ECP-2007-LANG-617001

FLaReNet

Deliverable D4.1

Identification of problems in the use of LR standards and of standardization needs

Deliverable number/name	*D4.1 – Identification of problems in the use of LR standards and of standardization needs*
Dissemination level	*Public*
Delivery date	*30 September 2009*
Status	*Final*
Author(s)	*Gerhard Budin*

*e*Content*plus*

This project is funded under the *e*Content*plus* programme[1], a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

1 OJ L 79, 24.3.2005, p. 1.

Table of Contents

1 Introduction

The purpose of this document is to identify problems that are encountered when Language Resource (LR) standards are used and to identify and describe needs for such LR standards.

In order to reach this goal, it is unavoidable to first identify existing LR standards focusing on their scopes and application areas in order to subsequently assess their actual use by different user groups and the problems they have encountered in this. These problems need to be analysed and conclusions have to be drawn for further work (e.g. for improving, updating, extending, modifying, etc. such standards). From this analysis an assessment is derived concerning the lack of LR standards for different user groups and application areas. The empirical input to this report has been coming from most diverse sources, ranging from previous documents on language technology standards, reports from workshops and conferences on this or similar topics (including the proceedings of the FlaReNet conference held in February 2009 in Vienna), discussions with representatives from various user groups listed below, up to an analysis of the actual use of LR standards and the (more or less) obvious and visible problems that are encountered. In addition, input has also been gathered from relevant standards bodies or organizations that include the creation and/or promotion of LR standards, such as ISO (not only TC 37, but also many other TCs), CEN (incl. CEN/ISSS), W3C, TEI, CLARIN (esp. WP 5), OMG, OASIS, LISA, translators organizations, domain-specific organizations (e.g in eHealth), etc. Another important dimension in this analysis is that of language. Nowadays we see a few languages that are relatively well equipped with language resources and thus with LR standards, while many other languages are relatively "under-resourced" both in terms of available language resources (text corpora, terminologies, tree banks, etc.) and LR standards needed to use or operate such standards. An additional problem is that best practices are not always known to others and that concrete examples of using LR standards are not documented in such a way that they can be re-used.

2 Scope of LR standards and user group categories

Given the broad scope of the FlaReNet project, it is obvious to define the scope of LR standards as broad.

As main user communities or user groups we identify the following professional activities (in a non-exhaustive list):

- Translators/interpreters
- Operators/managers of computer-assisted translation environments
- Operators of machine translation systems
- Localizers/localization managers
- Operators of localization management systems
- Technical writers
- Lexicographers, editors and publishers of dictionaries

- Terminology managers
- Corpus linguists, computational linguists, and language engineers in industrial production and/or research environments
- Other user groups and professional activities

What is referred to above as "industrial production and/or research environments" covers a broad spectrum of branches of industry such as information technologies/computer industries, automation/robotics, telecommunications, data mining, information retrieval, etc., and basically all economic sectors that are systematically supported by information technologies, increasingly referred to as eCommerce, eHealth, eLearning, eGovernment, eEnvironment, etc. In all these sectors we find language resources being produced and used, and thus LR standards becoming relevant for all user communities in these fields. This group "corpus linguists, computational linguists, and language engineers" also includes the specific group of technology providers, i.e. developers of machine translation systems, developers of any other language resource processing software, compilers of electronic language resources (i.e. other than "classical" dictionaries and terminology databases, that is to say developers of tree banks, electronic lexica, large-scale text corpora, etc.).

It also should be noted that all other user groups listed above (i.e. other than "corpus linguists, computational linguists, and language engineers") are also found in all sectors of the economy referred to in the previous paragraph, but with distinctively different roles and perspectives on language resources.

The user groups listed above are very heterogeneous with respect to their professional cultures and traditions, their degree of computer literacy, their ways of using language resources and LR standards, their application specific needs and requirements to LR standards.

The language dimension mentioned above is horizontal to the domain dimension, i.e. domain-specific language resources are developed and/or needed for many different languages, not only in English. By the same token, a language community needs language resources in many different or all domains of society. Increasingly, with a growing impact of globalization, we need both, LR in all languages and in all domains, thus the need for further developing LR standards is rapidly growing.

3 LR standards

The most straightforward way to identify the problems that user groups encounter in using specific LR standards is to list each of them that is considered relevant under the scope described above and describe the problems that have been reported to us.

At this point of the project it is considered more useful in this report D 4.1 to pinpoint at a set of LR standards that represent major types of standards and that are used in selected, but typical user groups in order to focus on the nature of the problems and the types of problems encountered. Therefore we proceed from the most specific LR

standards to the most generic ones (i.e. foundational standards that are not LR standards as such but that are also used in such contexts and that build the basis for language-specific standards (typically XML, RDF).

Many LR standards are in the process of development or are being revised at the moment. Thus it is not easy, for a number of reasons, to clearly separate problems encountered with existing standards from standards needs that have not yet been covered. In order to organise this document in a most simple and more readable way we group those standards together, but for each standard we mention their stage of development and we focus on problems that may be derived from this temporal perspective.

3.1 LR-specific standards for linguistic data processing

For academic purposes the TEI Guidelines (current version P5) has been a well established and widely used resource of LR-specific standards mainly for corpus analysis, markup and annotation. But TEI is hardly known in industrial communities (with a few exceptions) and completely foreign to professional groups such as localizers and translators. We see great potential in using TEI Guidelines in industrial contexts.

National corpus encoding and research work has led to a number of formats and tagsets that are widely used and have become kind of de facto standards, only to mention 3 of them here:

- XCES is an XML based corpus format that is widely used to create text corpora with multilevel annotations (US)
- TIGER is a flexible format to represent syntactical annotations on texts (Germany)
- Penn TreeBank is a format to represent syntactical annotations on texts (US)

It is felt that such results should flow into global standardization efforts. On the basis of many years of pre-normative research in different projects and initiatives such as EAGLES, LIRICS, etc., ISO TC 37 (originally for terminology, see under 3.3) had widened its scope to include language resource management and created a new sub-committee, no. 4, to turn these results into ISO standards. This is a long-term strategy and will bear industrial fruit only in a few years from now, while at the moment the testing and consolidation phase is in full swing: TC 37 SC 4 is very international, bringing together the above mentioned US-experts with their standards and best practices with the European traditions of EAGLES etc. and with East Asian best practices in the field.

ISO/TC 37/SC 4 has elaborated (together with TEI and other academic communities) a standard on the representation of feature structures: ISO 24610-1:2006 Language resource management -- Feature structures -- Part 1: Feature structure representation (FSR). Some components (e.g. to express constraints and concrete feature values) are not yet finished or not yet part of the standard. It will take quite a while until enough convincing test cases in industrial environments have been collected to show that this standard should be used in non-academic production environments.

