/ns/GrAF/1.0/">
    <graphHeader>
       <labelsDecl>
           <labelUsage label="fullTextAnnotation" occurs="1"/>
           <labelUsage label="Target" occurs="171"/>
           <labelUsage label="FE" occurs="372"/>
           <labelUsage label="sentence" occurs="32"/>
           <labelUsage label="annotationSet" occurs="171"/>
           <labelUsage label="NamedEntity" occurs="32"/>
       </labelsDecl>
       <dependencies>
           <dependsOn type="fntok"/>
       </dependencies>
       <annotationSpaces>
       <annotationSpace as.id=”FrameNet" default=”true”/>
       </annotationSpaces>
    </header>
    …
    <node xml:id="fn-n156"/>
    <a label="FE" ref="fn-n156">
        <fs>
            <f name="FE" value="Speaker"/>
            <f name="rank" value="1"/>
            <f name="GF" value="Ext"/>
            <f name="PT" value="NP"/>
        </fs>
    </a>
    <edge xml:id="e233" from="fn-n156" to="fn-n133"/>
    <edge xml:id="e232" from="fn-n155" to="fn-n156"/>
    <edge xml:id="e231" from="fn-n132" to="fn-n155"/>
   …

   <!-- Segment of gesture annotation -->
   <region xml:id="r1" anchors="980 9190"/>
   <region xml:id="r2" anchors="980 993"/>
   <!-- Each anchor corresponds to a timeslot -->
   <node xml:id="a232">
    <link targets="r1"/>
   </node>
   <node xml:id="a233">
```

```
 <link targets="r2"/>
</node>
<a label="R Gesture Units 1" ref="a232"/>
<a label="preparation" ref="a233"/>
```

3.5 XML elements for the resource header

3.5.1 Specifications

The high level XML element structure for the resource header is as follows:

```
resourceHeader = fileDesc, encodingDesc, resourceDesc, revisionDesc
fileDesc       = titleStmt, editionStmt, extent, publicationStmt+
encodingDesc   = projectDesc, samplingDesc, editorialDecl,
                 classDecl
resourceDesc   = fileStruct, annotationSpaces, annotationDecls,
                      media, anchorTypes, groups?
revisionDesc   = change+
```

Elements in the resource header (required attributes in bold) are defined in Table 4.

Table 4 — Elements in the resource header

<resourceHeader>	Root element of the resource header.
Attributes	***@xmlns***="http://www.xces.org/ns/GrAF/1.0/"
	***@xmlns*:*xlink*=**http://www.w3.org/1999/xlink
	***@docID* [string]**: Identifier for this header.
	***@version* [string]**: Version of the header (not the resource to which it applies).
	@creator [string]: Identification of the person or entity responsible for creating the header.
	***@date.created* [string]**: Creation date in ISO 8601 format.
	@date.updated [string]: Most recent date of update in ISO 8601 format.
	***@type* [string]**: Type of resource (corpus, lexicon, etc.)
<fileDesc>	General description of the resource.
<titleStmt>	Title and related information.
<title>	Title of the resource.
<funder>	Entity or agency providing the funding for creation of the resource.
<respStmt>	Responsibility declarations.
<resp>	Entity or individual that created the resource.
Attribute	***@xlink*:*href* [URL]**: PID of the entity or individual that created the resource.
<editionStmt>	Practices followed in creating the resource.

Table 4 (*continued*)

Attribute	**@*version* [string]**: Version number of the current edition of the resource.
<extent>	Indicates the size of the resource.
Attributes	**@*count* [numeric]**: Size of the resource as a number of words, bytes, etc.
	@*unit* [string]: Word, byte, character, token, etc.
<publicationStmt>	Information about obtaining the resource.
<distributor>	Name of the resource distributor.
<pubAddress>	Street address of the resource distributor.
<phone>	Phone number of the resource distributor.
<fax>	Fax number of the resource distributor.
<eAddress>	Electronic address of the resource distributor.
Attribute	**@*type* [string]**: Type of address (email, URL, etc.).
<pubDate>	Date of publication of the resource.
Attribute	**@*iso8601* [string]**: Publication date in ISO 8601 format.
<availability>	Terms of availability of the resource, in terms of cost, licensing, etc.
Attributes	**@*xlink*:*href* [URL]**: PID where the resource may be obtained.
	@*status* [string]: General terms of availability, e.g. “free”, “restricted”, “licensed”.
	@*license* [URL]: PID of the license agreement for the resource.
<idno>	Number uniquely identifying the resource, e.g. ISBN, catalogue number.
Attribute	**@*type* [string]**: Type of the identifier, e.g. ISBN, LDC Catalog number.
