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 iv
Introduction v
1 Scope 1
2 Terms and definitions 1
3 LAF specification 3
3.1 Overview 3
3.2 LAF data model 3
3.3 LAF architecture 4
3.4 XML pivot format 6
3.5 XML elements for the resource header 11
3.6 Elements in the primary data document header 16
Bibliography 19

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