As a result of the LIRICS project, LMF was published as ISO 24613 in 2008, called Lexical Markup Framework. The need for such a standard was seen long time ago in order to overcome fragmentation in terms of developing and using categories and value sets for lexical attributes and models of lexical structures. LMF is new and not yet widely adopted. It has an abstract style understandable for computational linguists only. So this is the target group of this document. It is open whether language industry will adopt it for their corporate operations on lexical resources. The standard certainly needs real-life examples from different early adopters and showcases of how to use it and what is the return on investment on its adoption in existing work environments. There are a number of open issues to be resolved (e.g. representation of ontologies), modelling constraints and their representation.

In the draft standardization action plan in CLARIN there are, in addition to the ones mentioned above, further standards projects grouped under the heading „Standards for Structural Interoperability":

The Morpho-syntactic Annotation Framework (MAF) (current stage ISO DIS 24611) contains a model as well as a format for the representation of morpho-syntactic annotation for tokens and word forms. It provides a tag set based on feature structure libraries. The document is still immature in several respects. At the moment it is more an empty shell that needs to be filled with data, i.e. morpho-syntactic tags actually used. It will take years until critical mass of data will be available for large-scale industrial applications.

Similarly, the Syntactic Annotation Framework (SynAF) (ISO/CD 24615) is also very much under development and will take years to come of age. It provides a generic model for representing both constituent and dependency based syntactic annotation and has been inspired by initiatives like Tiger which is very close to SynAF. A lot of work still needs to be done before any industrial application becomes feasible.

There are several other work items in TC 37 SC 4, such as one on word segmentation, in particular for East Asian languages such as Chinese, Korean, Japanese, a burning need not only in academic corpus analysis projects but also in industrial large-scale applications. This standard is expected to become applicable also in industrial environments rather soon as a wide community of research teams (academic and non-academic) are working out solutions for this burning problem.

In addition to further annotation standards (e.g. semantic annotation) that have been embarked upon in TC 37 SC 4, there is a new project called MLIF – Multilingual Information Framework. This project is integrative and workflow oriented in nature, referring to a number of existing standards in language industry (see 3.2 below in the context of the OAXAL architecture). Potentially this standard, once fully specified, can become a very useful document for multilingual information workflows. It would fit to the OAXAL reference architecture that will be dealt with in chapter 3.2 as a generic workflow-oriented information management guideline, also with reference to multimodal information.

Another important area of de facto standards is the evaluation of machine translation processing results. The best known algorithm is BLEU (bilingual evaluation understudy, developed in IBM), by Kishore Papineni Salim Roukos Todd Ward Wei-Jing Zhu. (see also references Papineni 2001, KIS2002, KUL2004, TUR2003). There is current research work to improve such algorithms, further develop them or develop new ones. They are often used as gold standards against which MT systems are evaluated in order to be able to compare such systems to each other. We can expect more results from this research field also in terms of usable standards for advancing MT industry.

The standardization action plan from CLARIN discusses more NLP-related initiatives to create standards (metadata management, long-term preservation, tagsets, web services, etc.). Currently these initiatives are not yet mature enough to be relevant for industrial application. In the long run such initiatives are certainly needed for large-scale resource infrastructures with true interoperability management among networked language resources of any kind.

The MLIF standard inititiative is also an example of explicitly covering multi-modality in information presentation. It refers to standards such as SMIL (Synchronized Multimedia Integration Language) that have become highly relevant in multimedia (multimodality) processing and standards such as and VoiceXML that are used for speech processing.

Larson describes EMMA (Extensible MultiModal Annotation markup language) from 2009 as the main element of a multimodal standard framework developed by the W3C (several working groups dedicated to developing and maintaining these standard recommendations. He lists in his overview of multimodal standard languages the following:

- “Extended Hypertext Markup Language (XHTML)10.an XML version of HTML for presenting visual information on screens
- Speech Synthesis Markup Language (SSML)11.an XML-based language used to render text as speech
- Scalar Vector Graphics 1.2 (SVG)12.an XML-based language for writing two-dimensional vector and mixed vector/raster graphics
- Synchronized Multimedia Integration Language 2.0 (SMIL)13.an XML-based language for writing interactive multimedia presentations” (Larson 2005)

There are also other W3C recommendations to be listed here:

- Pronunciation Lexicon Specification (PLS) Version 1.0, a W3C Recommendation from 14 October 2008
- Semantic Interpretation for Speech Recognition (SISR) Version 1.0

Although speech processing and multimodal processing have existed for quite a long time, it seems that only now they are taking off the ground, due to industrial developments enabling large groups of persons to use and develop multimodal applications, the rapid development of telecommunications industry and of multimedia industry. All these developments have been made possible by intensive research as well as R&D work. The International Telecommunications Union (ITU) has also been active in

this field for many years and has produced important standards on the technical level of speech processing (ITU-T Speech and audio coding standardisation).

In a parallel project on language research infrastructures (CLARIN) and its work package 5 on language resources and technologies, we also deal with LR standards, mainly from a research perspective, i.e. the needs and requirements of linguists in applying LR standards in their research work and linguists actively participating in and contributing to developing new LR standards or improving existing ones. In a CLARIN workgroup we have identified LR standards in several groups, most of them are covered in this document D 4.1 in FLaReNet, but there we also discuss the more foundational standards such as XML itself. In ongoing coordination and concertation work between the two projects, we are currently discussing the research-oriented dimension of LR standards in the CLARIN working groups in its work package 5 and include this dimension here in work package 4 in FLaReNet in the overall landscape (see the working document "CLARIN standardization action plan", 2009).

3.2 Translation & localization standards

LISA (Localization Industry Standards Association) has been founded around 1990 and has been actively promoting the development and application of practical standards in this industry. Since no such standard is ever adopted by a company without a visible return on investment, the whole approach in LISA, in particular in its standards development group OSCAR, is always needs- and problem-driven. It follows the principle of economy to use standards is little as possible and as much as necessary. As it turns out, there are quite a number of standards that have proven to be necessary for industrial purposes.

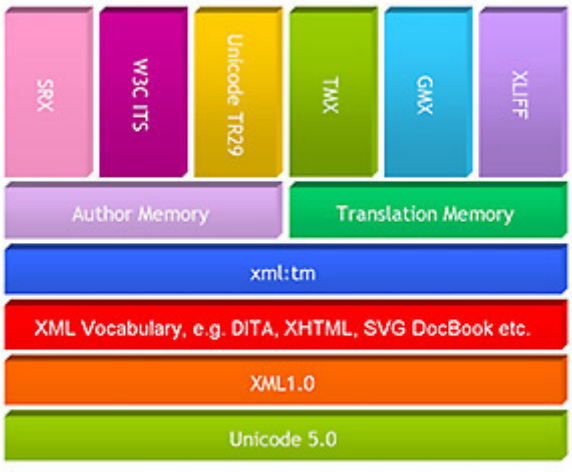
In this report we will not look at all LISA standards efforts but focus on its most recent, and most comprehensive, most strategic initiative, called OAXAL. The following approach is a promising reference architecture for localization management: OAXAL – Open Architecture for XML Authoring and Localization Reference Model**.** It is not only a reference model, but also a newly founded OASIS reference architecture technical committee under the same name.