<encodingDesc>	Encoding practices used to create the resource.
<projectDesc>	Prose description of the project that created the resource.
<samplingDecl>	Information concerning the choice of data included in the resource.
<editorialDecl>	Editorial practices applied in creating the resource.
<transduction>	Transduction practices applied to transform the data from its source format (e.g. PDF, recorded speech).
<correction>	Correction practices applied to the data in its source form, such as spelling correction or normalization of special characters.
<segmentation>	Unit(s) of segmentation applied to the data for the purposes of annotation.
<classDecl>	Declaration(s) of classifications applied to portions of the resource.
<taxonomy>	Set of categories used to classify parts of the resource, such as genre labels for text.
Attribute	**@*xml*:*id* [ID]**: Identifies the taxonomy. Used as a prefix to the category ID when multiple taxonomies are referenced from a document that is part of the resource.
<category>	Category used to classify components of the resource. Category elements may be nested to indicate sub-category relations.
Attribute	**@*xml*:*id* [ID]**: Identifier for the category.
<catDesc>	Prose description of the category.
<resourceDesc>	Objects and descriptors used in the resource.
<directories>	Directories (folders) appearing in this resource.
<directory>	Directory in this resource. This element may contain one or more nested <directory> elements that describe sub-directories of this directory.

Table 4 (*continued*)

Attributes	***@ xml*:*id* [ID]**: Unique identifier for the directory.
	***@d.name* [string]**: The directory name.
	@f.ids [IDREFS]: Space-delimited list of file IDs indicating the file types that may appear in this directory.
	@root [yes | no]: Indicates whether this directory is the root directory for the resource; if not specified, the default is *no*.
<d.desc>	Prose description of the directory contents.
<fileStruct>	File types included in the resource.
<fileType>	Description of a file type in terms of its contents.
Attributes	***@ xml*:*id* [ID]**: Unique identifier for the filetype.
	***@medium* [IDREF]**: Medium type as defined in the resource header.
	@f.suffix [string]: Suffix applied to filenames to identify the filetype. Note that this is appended to the filename itself, not applied as a file extension. File extensions may be defined on a <medium> element.
	@a.ids [IDREFS]: Type(s) of annotations included in this file type
	@required [yes | no]: Indicates if this file type is required to be present for each primary data document in the resource. Default is yes.
<requires>	File types required to process this file type.
Attribute	***@f.id* [IDREF]**: Identifier of file type to which this file type contains references.
<annotationSpaces>	Annotation spaces used in this resource.
<annotationSpace>	Defines a unique name that may be associated with an annotation to enable the disambiguation of homonym identifiers residing in different namespaces, or to identify a logical grouping of annotations to be retained when merged with other annotations of the same data.
Attributes	***@xml*:*id* [ID]**: The annotation space identifier.
	***@xlink*:*href* [URL]**: PID where the description of this annotation space is located, typically, the annotation scheme documentation or project site.
<annotationDecls>	Groups annotation declarations.
<annotationDecl>	Describes an annotation type, its provenance, documentation, etc.
Attributes	***@xml*:*id* [ID]**: ID identifying the annotation.
	***@as.id* [IDREF]**: Annotation space to which the annotation belongs.
<a.desc>	Prose description of the annotation space.
Attribute	*@xml:lang* [string]: ISO 639 code(s) for the language(s) of the description.
<a.resp>	Entity that produced the annotations.
Attribute	***@xlink*:*href* [URL]**: PID of description of the responsible entity.
<a.method>	Method by which the annotations were produced.
Attribute	***@type* [string]**: Method by which the annotations were produced. *Values*: manual, automatic, automatic-validated, other.
<a.doc>	External documentation for the annotation space.
Attribute	***@xlink*:*href* [URL]**: PID of documentation for the annotations
<a.schema>	Formal annotation schema for this annotation space. Multiple <a.schema> elements may be provided if several external schemas exist.

Table 4 (*continued*)

Attributes	***@xlink*:*href* [URL]**: PID of annotation schema.
	@type [string]*:* Schema type, e.g. fsDecl (feature structure declaration).
<media>	Groups declarations of media types used in the resource.
<medium>	Medium used in the resource.
Attributes	***@xml*:*id* [ID]**: Unique identifier for reference from annotation documents and document headers.
	***@type* [string]**: Description of the medium, e.g. text/plain, text/xml.
	***@encoding* [string]**: Encoding system used, e.g. UTF-8.
	***@extension* [string]**: File extension used in the resource to identify files containing data in this medium.
<anchorTypes>	Anchor types defined for the resource.
