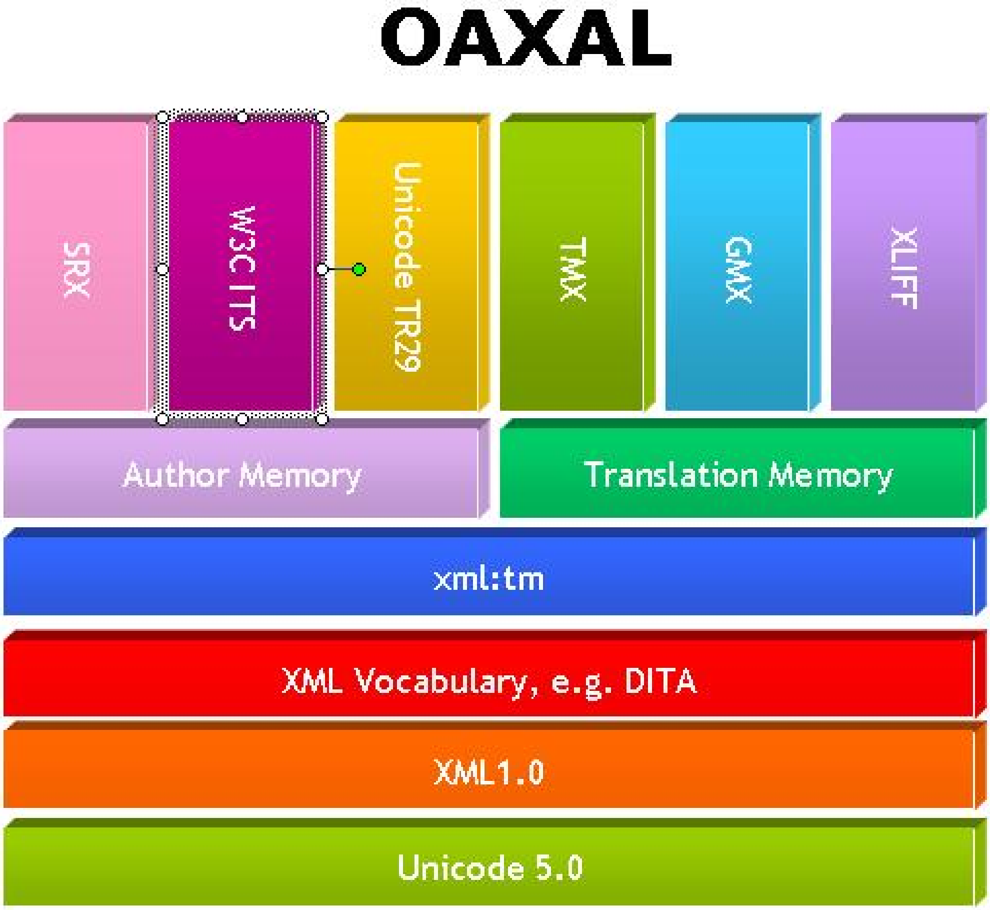
LREC 2008 Tutorial

LISA Standards

Standards and Guidelines for Multilingual Text Processing

Standards play an important role in today's localization and translation industry. By providing agreed upon ways of representing linguistic data they allow various applications to interoperate and share data. The Localization Industry Standards Association (LISA, http://www.lisa.org) has been particularly active in development of standards for linguistic data, through the LISA OSCAR group and other parts of LISA.. This tutorial will describe various LISA standards and specifications and their relationship to other standards initiatives.

One way to relate LISA standards to a bigger picture of multilingual text processing is through a cluster of standards called OAXAL (Open Architecture for XML Authoring and Localization) and through a set of standards relating to translation quality assurance.. Below is a figure that shows the various standards that make up OAXAL, followed by a prose description of how they fit together.



- In **OAXAL**, a source text (assumed to be in marked up using some XML vocabulary with Unicode as the encoding option) is authored with translation (or localization) in mind, rather than treating authoring and translation as entirely separate processes. The XML schema (whether it uses the DTD, XSD, or RNG schema definition language) to which the source text conforms should use ITS (a standard from the W3C - http://www.w3.org/TR/its/) to facilitate translation. Preferably, though not necessarily, the source text is segmented using xml:tm namespace elements added in to the source text (xml:tm is a LISA OSCAR standard). Among other things, xml:tm allows an embedded record of revisions of source and target text segments. During authoring and during translation, various memory tools allow re-use of previous authored segments and their translations. Terminology lookup is also an important aspect of authoring and translation. TBX (a LISA OSCAR standard) is useful for transferring terminological data between software systems. Some translation tools accept XLIFF files as input (XLIFF is an OASIS standard - http://docs.oasis-open.org/xliff/v1.2/cs02/xliff-core.html). A source text, without or without xml:tm segmentation, can be converted to XLIFF format if appropriate. XLIFF is an XML markup language (that the representation of a b-text (a text and its translation, segmented and aligned). The source text and the target text each have a length, which is typically measured in characters or words. GMX-V (another LISA OSCAR standard), which is based on Unicode TR-29, defines a uniform method of measuring and representing the length (also called "volume") of a text. When the text has been translated, the resulting bi-text is often converted to a traditional translation memory (TM) in which the translation units (a segment of source text paired with a segment of target text) are separated from the context of their adjoining sentences and indexed for retrieval. The TMX and SRX standards facilitate the exchange of translation memory data among various tools.

