LREC 2008 Workshop

Uses and usage of language resource-related standards

PROCEEDINGS

Edited by

Andreas Witt, Felix Sasaki, Elke Teich,
Nicoletta Calzolari, Peter Wittenburg

May 27, 2008

LREC 2008 Workshop

Proceedings of the LREC 2008 Workshop
“Uses and usage of language resource-related standards”

Edited by Andreas Witt (Tübingen, Germany), Felix Sasaki (W3C), Elke Teich (Technische Universität Darmstadt), Nicoletta Calzolari (ILC Pisa), Peter Wittenburg (MPI Nijmegen)

May 27, 2008

Organisers

- Nicoletta Calzolari, ILC Pisa
- Felix Sasaki, W3C
- Elke Teich, Technische Universität Darmstadt
- Andreas Witt, Tübingen University
- Peter Wittenburg, MPI Nijmegen

Programme Committee

- Nuria Bel (Universitat Pompeu Fabra Barcelona, Spain)
- Lou Burnard (University of Oxford, UK)
- Gerhard Budin (Vienna University, Austria)
- Key-Sun Choi (KAIST, Republic of Korea)
- Thierry Declerck (DFKI, Germany)
- Peter Fankhauser (FhG-IPSI, Darmstadt, Germany)
- Gil Francopoulo (Tagmatica, France)
- Thilo Goetz (IBM, Germany)
- Yoshihiko Hayashi (Osaka University and NICT, Japan)
- Uli Heid (Stuttgart University, Germany)
- Erhard Hinrichs (Tübingen University, Germany)
- Nancy Ide (Vassar College, New York, USA)
- Fotis Jannidis (Technische Universität Darmstadt, Germany)
- Kiyong Lee (Korea University, Republic of Korea)
- Christian Lieske (SAP AG, Germany)
- Monica Monachini (ILC Pisa, Italy)
- Laurent Romary (MPG, Germany)
- Justus Roux (University of Stellenbosch, South Africa)
- Yves Savourel (Enlaso, USA)
- C.M. Sperberg-McQueen (W3C, USA)
- Sue Ellen Wright (Kent State University, OH, USA)

Preface

In recent years more and more standards related to the representation and processing of language resources (LRs) have been issued by different organizations (e.g., ISO, Oasis, LISA, W3C, TEI etc.). While some of them have been designed explicitly for the purpose of modeling language (e.g., LAF, OLAC, IMDI, etc), other standards in use in the community are originally geared to the representation of properties of texts (e.g., TEI). Still others have been defined with broader, not inherently language or text-related purposes in mind. Recently, standards for resource management are also emerging.

Obviously, the success of a standard is dependent on how well it is received in a community. The present workshop provides a platform for an exchange of experiences between the developers of standards and the users of standards. There are eight papers that report and reflect on the usage of standards in language resource building and deployment: Lyding, Bel & Bel and Vasiljevs & Rirdance present experiences with standards in the contexts of multilingual data, lexical resources and terminology banks, respectively; DeCamp and Francopoulo et al. report on standards for language names and linguistic categories (both in ISO) and Wittenburg & Broeder discuss metadata representation in DMD; Wright & Summers and Quin et al. take the perspective of general standards for data/document representation and their relation to LR-related standardization.

The workshop concludes with a panel on current challenges in language resource-related standardization, featuring some key figures from the standardization community including Sue Ellen Wright (ISO), Henry S. Thompson (W3C) and Laurent Romary (TEI).

Andreas Witt, Felix Sasaki, Elke Teich, Nicoletta Calzolari & Peter Wittenburg May 2008

Programme

09.00 – 09.45	Liam Quin, Felix Sasaki, C.M. Sperberg-McQueen, Henry S. Thompson Architectural Specifications for the WorldWideWeb and their Role for Language Resources
09.45 – 10.30	Jennifer DeCamp Standards for the Representation of Language Names – A Case Study
11.00 – 11.45	Núria Bel & Santiago Bel Measuring standards in Lexical Resources
11.45 – 12.30	Sue Ellen Wright & Dave Summers Crosswalking from Terminology to Terminology: Leveraging Semantic Information across Communities of Practice
12.30 – 13.15	Gil Francopoulo, Thierry Declerck, Virach Sornlertlamvanich, Eric de la Clergerie, and Monica Monachini Data Category Registry: morpho-syntactic and syntactic profiles
14.30 – 15.15	Peter Wittenburg & Daan Broeder On the Relevance, Standards and Usage of Metadata for electronic language resources
15.15 – 16.00	Verena Lyding Applying the Corpus Encoding Standard to a quadrilingual corpus of legal texts
16.30 – 17.15	Andrejs Vasiljevs & Signe Rirdance Application of terminology standards for a multilingual term bank: the EuroTermBank experience
17.15	Panel discussion with Sue Ellen Wright and Nicoletta Calzolari (both ISO), Henry S. Thompson (W3C), Laurent Romary (TEI) and Peter Wittenburg (MPI), Chairs: Elke Teich and Andreas Witt