<anchorType>	Defines a type of anchor used to ground annotations in primary data, e.g. character-anchor, image-region, timestamp, xpath-expression.
Attributes	***@xml*:*id* [ID]**: Unique identifier for reference from annotation documents.
	***@medium* [IDREF]**: Medium to which this anchor type applies, e.g. text.
	@default [yes | no]: Indicates whether or not this anchor type is the default. If the attribute is not present, *no* is assumed.
	***@xlink*:*href* [URL]**: PID of a formal description of the anchor type.
<groups>	Definitions of groups of annotations; see 3.5.2 for details.
<group>	Logical group of annotations (e.g. layer, tier).
Attribute	***@xml*:*id* [ID]**: Name identifying the group.
<g.desc>	Prose description of the group.
Attribute	*@xml:lang* [string]: ISO 639 code(s) for the language(s) of the description.
<g.member>	Member of the group.
Attributes	***@type* [annotation | type | file | enumeration | expression]**: The method by which the group is identified.
	***@value* [string]**: a string representing an annotation label and/or an annotation space, an enumeration of annotation IDs, one or more fileType IDs, or an expression for navigating to annotations in the group, or the ID of another group.
	@xml:base [URL]: A URL or relative URL for the file containing the IDREFS in the *@value* attribute. Only applies if the value of the *@type* attribute is "enumeration".
<revisionDesc>	Documentation of revisions.
<change>	Information about a particular change made to the resource
<changeDate>	Date of the change in ISO 8601 format.
<respName>	Identification of the person responsible for the change.
<item>	Description of the change.

3.5.2 Groups

The <group> element can be used to define one or more groups of annotations that are to be regarded as a logical unit for any purpose. The most common use of groups is to associate annotations that represent a “layer” or “tier”, such as a morphosyntactic or syntactic layer. However, grouping can be applied to virtually any set of annotations. GrAF provides five types of grouping mechanisms.

a) *Annotation*: annotations with specific values for their labels (as given on the @*label* attribute of an <a> element in an annotation document) and/or annotation space. Wildcards may be used to select sets of annotations with common labels or annotation spaces, e.g. *:tok selects all annotations with the label *tok*, in any annotation space (designated by *:), xces:* selects all annotations in the xces annotation space.

b) *Type*: annotations of a specific type or types, by referencing the ID of an annotation declaration defined in the resource header.

c) *File*: annotations appearing in a specific file type or types, by referring to the ID of a file type defined in the resource header;

d) *Enumeration*: enumerated list of annotation IDs appearing in a specified annotation document.

e) *Expression*: path expression for navigating through annotations to find a specific value or values — for example, the expression @speaker='alice' would choose all annotations with a feature named *speaker* that has the value “Alice”.

f) *Group*: another group or set of groups. This can be used to group several enumeration groups whose IDs appear in different annotation documents.

EXAMPLE

```
<groups>
    <group xml:id = "g.token">
        <!-- all annotations in any annotation space with label "tok" -->
        <g.member value = "*:tok" type = "annotation"/>
    </group>
    <group xml:id = "g.example">
        <!-- all annotations of type logical -->
        <g.member value = "a.logical" type = "type"/>
        <!-- all files of containing entity annotations -->
        <g.member value = "f.entities" type = "file"/>
        <!-- all annotations with a feature "speaker" with value "Alice" -->
        <g.member value = "@speaker = 'alice'" type = "expression"/>
        <!-- annotations with ids "id_1" to "id_n" in file "myfile.xml"-->
        <g.member xml:base = "myfile.xml" value = "id1 id2 ... idN"
                        type = "enumeration"/>
        <!-- the annotations included in group g.token, defined earlier -->
        <g.member value = "g.token" type = "group"/>
    </group>
</groups>
```

3.6 Elements in the primary data document header

3.6.1 Specifications

The high-level XML element structure for the primary data document header is as follows:

```
documentHeader = fileDesc, profileDesc, revisionDesc
fileDesc       = titleStmt, extent, sourceDesc
profileDesc    = language, textClass, particDesc, settingDesc
dataDesc       = primaryData, annotations
revisionDesc   = change+
```

Elements in the primary data document header are described in Table 5 (required attributes in bold).

Table 5 — Elements in the primary data document header

<documentHeader>	Root element for primary data headers
Attributes	***@xmlns*=**"http://www.xces.org/ns/GrAF/1.0/"
	***@xmlns*:*xlink*=**http://www.w3.org/1999/xlink
	***@docID* [string]**: Identifier for this header.
	***@version* [string]**: Version of the header (not the resource to which it applies).
	@*creator* [string]: Identification of the person or entity responsible for creating the header.