- **xml text memory** (xml:tm) allows revision and translation memory to be stored directly in XML documents via the XML namespace mechanism, thus making linguistic assets available to any process that requires them. A relatively new standard, xml:tm interfaces with DITA, TMX, and other standards to allow for integration of multilingual concerns during the authoring and revision phases. See http://www.lisa.org/XML-Text-Memory-xml.107.0.html for more information on xml:tm.

- **Translation Memory eXchange** (TMX) is the industry standard for the exchange of translation memory databases. LISA research has shown that some organizations have millions of dollars invested in translation memory data. This investment, coupled with ongoing consolidation and changes in the lineup of TM vendors and products, makes the ability to reuse TM data and exchange TM databases of vital importance if these assets are to maintain their value. See http://www.lisa.org/Translation-Memory-e.34.0.html for more information on TMX.

- **Segmentation Rules eXchange** (SRX) provides a regular expression-based XML formalism to describe how text is split into “segments” (such as sentences or paragraphs). When coupled with TMX, it allows linguistic tools to describe how they segmented text during the translation process so that other tools can emulate this behavior in order to better use text produced by them. It also has potential use in any field that requires text segmentation. Planned submission to ISO will focus on allowing national bodies to define default segmentation rules sets for their languages in order to better facilitate use of linguistic tools with those languages. See http://www.lisa.org/Segmentation-Rules-e.40.0.html for more information on SRX.

- **Term-Base eXchange** (TBX) is the industry standard for representing terminology databases in XML format. Terminology is a key component to achieving quality translations (LISA research has demonstrated that terminology errors result in negative impression of product quality on par with seemingly more serious functional errors), yet many organizations do not manage terminology or use simple tools like Microsoft Excel to store their terminology. TBX, which is now a joint work item between LISA and ISO (ISO 30042), simplifies the use and dissemination of terminological data and helps organizations leverage these assets during the translation process. See http://www.lisa.org/Term-Base-eXchange.32.0.html for more information on TBX.

- **Global information metrics Management eXchange - Volume** (GMX-V) is the first part of a proposed tripartite standard for representation of metadata about localization and translation projects. GMX-V specifically addresses the need for consistent, verifiable, and cross-platform word and character counts. In extreme cases the word counts provided by various text-processing applications can vary by as much as 30%, making any estimations of work or cost uncertain and contingent upon the tool being used. While such extreme variations are not normal, smaller variations can result in substantial cost differentials if two parties agreed to translation or other related tasks based on different assumptions about volume. While GMX-V will not replace application-specific word counts that are suitable for specific tasks, it does provide a mechanism for parties to agree upon costs up front with certainty. Forthcoming portions of the GMX standard will focus on textual complexity (e.g., grammatical and lexical complexity) and pre-negotiated quality requirements (in coordination with ISO Technical Committee 37 projects).

Translation Quality Assurance

- Quality Assurance (QA) has traditionally been difficult in the language industries because evaluations are subjective and subject to personal preference. Thus one individual might find nothing objectionable in a translation while another might consider it very bad. With support for a variety of quality metrics (including SAE J2450 and custom error profiles), the LISA QA Model provides a way for more objective quality decisions to be made. See http://www.lisa.org/LISA-QA-Model-3-1.124.0.html for more information on the LISA QA system. The tutorial will provide background information on translation quality assurance by discussing the American and European translation quality standards and the ISO project to complement various national and regional translation quality standards through a common set of translation parameters that provide a framework for creating project specifications, which are in turn the basis for a quality metric.

Conclusion

A standards-based workflow that includes LISA standards can offer significant technical and business benefits to implementers by reducing dependence on specific tool or services vendors and allowing choice in linguistic tools and processes to be made based on competitive differences rather than on technical lock-in. They play an important role in the development of tools for dealing with multilingual text, but also address areas that impact any individual or organization interested in processing text: segmentation of text, storing of text histories in XML, word counts, etc.

Tutorial Presenter

Alan K. Melby

Brigham Young University, Provo
Board Member, American Translators Association (ATA)
Member, LISA Open Standards for Container/content Allowing Reuse (OSCAR) standards steering committee
email: melbyak@yahoo.com
web: http://www.ttt.org

Alan K. Melby has worked in the translation/localization industry since the 1970s, starting in machine translation, switching in the 1980s to productivity tools for human translators. In the 1990s his focus shifted to work on translation-related standards. He holds a PhD in computational linguistics from Brigham Young University under the direction of William J. Strong, director of the Acoustics Research Group. He received the Eugen Wüster Prize for contributions to the field of terminology in 2007.

Tutorial Contributors

Arle R. Lommel
Chair, Open Standards for Container/content Allowing Reuse (OSCAR) standards committee, The Localization Industry Standards Association
LISA delegate to ISO TC 37 and TC 46
email: arle@lisa.org
web: http://www.lisa.org

Kara Warburton
Head Terminologist, IBM
Member of ISO TC37, SC3
Member of LISA Terminology Special Interest Group