Table of Contents

Architectural Specifications for the World Wide Web and their Role for Language Resources

L. Quin, F. Sasaki, C.M. Sperberg-McQueen, H.S. Thompson

W3C/MIT, W3C/Keio, W3C/MIT, W3C/ERCIM
liam@w3.org, fsasaki@w3.org, cmsmcq@acm.org, ht@inf.ed.ac.uk

Abstract

This paper describes specifications which have been (or are being) developed within the Architecture Domain of the World Wide Web Consortium. This Domain is responsible for many of the core technologies for the World Wide Web, including XML. We will describe XML-related technologies in five areas: validation, full-text analysis, declarative descriptions of XML processing, layout, and Internationalization, focusing on how they are particularly suited for the representation and processing of language resources. The paper also includes a broad overview of the standardization process which underlies the development of these and other W3C technologies.

1. Introduction: A Brief Overview of W3C and its Process

W3C[1] is an international consortium with the mission to develop Web standards, with contributions from W3C member organizations, the W3C staff, and the public.

W3C is working on a technology stack informally described at <http://www.w3.org/Consortium/techstack-desc.html>. The Work is organized in Activities like the XML Activity or the Internationalization Activity, which are parts of domains (Architecture, Interaction, Technology and Society, Ubiquitous Web, Web Accessibility). Work items are described in charters for Working Groups, Interest Groups, or Incubators. The difference between these is their scope. Working Groups and Interest Groups concentrate on royalty-free specifications for Web technologies ("Recommendations") and guidelines for their use ("Best Practices"); Incubator Groups concentrate on other, experimental topics which may be input to standardization efforts in the future.

In addition to W3C Activities there are the W3C Advisory Board and the Technical Architecture Group (TAG)[2]. The former provides guidance about management, legal matters etc., whereas the latter helps to build consensus on fundamental principles of Web Architecture (Jacobs and Walsh, 2004).

W3C as an organization is formally attached to three hosts: the Massachusetts Institute of Technology (MIT) in the USA, the European Research Consortium for Informatics and Mathematics (ERCIM) in France, and Keio University in Japan. In addition there are W3C offices in many countries to promote adoption of W3C technologies. The W3C membership can propose and drive new work items, has early access to new materials, and uses W3C as a community platform to decide new technology directions.

The development of royalty-free specifications relies on the W3C process (Jacobs, 2005) and the W3C patent policy (Weitzner, 2004). The latter describes licensing and patent disclosure requirements for the participation in W3C work and is an important factor for many organizations to decide about their engagement in W3C.

Two key aspects of W3C Recommendation development are that we aim to reach consensus, within the W3C and the public, about the actual features of new technologies, and it is required to demonstrate several interoperable implementations before publishing a Recommendation. The need for consensus sometimes slows the development process for a Recommendation, but it significantly increases the likelihood that the result will actually be accepted, implemented, and deployed in the community. An example is the XML Query language XQuery 1.0. Between the publication of the first public draft and the final publication of the Recommendation, about six years elapsed. However, during that period a large developer and user community took shape, and the Recommendation was published with around 40 implementations.

2. XML Validation: XSD 1.1

XSD, the XML Schema Definition language, is a metalanguage for defining XML vocabularies. Essentially, the author of an XSD schema provides a (regular right-part) document grammar for documents which use the vocabulary; unlike some other XML schema languages, XSD makes first-class citizens out of the *types* associated with elements and attribute. Types may be defined by restricting or extending other types, so that the class hierarchies usual in object-oriented design have a relatively natural representation in XSD schemas. A typical schema consists primarily of the definition of simple and complex types and the association of elements and attributes with types.

A number of primitive datatypes (or 'simple types') are provided: strings, booleans, decimal numbers, floating-point numbers, date-time stamps of varying precision (date-time, date alone, year, year plus month, time alone, etc.), URIs, and some others. From these, a number of other built-in types are derived by restriction, including integers and various subtypes of integer (long, short, byte, positiveInteger, etc.).

Modularization facilities are also provided for using several vocabularies in conjunction. In the usual case, one or more

[1] A general introduction to W3C can be found at <http://www.w3.org/Consortium/about-w3c.html>.

[2] See <http://www.w3.org/2002/ab/> for a description of the Advisory Board; for the TAG, see <http://www.w3.org/2001/tag/>.

schema documents are used to define the vocabulary associated with a given namespace (see (Bray et al., 2006)), and a schema is constructed by consulting one or more schema documents. The schema thus constructed will often consist of components from several namespaces.