It is interesting for several reasons:

1. it is not limited to localization in the narrower sense, but open to all authoring and publishing processes in dynamic workflows in multilingual communication environments
2. it is consistently based upon and oriented towards open standards
3. it is component-based and flexible, yet integrative in nature
4. it is practice-oriented and driven by practical needs arising from real-life localization industry workflows

The following paragraphs on OAXAL are taken from Andrzej Zydroń: OAXAL: What Is It and Why Should I Care? (Andrzej Zydroń, CTO, XML-INTL Ltd. and Member, LISA OSCAR), 2008:

"OAXAL is made up of a number of core standards from W3C, OASIS and, last but not least, LISA, the Localization Industry Standards Association. The following is a diagrammatic representation of the OAXAL standards component stack:"



OAXAL standards component stack

The foundational standard in this model is Unicode (equivalent to ISO 10646, driven and maintained by the Unicode consortium), as it is imperative to use character coding standards that really support multilingualism. Related to Unicode are ISO standards on language identifiers and script identifiers (see below). At the highest layer there is also reference to Unicode Technical Report 29 that is one of the most important implementation guidelines e.g. for tokenizers to be created and other programming and implementation work.

The next layer is XML. Now that Microsoft has increasingly adopted XML in their latest product versions, XML will take off the ground also in those IT sectors that have so far been reluctant in adopting XML. This is good for all LR standards, not only for those that are anyway based on XML, but also for those that have other origins provided that they offer open exchange formats and transformation routines from and to XML. The problem with XML is that different "flavours" of it exist in different communities (similar to what has happened in the worlds Oracle and Unix user communities over many years). An important principle in multilingual environments (translation, localization) for XML use is UTF 8 or UTF 16 character coding of documents.

For the localization field the Internationalization Tag Set (ITS) by W3C is another important implementation basis. "It allows for the declaration of Document Rules for localization. In effect, it provides a vocabulary that allows the declaration of the following for a given XML document type such as DITA:

- Which attributes are translatable.
- Which elements are 'in line', that is, they do not break the linguistic flow of text, e.g., 'emphasis' elements.

- Which inline elements are 'sub flows', that is, although they are inline, they do not form part of the linguistic flow of the encompassing text (e.g., 'footer' or 'index' markers).

W3C ITS provides much more, including a namespace vocabulary that allows for finetuning localization for individual instances of elements within a document instance. W3C ITS is therefore at the core of localization processing.“ (Zydron 2008)

“Standard XML Vocabularies
DITA, DocBook, XHTML, SVG – all of these standards dramatically reduce the cost of XML adoption. One of the factors that initially limited the adoption of XML was the high cost of implementation since XML DTD and/or Schema definition is neither simple nor cheap. Not only can costs be reduced dramatically, but as is the case with DITA, these standard tools and utilities often introduce key advances in the way we understand, build and use electronic documentation.“ (Zydron 2008).

In the field of technical authoring DITA has rapidly taken off the ground and is being implemented in many production environments in industry. It can be expected that in the years to come also localization industry will profit from its implementation, most probably as a building block in applications of this OAXAL reference architecture or components and reduced configurations of it.

“xml:tm
xml:tm (xml:text memory) is a key standard from LISA OSCAR. xml:tm introduces a revolutionary approach to document localization. Think of xml:tm as the standard for tracking changes in a document. It allocates a unique identifier to each translatable sentence or standalone piece of text in an XML document. It is a core element of OAXAL, as it links all of the other standards into an elegant, integrated system. At the core of xml:tm are the following concepts, which together make up 'Text Memory':

- Author Memory
- Translation Memory

You can think of Author Memory in terms of change tracking, but also as a way to insure authoring consistency – a key concept in improving authoring quality and reducing translation costs. As far as Translation Memory (TM) is concerned, xml:tm introduces a revolutionary approach to document localization. It is very rare that a standard introduces such a fundamental change to an industry. Rather than separating memory from the document by storing all TM data away from the document in a relational database, xml:tm uses the document as the main repository with no duplication of data. This approach recognizes, fundamentally, that documents have a lifecycle, and that within that life cycle they evolve and change, and that at regular stages in that cycle, they require translation.“ (Zydron 2008).

As Zydron points out, the distinction between author memory and translation memory is important here for a new perspective on document authoring and management. It will certainly take some time for tool providers to adopt this approach more explicitly and pro-actively, but implicitly this distinction has been made already in some production

environments. As the OSCAR group within LISA consists of tool developers and some major customer companies in localization, chances are good that this text memory approach will be widely adopted.

In fact, this approach mirrors a general trend in translation industry and multilingual information environments in general where the clear-cut distinction between translation (based on a source text) and multilingual authoring (not based on a source text but parallel text production in several languages based on a clearly defined globalization strategy) does not exist anymore. It also reflects the insight in translation studies and in translators who critically reflect what they are doing that what we translate is not a source text but content that happens to be expressed in a text that may be considered a source text, but where the "target text" is very much governed by the communication and information goals defined for target audiences by those who have asked for a translation.

The following standards developed by the LISA OSCAR group, SRX, TMX and GMX are also very practice oriented and have all arisen from practical problems that asked for industry solutions. It should also be mentioned, though, that quite some pre-normative research went into the development of these standards, in particular the first one from a chronological point of view, ie. TMX.

"TMX (Translation Memory eXchange) is the original standard from LISA OSCAR. It helped break the monopoly that proprietary systems had over translation memory content. TMX allows customers to change systems and Language Service Providers without loosing their TM assets." (Zydron 2008).

Nowadays all tool providers in the field of computer-assisted translation that use this "text memory" approach have implemented TMX as input/output options. This has certainly been a very successful development and implementation of a language technology standard. The disadvantage we see is rather the fact that the technology of memory-based translation systems has hardly changed for a long time and that usability of all such systems is low, thus user dissatisfaction (in translators who have to work with every day) is high. But this usability issue is not to be attributed to the standard, rather to the lack of motivation to invest in new technologies in translation tool provider industry.

TMX alone soon proved to be under-specified and not enough for consistent workflow support. Thus SRX and GMX soon followed:

"SRX (Segmentation Rules eXchange) is the LISA OSCAR standard for defining and exchanging segmentation rules. SRX uses an XML vocabulary to define the segmentation rules for a given language and to specify all of the exceptions. SRX uses Unicode regular expressions to achieve this. The key benefit of SRX is not so much exchange, as the ability to create industry-wide repositories for the segmentation rules for each language." (Zydron 2008).

Some companies have started to use SRX in a networked way. The years to come will show whether it is successfully adopted by the industry communities. Segmentation is a

general methodological problem for the text memory-based translation approach. While the focus on the sentence or parts of sentences has been mostly unquestioned, increasingly translators are discomforted as they would need flexible supra-sentential (text-oriented) segmenting. After all, what we translate is not sentences, but texts.