	***@date.created* [string]**: Creation date in ISO 8601 format.
	@*date.updated* [string]: Most recent date of update in ISO 8601 format.
<fileDesc>	
<fileName>	Name of the file containing the primary data document in the resource. This name provides the base for filenames of associated annotation documents.
<extent>	Size of the resource.
Attributes	***@count* [numeric]**: Value expressing the size.
	***@unit* [string]**: Unit in which the size of the resource is expressed (words, bytes, tokens, etc.).
<sourceDesc>	Information about the source document.
<title>	Title of the primary data document.
<author>	Author of the primary data document.
Attribute	@*sex* [*male, female, unknown*]*:* Author's sex.
Attribute	@*age* [string]*:* Author's age.
<source>	Source from which the primary data was obtained.
Attribute	***@type* [string]**: Role or type the source with regard to the document (e.g. distributor, contributor, publisher, Web).
<distributor>	Distributor of the primary data (if different from source).
<publisher>	Publisher ("self" for individuals) of the source.
<pubAddress>	Address of publisher.
<eAddress>	Email address, URL, etc. of publisher.
Attribute	***@type* [string]**: Type of electronic address, such as email or URL.
<pubDate>	Date of original publication.

Table 5 (*continued*)

Attribute	*@iso8601* [string]: Date of publication in ISO 8601 format.
<idno>	Identification number for the document.
Attribute	***@type* [string]**: Type of the identification number (e.g. ISBN, LDC Catalog number).
<pubName>	Name of the publication in which the primary data was originally published (e.g. journal in which it appeared).
Attribute	*@type* [string]: Type of publication (e.g. journal, newspaper, anthology, electronic publication).
<documentation>	PID where documentation concerning the data may be found (e.g. description of recording procedures, participants, etc., for spoken data or corpus collection procedures).
<profileDesc>	
<langUsage>	
<language>	Language(s) of the primary data.
Attribute	*@iso639* [string]: ISO 639 code(s) for the language(s) of the primary data.
<textClass>	
Attribute	***@catRef* [IDREFS]**: One or more categories defined in the resource header.
<subject>	Topic of the primary data.
<domain>	Primary domain of the data.
<subdomain>	Subdomain of the data.
<particDesc>	Participants in an interaction.
<person>	One participant in an interaction.
Attributes	***@id* [ID]**: Identifier for reference from annotation documents.
	@age [numeric]: Age of the speaker.
	@role [string]: Role of the speaker in the discourse.
	@sex [string]: One of male, female, unknown.
<settingDesc>	The setting or settings within which a language interaction takes place, either as a prose description or a series of setting elements.
<setting>	Particular setting in which a language interaction takes place.
Attribute	*@who* [IDREFS]: Reference to person IDs involved in this interaction.
<time>	Time of the interaction.
<activity>	What a participant in a language interaction is doing other than speaking.
<locale>	Place of the interaction, e.g. a room, a restaurant, a park bench.
<dataDesc>	
<primaryData>	Provides the location of the primary data document.
Attributes	***@loc* [string]**: Relative path or PID of the primary data document.
	@loctype [*relative*, URL]: Indicates whether the primary data path is a fully specified path (PID) or a path relative to the location of this header file. Default is *relative.*
	***@f.id* [IDREF]**: File type via reference to definition in the resource header.
<annotations>	
<annotation>	Annotation document associated with the primary data document this header describes.

Table 5 (*continued*)

Attributes	**@*loc* [string]**: Relative path or PID of the annotation document.
	@*loctype* [*relative*, URL]: Indicates whether the path is a fully specified path (PID) or a path relative to this header file. Default is *relative.*
	@*f.id* [IDREF]: File type via reference to definition in the resource header.
<revisionDesc>	Documentation of revisions.
<change>	Information about a particular change made to the document.
<changeDate>	Date of the change in ISO 8601 format.
<respName>	Identification of the person responsible for the change.
<item>	Description of the change.

