May 2009

Relevance

CLARIN community

for all communities

Standards for Text Encoding

Abstract

In this CLARIN ShortGuide to Standards for Text Encoding we give an overview of the current state of general standards for text encoding, such as UNICODE, XML, Character Encoding, etc. as well as standards specific to textual resources, such as the ISO standards established by the TC37 SC4 committee, the Text Encoding Intiative (TEI) and the Consortium developing the Corpus Encoding Standard for XML (XCES).

Why are they important?

CLARIN wants to develop a common, pan-European infrastructure for language resources, tools and services for use in the humanities. A key requirement of such a common infrastructure is the ability to exchange information across different resources and across different services that extract linguistic information or annotate language resources at different analysis levels. In short, the CLARIN infrastructure needs to be interoperable. This, in turn, presupposes that language resources and tools need to utilize, as much as possible, common data formats that are compliant with encoding standards and best practises.

Apart from multi-modal and spoken materials, textual material plays a central role in such an eHumanities infrastructure. This overview of existing standards for text encoding and their current state of development is neiter an exhaustive list nor a statement about the quality of the discussed standards, but a snapshot of current activities, of the standards in common use and of best practices in the encoding and interchange of textual language resources (text corpora, lexica, etc.). For information about the important issue of metadata standards please refer to the CLARIN documents dedicated exclusively to this issue available at the CLARIN website: on http://www.clarin.eu. / documents/short-guides/.

How will CLARIN promote standards?

CLARIN will inform its members, but also the larger user community of language resources and tools about pertinent standards and best practices and about current developments in relevant standardization.

CLARIN will rely as much as possible on existing standards, guidelines and best practices. Several CLARIN partners have been long-standing members of standards organizations so that there is easy information flow between these activities and the CLARIN membership. Apart from officially recognized standards, CLARIN will also track de-facto standards and best-practise encodings for those areas of language resources and tools for which no published standards are available yet. For standards that pertain to neighboring disciplines in the humanities, CLARIN will seek active collaboration with standards organizations and initiatives in these fields.

It is important that the specification of standards is easily accessible so that developers of language resources and tools can refer to them and, if appropriate, give feedback to the relevant standards committees. Ideally, existing standards are already followed by a large community of users. For these well-established standards, CLARIN will ensure that data conversions tools are available.

How can they be used?

Standards serve different functions in the creation or curation of language resources and tools:

(1) For existing resources and tools with often heterogeneous data formats, standards can be used for the specification of interchange (pivot) formats. This will facilitate the development of conversion tools that translate heterogeneous data formats

into such pivot formats and back. The availability of such pivot formats will greatly simplify the specification and implementation of flexible processing chains for language resources.

(2) For the development of new resources and tools, standards should be followed as much as possible. Creators of new corpora and lexical resources, for example, should consult the ISO standards pertaining to language as well the TEI and XCES guidelines on text encoding. Adherence to such established or emerging standards will minimize the need for data conversion in the first place.

(3) Other standards may guide the interaction and exchange of resources and tools that have been created by other initiatives world-wide. Through the process of such international collaborations, best practices will continue to emerge that are empirically validated and that are broadly accepted by the scientific community.

How does it work?

International standardization committees, such as the ISO TC37/SC4 Language Resource Management, the W3C Consortium, and other standardization bodies play a central role in developing guidelines for the processing, storage and modelling of language resources to facilitate reusability, merging and comparison of linguistic information independent of the language used. Please consult the Annex for more details.

The ISO Process

A standard developed by the International Standardization Organization (ISO) goes through six stages from the first proposal to the final publication of the International Standard. Initial approval of a new work item leads to the formation of a technical committee (TC) and subcommittees (SC) for specific work items. These expert groups prepare working drafts (WD) which- if approved by the relevant TC/SC members- will lead to the registration of an approved work item (AWI, Preparatory stage). The ISO registration process then starts with the first Committee Draft (CD), which will be distributed in form of a draft International Standard (DIS) to all ISO member bodies for commenting and voting (Committee stage). Once approved for submission as a Finalized Draft International Standard (FDIS, Enquiry stage), it has to pass a final vote by all ISO members (Approval stage) to become an ISO Standard. The new International Standard will be published by the ISO Central Secretariat (Publication stage).

The Text Encoding Initiative

The Text Encoding Initiative (TEI) Consortium is a non-profit organization supported by the academic community. It develops and maintains guidelines for the representation of digital text resources in form of XML encoding schemes as published along with documentation on the the TEI web site (http://www.tei-c.org). The TEI guidelines define close to 500 elements arranged in modules according to the kind of information or the resource type. Apart from the specific modules, TEI P5 also supplies general modules, such as the TEI header for metadata about the resource, and generic elements common to all kinds of texts.