"GMX (Global information Management Metrics Exchange) is a three-part standard from LISA OSCAR that focuses on translation metrics. GMX/V defines what constitutes word and character counts, and allows for the exchange of metrics information within an XML vocabulary. […] before GMX/V, there was no standard for word or character counts. GMX/V defines a canonical form for counting words and characters in a transparent and unambiguous way. The two associated standards, yet to be defined, will be GMX/C for complexity and GMX/Q for quality. Once the three GMX standards are available, they will provide a comprehensive way of defining a given localization task." (Zydron 2008)

Translation metrics has become an important topic for industry-driven standardization efforts.

"XLIFF (XML Localization Interchange File Format) is an OASIS standard for the exchange of data for translation. Rather than having to send full unprotected electronic documents for localization, with the inevitable problems of data and file corruption, XLIFF provides a loss-less way of round tripping text to be translated. Language Service Providers, rather than having to acquire/write filters for different file formats or XML vocabularies, have merely to be able to process XLIFF files, which can include translation memory matching, terminology, etc." (Zydron 2008).

XLIFF is workflow oriented and integrative in nature and is increasingly adopted in industry.

The OAXAL reference model is open and will further be extended. TBX, for instance, will be added to the model (see chapter 3.3 for details on TBX). OAXAL is certainly useful already from a conceptual and strategic point of view, as it invites decision makers in industry not to take a look at each individual standard in an isolated way but rather to look at the whole model from a workflow and integration perspective. Then one can decide which building blocks or components are actually relevant for a particular implementation and application scenario.

There are other translation oriented standards, but not so much focused on LR but rather on translation service provider quality management aspect, e.g. EN 15038 on translation services. But ongoing standards development work in ISO TC 37 SC 2/ WG 6 now also focuses on translation metrics similarly to the domain-specific standard from the automotive industry, SAE J2450:2001, Translation Quality Metric. It provides a method for assigning weighted rating values to errors in translation processes.

3.3 Terminology standards (incl. lexicography)

A terminology-related standard also adopted and further developed by LISA OSCAR is TBX, which stands for TermBase eXchange format. Unlike TMX, TBX has not yet been

fully adopted by tool developers, but more recently there is more openness and willingness by tool developer companies to support it in their terminology management environments not only as an exchange format, but also a basis for data modelling for terminological databases.

Originally developed in the EU-funded project SALT (2000-2002), TBX was then handed over to LISA OSCAR. In parallel, ISO/TC 37/SC 3 started a working group to develop TMF, the terminology markup framework that was soon published as ISO 16642. TMF is not limited to XML, but also has UML notation and modelling options. TMF is the overarching foundational standard for all forms of terminology markup. TBX is one TMF-conformant markup language for terminological data.

The problem with TMF is that is highly abstract and has no direct industrial implementation need. It is a long term strategic vision from a meta-modelling perspective. This makes it quite abstract for most user communities except for those who do research in this field or for those who professionally deal with meta-modelling and meta-data modelling in large IT environments. But TBX, as one of its concrete manifestations, is much more concrete and applicable. In an improved and cleaned-up version, it recently also became an ISO standard in ISO/TC 37/SC 3 under the code IS 30042. This successful cooperation between LISA and ISO is similar to the one between Unicode and ISO in the area of character coding. In the last 2 years TBX has been increasingly adopted and recognized as a useful standard.

ISO 1951 is a standard for (more traditional) dictionary publishing houses. It has been developed by a group of experts working at various publishing houses and is currently been used by at least one global player in the field of dictionary publishing, mainly located in Germany. It focuses on traditional mono- and multilingual lexicography, also in terms of workflows and data modelling. A wider adoption of this standard is not imminent, as comparable publishing houses have their own, traditional workflows; dictionary compilers are hesitant and reluctant to adopt such standards, only when they are embedded in standard corporate tools and workflows.

In the wide field of terminology there are many more standards. Terminological resources are a type of language resource that is usually not so much discussed in the academic corpus linguistic community, since terminology managers in companies or international organizations are usually not corpus linguists by training, but rather translators, translator-terminologists, information scientists, computer scientists or quite often domain experts who happen to be responsible for terminology in their company. Thus terminology standards traditionally are written in different language from the standards discussed in chapter 3.1, for instance.

In terminology there is an important distinction between methodological standards and data standards. ISO/TC 37 is the committee for “terminology and other language- and content-oriented resources”. It has been founded in 1936 and re-established after WWII in 1952. Only some years ago the traditional field of terminology standardization opened up to language resource management and their standards, which led to the establishment of a fourth sub-committee, TC 37/SC 4, where most of the standards discussed in 3.1 have been worked on (see above in 3.1).

Methodological standards in terminology include first of all ISO 704 (has existed since 1948 with continuous publications of new versions adapted to changing needs, the latest version is just being prepared for publication (official publication date expected in early 2010). ISO 704 is widely used by thousands of standards committees not only within the hundreds of ISO committees in different branches of industry, but also in other international organizations where standardization of their domain- or corporate terminologies needs to be done. Yet the standard is abstract in language, and we have the ambition to improve its language to make it more accessible to different user groups. ISO 704 deals with principles of writing (good) definitions, creating terms in any language, specifying language-independent (semiotic) criteria for technical terms in any domain, for concept systems, definitions, and other basic building blocks of terminological resources. Another widely used methodological standard is ISO 10241 that governs the preparation and layout of terminology standards, i.e. data standards. Thousand of such data standards exist, not only in ISO, but also in all other international, regional and national standards bodies and other regulatory authorities. While terminology standards have traditionally been published in paper-form as mono-, bi- or multilingual glossaries with terms systematically defined, increasingly terminology standards are treated as language resources in digital workflows and increasingly "exist" not on paper (only) but rather in databases that are made available to all user groups in public that are interested in them or who are required to use them for internal purposes of corporate language management, etc. It is hard to estimate the size of this type of language resources. It is a safe bet to estimate that standardized terminologies include several dozens of millions of terms in different languages in different domains. They are fragmented along corporate, domain, and language boundaries, which makes it difficult or rather impossible to get hold of ALL of them, if anybody would like to. Rather, those who create these domain-specific terminology resources are domain experts themselves and thus know exactly who their target group or constituency is and appropriately address them. The problem starts when such domain-specific terminologies leave their domain or their corporate/organizational context and are supposed to be used in different contexts. This is where problems start to arise, as definitions are often mirroring corporate language and corporate cultural traditions that are unknown or foreign in other application scenarios. People who want to collect such data and "sell" them or provide them to a public at large tend to ignore this fact. Then users run into problems as soon as they try to use such corporate terminologies in their own context.

An important new strategy by ISO Central Secretariat is to offer a central terminology database (ISO Concept Database) free of charge. It contains all terms and definitions contained in all ISO standards. This follows their "standards as database" strategy. This has important consequences as the division between the meta level of terminological principles and the object level of terminological data manifests itself in a completely different way in digital environments than on paper.

