INTERNATIONAL STANDARD

ISO 24613

First edition
2008-11-15

Language resource management — Lexical markup framework (LMF)

Gestion de ressources langagières — Cadre de balisage lexical (LMF)

Reference number
ISO 24613:2008(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 24613 was prepared by Technical Committee ISO/TC 37, *Terminology and other language and content resources*, Subcommittee SC 4, *Language resource management*.

ISO 24613 is designed to coordinate closely with ISO 12620, *Terminology and other content and language resources — Data categories — Specification of data categories and management of a Data Category Registry for language resources*[1)], and ISO 16642, *Computer applications in terminology — Terminological markup framework*.

1) To be published. (Revision of ISO 12620:1999)

Introduction

Optimizing the production, maintenance and extension of electronic lexical resources is one of the crucial aspects impacting human language technologies (HLT) in general and natural language processing (NLP) in particular, as well as human-oriented translation technologies. A second crucial aspect involves optimizing the process leading to their integration in applications. Lexical Markup Framework (LMF) is an abstract metamodel that provides a common, standardized framework for the construction of computational lexicons. LMF ensures the encoding of linguistic information in a way that enables reusability in different applications and for different tasks. LMF provides a common, shared representation of lexical objects, including morphological, syntactic and semantic aspects.

The goals of LMF are to provide a common model for the creation and use of electronic lexical resources ranging from small to large in scale, to manage the exchange of data between and among these resources, and to facilitate the merging of large numbers of different individual electronic resources to form extensive global electronic resources. The ultimate goal of LMF is to create a modular structure that will facilitate true content interoperability across all aspects of electronic lexical resources.

The LMF core package describes the basic hierarchy of information of a lexical entry, including information on the form. The core package is supplemented by various resources that are part of the definition of LMF. These resources include:

— specific data categories used by the variety of resource types associated with LMF, both those data categories relevant to the metamodel itself, and those associated with the extensions to the core package;

— the constraints governing the relationship of these data categories to the metamodel and to its extensions;

— standard procedures for expressing these categories and thus for anchoring them on the structural skeleton of LMF and relating them to the respective extension models;

— the vocabularies used by LMF to express related informational objects for describing how to extend LMF through linkage to a variety of specific resources (extensions) and methods for analysing and designing such linked systems.

Extensions of the core package which are documented in the annexes of this International Standard include:

a) machine readable dictionaries;

b) natural language processing lexical resources.

LMF extensions are expressed in a framework that describes the reuse of the LMF core components (such as structures, data categories, and vocabularies) in conjunction with the additional components required for a specific resource.

Types of individual instantiations of LMF can include such electronic lexical resources as fairly simple lexical databases, NLP and machine-translation lexicons, as well as electronic monolingual, bilingual and multilingual lexical databases. LMF provides general structures and mechanisms for analysing and designing new electronic lexical resources, but LMF does not specify the structures, data constraints and vocabularies to be used in the design of specific electronic lexical resources. LMF also provides mechanisms for analysing and describing existing resources using a common descriptive framework. For the purpose of both designing new lexical resources and describing existing lexical resources, LMF defines the conditions that allow the data expressed in any one lexical resource to be mapped to the LMF framework, and thus provides an intermediate format for lexical data exchange.

Language resource management — Lexical markup framework (LMF)

1 Scope

This International Standard describes the Lexical Markup Framework (LMF), a metamodel for representing data in lexical databases used with monolingual and multilingual computer applications.

LMF provides mechanisms that allow the development and integration of a variety of electronic lexical resource types[2]. These mechanisms will present existing lexicons as far as possible. If this is impossible, problematic information will be identified and isolated.

2 Normative references

The following referenced documents are indispensable for the application 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 639 (all parts), *Codes for the representation of names of languages*

ISO 1087-1, *Terminology work — Vocabulary — Part 1: Theory and application*

ISO 1087-2, *Terminology work — Vocabulary — Part 2: Computer applications*

ISO 12620, *Terminology and other content and language resources — Data categories — Specification of data categories and management of a Data Category Registry for language resources* [3]

ISO 15924, *Information and documentation — Code for the representation of names of scripts*

3 Terms and definitions

For the purposes of this International Standard, the terms and definitions given in ISO 1087-1, ISO 1087-2 and the following apply [4].

3.1
abbreviated form
form (3.14) resulting from the omission of any part of the **full form** (3.16) of the same **lexeme** (3.25)

2) LMF supports existing lexical resource models such as the Genelex [9], the EAGLES International Standards for Language Engineering (ISLE) [5] and Multilingual ISLE Lexical Entry (MILE) models [6].

3) To be published. (Revision of ISO 12620:1999)

4) It is worth noting that we have purposely avoided defining and using highly controversial terms such as “word”, “morpheme”, “base”, “fusion”, “ergative”, “paradigm”, and “collocation”.