XML Corpus Encoding Standard (XCES)

XCES is the XML version of the Corpus Encoding Standard and instantiates the EAGLES Corpus Encoding Standard (CES) DTDs for linguistic corpora. CES was developed by the Export Advisory Group on Language Engineering Standards (EAGLES) as a set of encoding standards for corpus-based work in natural language processing applications. Documentation is available under www.cs.vassar.edu/CES/.

Who is responsible?

CLARIN Work Package 5
Erhard Hinrichs
Tuebingen University
Seminar fuer Sprachwissenschaft,
Abt. Computerlinguistik
Wilhelmstr. 19, 72074 Tuebingen, Germany
Website: www.clarin.eu
Email: eh@sfs.uni-tuebingen.de

The SfS at the University of Tuebingen is leading work package 5 of CLARIN and will implement this procedure and is for the moment acting as the central contact for questions about standards in text encoding.

Other Available ShortGuides

at www.clarin.eu/documents/short-guides

» CLARIN Centers

» Concept Registry Service

» Component Metadata

» Persistent Identifier Service

» Long-term Preservation and Access

» Standards and Best Practices

» Trust Domain

» Virtual Collections

The last version of this document is maintained at the CLARIN Web-Site under documents: www.clarin.eu/documents

Annex: Standards in Text Encoding

Standard	Description	Responsibility	Current Standard
Unicode	Universal Character Set consisting of a repertoire of ca. 100.000 characters	ISO 10646 Working Group (SC 2/WG 2), Unicode Consortium	Unicode 5.1.0
Language Codes	Codes for the representation of names of languages	TC 37/SC 2/WG 1	ISO 639-x, etc.
Character Encoding	Coded Character Set	JTC 1/SC 2	ISO/IEC 8859-X, etc.
XML	Extensible Markup Language SGML (ISO 8879) extended by TC2 ISO/IEC JTC 1/SC 34 N 029:1998-12-06	W3C XML Working Groups	XML 1.0 XML 1.1
TermLR	Terminology for Language Resource Management	ISO/TC 37/SC 4 WG1	ISO/WD 21829
LMF	Lexical Markup Framework	ISO/TC 37/SC 4/WG 4	ISO 24613:2008
LAF	Linguistic Annotation Framework	ISO/TC 37/SC 4/WG 1	ISO/DIS 24612
MAF	Morpho-Syntactic Annotation Framework	ISO/TC 37/SC 4 WG2	ISO/DIS 24611
FSR	Feature Structure Representation	ISO/TC 37/SC 4 WG1	ISO/24610-1:2006
FSD	Feature Structure Description	ISO/TC 37/SC 4 WG1	ISO/DIS 24610-2
WordSeg1	Word segmentation of written texts for mono-lingual and multi-lingual information processing — Part 1: Basic concepts and general principles	ISO/TC 37/SC 4 WG2	ISO/DIS 24614-1
WordSeg2	Word segmentation: Part 2 Chinese, Japanese, Korean	ISO/TC 37/SC 4 WG 2	ISO/WD 24614-2
SemAF	Semantic Annotation Framework	ISO/TC 37/SC 4 WG 2	ISO/DIS 24617-1 Part 1: Time and events
SynAF	Syntactic Annotation Framework	ISO/TC 37/SC 4/WG 2	ISO/DIS 24615
MLIF	Multilingual Information Framework	ISO/TC37/SC 4/ WG3	ISO/WD 24616
DCR	Data Category Registry	ISO/TC 37/SC 4 WG1	ISO 12620:1999 ISO/DIS 12620.2
TEI	Text Encoding Initiative	TEI Consortium	TEI P5 Guidelines (1)
XCES	XML Instantiation of the Corpus Encoding Standard	XCES Schema /DTD /namespaces released	http://www.xces.org/schema/2003 (2)

(1) TEI: The most relevant chapters for text corpus encoding are:

"The TEI Header"

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

"Linking, Segmentation, and Alignment"

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html

"Simple Analytic Mechanisms"

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AI.html

"Feature Structures"

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html

"Default Text Structure"

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html

"Language Corpora"

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html

The TEI-roma tool (http://www.tei-c.org/Roma/) can be used to create or validate TEI customizations or to generate schemata from existing customizations.

(2) There is no XCES documentation available yet, but examples of projects using the XCES Standards and documentation of their implementation can be found under:

ANC - American National Corpus

http://americannationalcorpus.org/SecondRelease/encoding.html

IPI PAN Corpus at the Polish Academy of Science

http://korpus.pl/index.php?lang=en&page=download

Written corpora of the Institut für Deutsche Sprache in Mannheim

http://www.ids-mannheim.de/kl/projekte/korpora/textmodell.html.