For some standards, this division does not exist any more, for example ISO 12620, governing data category definition and management for language industry in general. It is originally a terminological meta-standard, but now it has become an online work environment for registering, following a carefully specified methodology, data

categories with their names and definitions, value ranges and constraints. It is open to the public, yet well organized in its governance structure of expert groups taking informed decisions in data category maintenance. The underlying system has been called DCR (Data Category Registry), while the operational interface is called ISOcat. It has been adopted as an important building block in strategic LR infrastructure projects such as CLARIN, but is also used by large corporations and international organizations for their corporate data category subsets they are maintaining in a standardized way. This is a crucial step towards consistency and interoperability beyond corporate boundaries of global information exchange, eCommerce, and any other environment, irrespective of competitive or collaborative relationships among different stake holders, customers, users, data creators, etc.

ISO 639 (dealing with language identifiers) is another example of the new standards-as-database approach. So far it exists in 6 different parts, dealing with different subsets of language identifiers (with these subsets being of very different sizes, ranging from 200 to dozens of thousands). Now it is being integrated into the ISO Concept Database. The governance structure is also being adapted accordingly.

There is a long list of methodological standards (12620 and 639 are both, methodological and a data standard) in the field of terminology, but for the purposes of this first analysis in D 4.1 it is enough to mention these few (704, 10241, 16642, 30042, 12620, 639), as they are quite different in nature and basically represent the different kinds of standards currently prepared in ISO TC 37 and used all over the world.

3.4 Domain-specific LR standards

In this section we discuss 2 examples from a rapidly growing field of domain-specific standards with a clear LR dimension.

a) XBRL is the eXtensible Business Reporting Language (http://www.xbrl.org)

   "XBRL is a language for the electronic communication of business and financial data which is revolutionising business reporting around the world. It provides major benefits in the preparation, analysis and communication of business information. It offers cost savings, greater efficiency and improved accuracy and reliability to all those involved in supplying or using financial data.

   XBRL stands for e**X**tensible **B**usiness **R**eporting **L**anguage. It is one of a family of "XML" languages which is becoming a standard means of communicating information between businesses and on the internet.

   XBRL is being developed by an international non-profit consortium of approximately 450 major companies, organisations and government agencies. It is an open standard, free of licence fees. It is already being put to practical use in a number of countries and implementations of XBRL are growing rapidly around the world." (from www.xbrl.org)

XBRL contains taxonomies in the form of ontologies. There is a whole section on how to create XBRL-conformant taxonomies and how to register them (http://www.xbrl.org/Taxonomies/). Guidance is also provided to user groups.

b) SBVR stands for "Semantics of Business Vocabulary and Business Rules" and is an initiative of OMG (http://www.omg.org/spec/SBVR/1.0/)

The scope of SBVR reads as follows: "This specification defines the vocabulary and rules for documenting the semantics of business vocabularies, business facts, and business rules; as well as an XMI schema for the interchange of business vocabularies and business rules among organizations and between software tools. This specification is interpretable in predicate logic with a small extension in modal logic. This specification supports linguistic analysis of text for business vocabularies and rules, with the linguistic analysis itself being outside the scope of this specification. This specification is applicable to the domain of business vocabularies and business rules of all kinds of business activities of all kinds of organizations. It is conceptualized optimally for business people rather than automated rules processing, and is designed to be used for business purposes, independent of information systems designs. This specification is applicable as input to transformations by IT staff into information system designs, using a combination of decisions from system architects and Platform Independent Model designers together with software tool function." (SBVR specification 1.0, 2008, p. 15)

Both examples reflect real needs of large user communities in specific domains – business and finance communities in these cases – for LR standards with a heavy focus on terminologies, language regulation (controlled language) and public registration of taxonomies. This type of domain-standard with LR aspects is needed today in all domains, large disciplines such as medicine, chemistry, biology, etc. have elaborated huge LR in terms of taxonomies, ontologies, thesauri, terminologies, all with very specific rules for creation and maintenance.

3.5 Foundational IT and Web standards with LR aspects

The following example is given under this rubric:

ISO/IEC 11179 (formally known as the ISO/IEC 11179 Metadata Registry (MDR) standard)

An excerpt of the Wikipedia entry specifically designed by the responsible ISO committee on this standard reads:

"ISO/IEC 11179 is the international standard for representing metadata for an organization in a Metadata Registry.

Strategic nature

Today organizations are required to exchange data quickly and precisely between computer systems using enterprise application integration technologies. Completed transactions must also be regularly transferred to separate data warehouse and business rules systems with specialized structures designed to make data retrieval efficient. Many industry experts feel that this can only be done efficiently if data is precisely defined and automated tools are created to exchange data between remote computer systems. Precise exchange of data between computers is also a key driver behind the W3C's Semantic Web project.

ISO/IEC 11179 is one of the few mature standards for storing enterprise metadata in a controlled environment.

Structure of an ISO/IEC 11179 metadata registry

An ISO metadata registry consists of a hierarchy of "concepts" with associated properties for each concept. Concepts are similar to classes in object-oriented programming but without the behavioral elements. Properties are similar to Class attributes. ISO standards require that each concept and property have a precisely worded data element definition.

Structure of the ISO/IEC 11179 standard

The standard consists of six parts:
Part 1 - Framework
Part 2 - Classification
Part 3 - Registry metamodel and basic attributes
Part 4 - Formulation of data definitions
Part 5 - Naming and identification principles
Part 6 – Registration" (end of quote from the Wikipedia entry)

This generic multi-part standard is highly relevant to LR standardization as a foundational standard. And due to a successful cooperation between ISO TC 37 and ISO/IEC JTC 1 SC 32/WG 2 (see http://metadata-standards.org/ for more details), there is mutual interoperability at this level of meta-standards: The 11179 framework heavily and consistently relies on terminological foundations laid down in ISO 704, while ISO 12620 (DCR/ISOcat), ISO 16642, ISO 639, and other TC 37 standards increasingly rely on the conceptual framework of 11179.

There are many other standards to be mentioned under this category, such as OWL and SKOS from W3C.

4 Needs for future LR standards

As we could see in chapter 3.1 in particular, LR-specific standards are either too much oriented towards academic research purposes or are not yet mature enough for industrial applications. It is not so much the topics that are the problem, rather the way

they have been addressed so far. It is not trivial at all to "translate" pre-normative research results into an industrial standard that is widely adopted.