Bibliography

[1] ISO 639-1:2002, *Code for the representation of names of languages — Part 1: Alpha-2 code*

[2] ISO 639-2:1998, *Code for the representation of names and languages — Part 2: Alpha-3 code*

[3] ISO/IEC 646:1991, *Information technology — ISO 7-bit coded character set for information interchange*

[4] ISO 3166-1:1997, *Code for the representation of names of countries and their subdivisions — Part 1: Country codes*

[5] ISO 8601:1988, *Data elements and interchange formats — Information interchange — Representation of dates and times*

[6] ISO 8879:1986, *Information processing — Text and office systems — Standard Generalized Markup Language (SGML)*. As extended by TC 2 (ISO/IEC JTC 1/SC 34 N 029:1998-12-06) to allow for XML

[7] ISO/IEC 10646-1:1993, *Information technology — Universal Multiple-Octet Coded Character Set (UCS) — Part 1: Architecture and basic multilingual plane*

[8] ISO 24610-1:2006, *Language resource management — Feature structures — Part 1: Feature structure representation (FSR)*

[9] ISO 12620:2009, *Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources*

[10] IDE, N. and ROMARY, L. (2004). International Standard for a Linguistic Annotation Framework

[11] IDE, N. and SUDERMAN, K. (2006). Merging Layered Annotations. In *Proceedings of Merging and Layering Linguistic Information Workshop*, held in conjunction with LREC 2006, Genoa

[12] IDE, N. and SUDERMAN, K. (2007). GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of the Linguistic Annotation Workshop*, held in conjunction with ACL 2007. Prague, pp. 1-8

[13] IDE, N. and BUNT, H. (2010). Anatomy of Annotation Schemes: Mappings to GrAF. In *Proceedings of LAW-IV: the Fourth Linguistic Annotation Workshop*, Uppsala, pp. 115-124

[14] IDE, N. and SUDERMAN, K. Bridging the Gaps: Interoperability for Language Engineering Architectures Using GrAF. *Language Resources and Evaluation*, Selected Papers from the Third Linguistic Annotation Workshop, Stede, M., and Huang, C.-R. (eds.)[1)]

[15] IDE, N. and SUDERMAN, K. The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging. *Language Resources and Evaluation*, Special issue on Standards for Language Resources and Language Technologies, Budin, G., Romary, L., Wright, S.-E. (eds.)[2)]

1) To be published.

ICS 01.020
Price based on 19 pages

This page deliberately left blank

British Standards Institution (BSI)

BSI is the national body responsible for preparing British Standards and other standards-related publications, information and services.

BSI is incorporated by Royal Charter. British Standards and other standardization products are published by BSI Standards Limited.

About us

We bring together business, industry, government, consumers, innovators and others to shape their combined experience and expertise into standards -based solutions.

The knowledge embodied in our standards has been carefully assembled in a dependable format and refined through our open consultation process. Organizations of all sizes and across all sectors choose standards to help them achieve their goals.

Information on standards

We can provide you with the knowledge that your organization needs to succeed. Find out more about British Standards by visiting our website at bsigroup.com/standards or contacting our Customer Services team or Knowledge Centre.

Buying standards

You can buy and download PDF versions of BSI publications, including British and adopted European and international standards, through our website at bsigroup.com/shop, where hard copies can also be purchased.

If you need international and foreign standards from other Standards Development Organizations, hard copies can be ordered from our Customer Services team.

Subscriptions

Our range of subscription services are designed to make using standards easier for you. For further information on our subscription products go to bsigroup.com/subscriptions.

With **British Standards Online (BSOL)** you'll have instant access to over 55,000 British and adopted European and international standards from your desktop. It's available 24/7 and is refreshed daily so you'll always be up to date.

You can keep in touch with standards developments and receive substantial discounts on the purchase price of standards, both in single copy and subscription format, by becoming a **BSI Subscribing Member**.

PLUS is an updating service exclusive to BSI Subscribing Members. You will automatically receive the latest hard copy of your standards when they're revised or replaced.

To find out more about becoming a BSI Subscribing Member and the benefits of membership, please visit bsigroup.com/shop.

With a **Multi-User Network Licence (MUNL)** you are able to host standards publications on your intranet. Licences can cover as few or as many users as you wish. With updates supplied as soon as they're available, you can be sure your documentation is current. For further information, email bsmusales@bsigroup.com.

BSI Group Headquarters

389 Chiswick High Road London W4 4AL UK

Revisions

Our British Standards and other publications are updated by amendment or revision.

We continually improve the quality of our products and services to benefit your business. If you find an inaccuracy or ambiguity within a British Standard or other BSI publication please inform the Knowledge Centre.

Copyright

Useful Contacts:

Customer Services
Tel: +44 845 086 9001
Email (orders): orders@bsigroup.com
Email (enquiries): cservices@bsigroup.com

Subscriptions
Tel: +44 845 086 9001
Email: subscriptions@bsigroup.com

Knowledge Centre
Tel: +44 20 8996 7004
Email: knowledgecentre@bsigroup.com

Copyright & Licensing
Tel: +44 20 8996 7070
Email: copyright@bsigroup.com

bsi.

...making excellence a habit.™