The problems related to translation and localization standards discussed in 3.2 is the other way round. They are industry-born and driven only by real-life needs and problems to be solved. So they lack long term visions. The hope is that the new OAXAL model architecture provides a strategic, multi-level basis for developing standards that also take into account long term trends in this industry section. As translation and localization are, in addition to their technical basis in tools, MT, etc. also and maybe primarily a service industry, providing professional language services to different customer groups. What is definitely needed for LR standards development in this area is *cooperation between two communities* that have had so far little to do with each other, the MT industry (MT technology providers and MT system users) on the one hand and the CAT-oriented translation and localization industry at large on the other hand. Only this cooperation can provide a broad basis, mixing advance research traditions with well established market-oriented translation & localization service provider industry groups, for well-balanced and more integrative LR standards that are industrially usable yet based on pre-normative research.

Terminology standards as discussed in 3.3 are in the process of re-orientation and re-modeling, developing towards standards as databases, online registries, digital data standards that are directly integratable in corporate workflows. This assessment is also true for standards in 3.4, domain-specific LR standards, while standards in 3.5 (OWL, SKOS) are rather similar to 3.1 standards, as the W3C work environment is essentially research-oriented in nature, which results in hesitation in industry to adopt ontologies and ontology standards in general.

At the European Language Resources and Technologies Forum in Vienna, February 12-13, 2009, session 4 was dedicated to "Openness, Sharing, and Standards". Summarising and analysing the Summary Report by Nancy Ide, the session Rapporteur, we can say that there was concensus on the crucial importance of interoperability in LR at all levels and for any future successful development, both in developing new language technologies and language resources and in their actual application in industrial practice. "In the long term, interoperability will be the cornerstone of a global network of language processing capabilities." (Ide, 2009, p. 1). Interoperability of resources, tools and frameworks has also become a major topic for research – and for creating standards. Thus, pre-normative research in this area has to be supported more pro-actively at all levels – international, EUROPEAN, cross-sectoral, for all languages, including all user communities. A number of initiatives, conferences, projects, etc. have recently been dedicated to this topic of interoperability (e.g. the international conference ICGL, attempts at merging existing annotation schemes (GATE, UIMA, E-Meld TILR, ISO TC 37/SC 4 activities, projects such as KYOTO, CLARIN, INTEROP-SILT, etc.). It was unanimously agreed that more coordination and cooperation is needed in order to harmonise the huge variety and diversity of annotation schemes, encoding schemes, wordnets, ontologies, etc. in all areas discussed above. A lot of scepticism was voiced in the community about existing standards and their lack of acceptance or lack of use. One reason is the lack of consensus in the research community in computational linguistics about representation formats for linguistic data, categorization, annotation, etc.

The following paragraphs are a direction citation from the session report, as they express what is important for this report:

> "The workshop participants suggested a variety of (not necessarily compatible or mutually exclusive) solutions for each of these problems, as summarized below.
>
> 1. Existing standards are not widely accepted/used. A compelling case was made for adopting a model for tool and resource development based on open advancement and collaborative development, where the community as a whole contributes components, modules, etc. to a common system or framework. Interoperability (at some level) is achieved as a necessary by-product. Other suggestions addressed possible changes to the standardization process itself. For example, it was generally agreed that the focus of standardization efforts should be on transformation between representation formats and linguistic annotation categories and schemes, rather than an attempt to establish a single standard for any of these phenomena. This allows researchers and developers to utilize formats and schemes that serve their needs and still have interoperability via transduction to formats suitable for other systems. It was noted that transduction can be fostered by identifying an underlying data model that can be realized superficially in a variety of formats/schemes. The point was also made that standards are often not used because of a lack of tools that support them; providing such tools and ensuring that they are easy to use is essential for widespread adoption.
>
> 2. Disagreement concerning theories/linguistic categories. The ISO Data Category Registry was pointed to as a major effort to address this problem, by providing a centralized repository for the identification and description of linguistic categories that are used in annotation and analysis. However, the ISO DCR does not at this time seek to establish a standard set of such categories, but only to provide a set of definitions / distinctions that can serve as a reference or a point of departure for defining new or variant categories. It was suggested that some steps toward standardization could be taken immediately by taking a "bottom up" approach and addressing only those areas where there is consensus, focusing on the "lowest common denominator" among phenomena. Other suggestions were to take the approach suggested for standards in general above, by establishing mappings/transductions among different categories (the difficulty of applying this to categories was acknowledged due to the lack of one-to-one correspondences in many cases). Finally, it was suggested that ontologies of linguistic information will be needed to provide the framework for establishing standard categories for linguistic annotation.
>
> 3. No standard representation format(s)/ frameworks. There was general consensus that in this area, there is a convergence of practice among several widely used formats, frameworks, and systems, relying on a UIMA-like architecture of configurable pipelines of language processing modules, and representing results using some surface format that serializes an underlying, adequately expressive abstract model for linguistic information. The emergence

of generic "pivot" formats (e.g., LAF, LMF) that realize the abstract model, into and out of which various serializations can be mapped (for interchange) is also contributing to convergence.

It was also suggested that rather than focusing on how things are represented at relatively low levels of analysis, we should focus on input/output formats for tools instantiated as web services.

4. Lack of Accessibility. Very few concrete solutions for the problem that resources and tools can be hard to find were suggested. The publicizing of LMF via Wikipedia was cited as a possible solution, and community outreach and education were recommended. Access rights pose another kind of obstacle, and it was clear from the conversation that there is a growing sentiment in support of open source development, and free and unfettered access to resources and tools. A federation of centers was suggested, which would negotiate access rights to data and software with commercial and other enterprises that hold them." (excerpt from Nancy Ide's session report)

The lessons for the FLaReNet project that were drawn focus on standards strategies (top-down vs. bottom-up, modular component standards rather than a single monolithic standard for all of LR, and focus in the short term planning on those areas where there is enough consensus so that chances are high that a widely accepted standard can be published in a short period of time. Long-term planning involves the creation of a generic LR standards framework and roadmap (D 4.2 and D 4.3 of this project).

5 Conclusions and further work

Due to the increasingly rapid change in science, technology, commerce, and other spheres of society, standards need to co-evolve at a higher speed. This is also true for language resource standards that have been developed so far. Standards maintenance has to be a process of change management, ideally in real time so that standards users can implement any change in a standard immediately.

There are several reasons why LR standards (across the wide and heterogeneous spectrum ranging from 3.1 to 3.5) are not used, as they might be:

- too research-oriented, not usable in industry, difficult to understand
- too abstract, lack concrete examples for implementation, lack of user scenarios or user guides
- too isolated, existing only on paper but not integratable in digital workflows,
- too cumbersome to implement, no return on investment in sight for implementers

Thus, for each LR standard, the return on investment and the possible motivations of users should be elaborated, ideally together with potential or real users (early adopters).

Emerging needs (content representation, ontologies, interactive language automation, ambient technologies, etc.) should be analysed in more detail (D 4.2).

The trend towards new forms and manifestations, in particular as embedded standards, such as standards as data or as tag sets, standards as formats, standards as databases, standards as software, standards as processes and workflows, standards as (web) services, interoperability protocols, standard data models, meta-models, resource registry standards, etc., will become more and more relevant to the whole area of LR standards.

The area of societal sectors using IT mentioned in the beginning (eHealth, eLearning, eGovernment, etc.) is one of the areas where new needs for LR standards are emerging most dynamically. This means that standards of the type as described in chapter 3.4 will become more and more frequent and deserve more attention that so far. Especially, as they tend to combine with other standards, in particular with terminology standards (3.3), but increasingly with the others (3.1, 3.2, 3.5). This calls for an integrative view on LR standards. Despite the fact that LR standards are so different in nature (even within groups they are very heterogeneous and different, in particular in 3.1), an overall strategy for all types of LR standards becomes necessary. Their common denominator is that they deal with language (or specific parts of it) in digital work environments.

A simple example is CALL (computer-assisted language learning) – now associated with eLearning: It covers all types of LR standards (3.1-3.5), as language learners use different kinds of corpora for different didactic purposes, often in academic translation programs or for further training of language professionals working in localization business, using domain-specific terminology standards in different languages, increasingly represented also in the form of ontologies.

More attention is needed for multimodal information processing and speech processing standards. W3C working groups and other communities have developed a whole framework of standards ("recommendations") that are rapidly taken up in industry just as well as in research communities of different scientific disciplines not limited to computer science and linguistics but covering humanities and social science disciplines such as media studies, cultural studies, sociology, etc.

In summarizing we can list the following actions and requirements for the area of interoperability and standardization in the field of language resources (in the wider sense):

- LR standards have to be made more operational (both, existing ones and those under preparation), with a specific view on different user communities – most users should not or do not want to know that they are using standards, they should operate in the background and they should be "inherent" to the language technology tools or more generic tools they use
- An interoperability framework for LR has to be developed (cross-domain, cross-purpose, cross-cultural, etc.) as part of D 4.2 and D 4.3 in this project
- Analyse the needs and requirements for harmonisation of existing standards

- Create an operational ecology of language resource standards that are easily accessible, re-usable, effective, and that contribute to semantic interoperability
- Contribute to and expand the EIF (European Interoperability Framework), e.g. in the context of eGovernment, eHealth, etc. where many of the existing LR standards can already contribute effectively to enhance data interoperability.
- Increase the acceptance of LR standards (and the need for them) in different communities, both research and industry communities, and to directly involve user communities in creating standards
- Develop a strategy for LR standards creation, taking into account both, bottom-up and.top-down approaches with an interactive process model needed
- Inform more pro-actively on best practices in implementing standards and in successful corporate language standards
- Develop a broader vision of LR standards with the inherent inclusion of multimedia, multimodal and speech processing applications
- Bring together research communities and industrial application communities for developing a joint vision on LR standards in general

A possible path is to develop a "language resource interoperability framework" (Budin 2009), with a clear focus on intergrating diverse standards beyond the scope of LR in the narrow sense but including many other standards that are relevant for complex workflows in large organizations, in highly distributed communities in all branches of industry, trade, science, and technology. The following chart illustrates this approach:

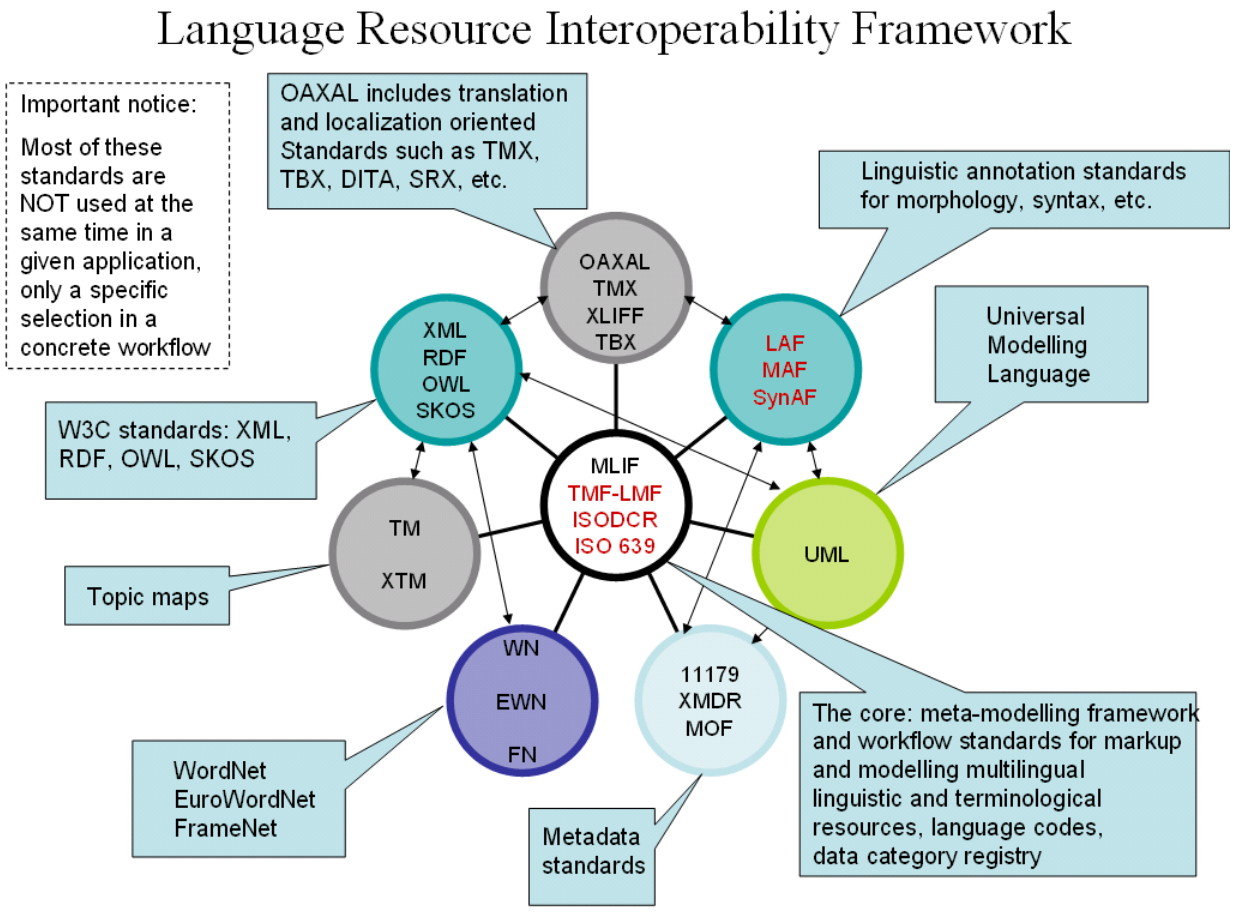


Figure 2: proposal for a language resource interoperability framework (Budin 2009)

This figure includes standards from many different communities and for different purposes. As the note in the chart points out, no single concrete application will involve ALL such standards, but there are many examples for concrete applications where SOME of these standards are indeed required for successful business or successful industrial product development as well as for research initiatives.

Acronyms used for standards, and the organizations responsible for them (as used in this figure)

Acronym	Full name	Responsible organization
XML:	Extensible Markup Language	W3C
RDF:	Resource Description Framework	W3C
SKOS:	Simple Knowledge Organization System	W3C
OAXAL:	Open Architecture for XML Authoring and Localization	LISA/OASIS
OWL:	Web Ontology Language	W3C
TM:	Topic Maps	Ontopia/ISO
XTM:	XML Topic Maps	Ontopia/ISO
WN:	WordNet	Uni Princeton
EWN:	EuroWordNet	EWN Cons.
FN:	FrameNet	ICSI Berkeley
TBX:	Termbase Exchange Format	LISA/ISO
TMX:	Translation Memory Exchange Format	LISA
MLIF:	Multilingual Information Framework	ISO
XLIFF:	XML Localization Interchange File Format	OASIS
TMF:	Terminology Markup Framework	ISO
LMF:	Lexical Markup Framework	ISO
LAF:	Linguistic Annotation Framework	ISO
EIF:	European Interoperability Framework	CEN/EU
ISO 11179:	Metadata Registry (MDR) standards framework	ISO
XMDR:	Extended Metadata Registry	ISO
MOF:	Metamodel Object Facility	ISO
UML:	Unified Modeling Language	OMG
DITA:	Darwin Information Typing Architecture	OASIS
ISO 639:	Language Codes	ISO
ISO/DCR:	Data Category Registry	ISO
ITS:	Internationalization Tagset	W3C

6 References

CLARIN Standardization action plan (working document, 2009)

Ide, Nancy, Session Report on Session 4 at the European Language Resources and Technologies Forum in Vienna, February 12-13, 2009, dedicated to “Openness, Sharing, and Standards” (2009)

Zydron, A. (2009). OAXAL: What Is It and Why Should I Care? Globalization Insider http://www.lisa.org/globalizationinsider/2008/09/oaxal_what_is_i.html?printerFriendly=yes

Budin, G. (2009). Global LRT infrastructures for enhancing globalization workflows (forthcoming), paper presented at LISA@Berkeley conference in August 2009, Berkeley, USA.

Some references to BLEU and other MT evaluation metrics:

KIS2002 - Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318. http://www1.cs.columbia.edu/nlp/sgd/bleu.pdf

KUL2004 - Alex Kulesza and Stuart M. Shieber, A learning approach to improving sentence level MT evaluation, in proceedings of the Tengh Conference on Theoretical and Methodological issues in Machine Translation, Baltimore, 2004 http://www.eecs.harvard.edu/~shieber/Biblio/Papers/kulesza-mt-evaluation04.pdf

TUR2003 - Joseph P. Turian, Luke Shen, and I. Dan Melamed, Evaluation of Machine Translation and its Evaluation a revised version of the paper to be presented at MT Summit IX, New Orleans, LA, 2003 http://nlp.cs.nyu.edu/publication/papers/turian-summit03eval.pdf

Kishore Papineni Salim Roukos Todd Ward Wei-Jing Zhu: Bleu: a Method for Automatic Evaluation of Machine Translation, IBM research paper, 2001

Multimodal language standards including speech (W3C):

Larson, James. Standard Languages for Developing Multimodal Applications. 2005

VoiceXML: Voice Extensible Markup Language 3.0, December 2008, W3C

Extended Multimodal Application (EMMA) language, 2009, W3C

Extended Hypertext Markup Language (XHTML)

Speech Synthesis Markup Language (SSML)

Scalar Vector Graphics (SVG)

Synchronized Multimedia Integration Language (SMIL)

Pronunciation Lexicon Specification (PLS) Version 1.0, a W3C Recommendation 14 October 2008

Semantic Interpretation for Speech Recognition (SISR) Version 1.0

Domain-specific language standards:

SBVR
Semantics of Business Vocabulary and Business Rules (SBVR), v1.0 OMG Available Specification available at: http://www.omg.org/spec/SBVR/1.0/PDF/

XBRL eXtensible Business Reporting Language at: www.xbrl.org

Foundational standards:

Information on the 11179 framework in Wikipedia under:
http://en.wikipedia.org/wiki/ISO/IEC_11179

List of ISO standards from ISO/TC 37 (30.9.2009)

ISO 704: 2000 Terminology work – Principles and methods (new version published in 2010)

ISO 860:1996 Terminology work – Harmonization of concepts and terms (new version published soon)

ISO 1087-1:2000 Terminology work – Vocabulary – Part 1: Theory and application

ISO 22134 Practical guide for socioterminology

ISO 639-1:2002 Codes for the representation of names of languages – Part 1: Alpha-2 code

ISO 639-2:1998 Codes for the representation of names of languages – Part 2: Alpha-3 code

ISO 639-3: 2008 Codes for the representation of names of languages Part 3: Alpha-3 code for comprehensive coverage of languages

ISO FDIS 639-4 Codes for the representation of names of languages Part 4: Implementation guidelines and general principles for language coding

ISO FDIS 639-5 Codes for the representation of names of languages Part 5: Alpha-3 code for language families and groups

ISO FDIS 639-6 Codes for the representation of names of languages Part 6: Extension coding for language variation

ISO 1951:2007 Presentation/Representation of entries in dictionaries – requirements, recommendations and information

ISO 10241:1992 International terminology standards – Preparation and layout (new version published in 2010 in 2 parts)

ISO 12199:2000 Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet

ISO 12616:2002 Translation-oriented terminography

ISO 15188:2001 Project management guidelines for terminology standardization

ISO 12615: 2005 Bibliographic references and source identifiers for terminology

ISO 22128: 2009 Quality assurance guidelines for terminology products

ISO 23185: 2009 Assessment and benchmarking of terminological holdings

ISO 1087-2:2000 Terminology work – Vocabulary – Part 2: Computer applications

ISO 12200:1999 Computer applications in terminology – Machine-readable terminology interchange format (MARTIF) – Negotiated interchange (to be revised)

ISO 12620:1999 Computer applications in terminology – Data categories (new version to be published in 2010, establishing ISO-DCR and ISOcat

ISO 16642:2003 Computer applications in terminology – Terminological markup framework

ISO TR 2009 12618 Computational aids in terminology – Design, implementation and use of terminology management systems

ISO CD 21829 Terminology for language resources

ISO DIS 23679-1 Word segmentation of written texts for mono-lingual and multi-lingual information processing – Part 1: General principles and methods

ISO CD 23679-2 Word segmentation of written texts for mono-lingual and multi-lingual information processing – Part 2: Word segmentation for Chinese, Japanese and Korean

ISO/CD 24610-3 Language resource management – Feature structures – Part 3: Word segmentation for other languages

ISO CD 24611 Language resource management – Morpho-syntactic annotation framework

ISO CD 24612 Language Resource Management – Linguistic Annotation Framework

ISO 24613: 2008 Language resource management – Lexical markup framework


FLaReNet