Because they are essentially document grammars with a few additional constraint mechanisms, schemas can vary in many of the same ways that grammars can vary. They can be tight or loose, over- or under-generate vis-a-vis some given body of material. In XSD, they can also be incomplete: wildcards and 'lax validation' can be used to define constraints on some elements and attributes, while leaving others undefined and unconstrained. Schemas are often used for validating XML data as a quality assurance measure, to detect typographic and tagging errors. They are very useful in this role, but schemas can also be used in other ways. Data binding tools read schemas and generate object-oriented code to read documents which conform to the schema and from them create objects of particular classes, or serialize objects of particular classes in XML which conforms to the schema.

Schemas can document the contract between data sources and data recipients: the source typically undertakes to provide only valid data, and the recipient undertakes to accept any and all valid data. In such scenarios, invalid documents are often simply rejected. In other cases, schemas can be used to document a particular understanding of a kind of document, capturing a simple view of the 'standard' realization of the document type without allowing for all variations. In dictionaries, for example, it is a commonplace observation that there are important regularities among entries, but that a small number of entries require rather unusual structures. Grammars which allow for all of the structural variations actually encountered in a dictionary often provide no very clear account at all of the regularities which apply in 99% or more of all cases. Grammars which capture the regularities clearly often do not accommodate the deviant structures which can appear in a small number of cases. (See (Birnbaum and Mundie, 1999) for fuller discussion.) In part to assist in handling such situations, XSD defines validity not solely as a Boolean property of documents, but describes validation as providing a much richer result: each element and each attribute is individually labeled as to validity or partial validity. XSD can thus be used either for conventional prescriptive grammars or for descriptive grammars which focus on capturing the salient regularities of the material; material with the typical structures described by the grammar can be handled in one process, while anomalous structures can be detected automatically (by their failure to be valid against the schema) and handled specially. XSD does not require that applications reject documents which are invalid or only partially valid.

XSD 1.1 (see (Gao et al., 2007) and (Peterson et al., 2006)) offers a number of enhancements to XSD 1.0, most visibly the addition of

- assertions
- conditional type assignment
- open content

Assertions allow the schema author to express constraints using XPath expressions: the assertions associated with any type are evaluated for each instance of the type, and if any assertion fails to evaluate to true, the instance is not valid against the type. The most common use of assertions will be to formulate co-occurrence constraints. When declaring an element with integer-valued attributes named *min* and *max*, for example, a schema author might wish to specify that the value of the one should be less than the value of the other. This is easily accomplished with an appropriate assertion:

```
<xs:assert test="@min le @max"/>
```

Some XML vocabularies specify two or more attributes for a given element with the proviso that at most one of them may occur. This can also be handled conveniently with assertions. To ensure that either attribute *a* or attribute *b* may appear, but not both, one might write:

```
<xs:assert test="not(@a and @b)"/>
```

To require additionally that at least one of the two must appear, one might write:

```
<xs:assert test="
  (@a or @b)
  and not(@a and @b)"/>
```

The assertions of XSD 1.1 are restricted in one important way: they can refer to attributes or descendants of the element being validated, but they cannot refer to its ancestors, to its siblings, or to any elements or attributes outside the element itself. This restriction helps preserve the design invariant that the validity of an element or attribute against a given type can be tested in isolation from the rest of the document. This provides a certain context-independence of type validity, which is useful in transformation contexts like XSLT or XQuery.

Another way to capture co-occurrence constraints is to make the assignment of a given type to an element depend upon conditions to be checked in the instance. XSD 1.1 provides a form of such *conditional type assignment* based on (Marinelli et al., 2004). Here, too, the conditions which govern type assignment are given in the form of XPath expressions. To specify, for example, that the type assigned to a *message* element depends upon its *kind* attribute, given appropriate definitions of *messageType*, *string-message*, *base64-message*, *binary-message*, and *xml-message*, one might write:

```
<xs:element name="message"
    type="messageType">
  <xs:alternative
      test="@kind='string'"
      type="string-message"/>
  <xs:alternative
      test="@kind='base64'"
      type="base64-message"/>
  <xs:alternative
      test="@kind='binary'"
      type="binary-message"/>
  <xs:alternative
      test="@kind='xml'"
      type="xml-message"/>
  <xs:alternative
      test="@kind='XML'"
      type="xml-message"/>
</xs:element>
```

The third major change in XSD 1.1 to be discussed here is the provision of methods for specifying what is sometimes called ‘open content’. In defining a document grammar, one may wish to specify, for example, that a particular element must have an *a*, a *b*, and a *c* among its children, in that order, without forbidding other material to appear before, after, or between these required elements. This can be done in XSD 1.0 with judicious use of wildcards, but experience shows that this method is error-prone and apt to fail for uninteresting technical reasons.[3] XSD 1.1 allows the schema author to specify (on a case-by-case basis) that types have open content; the schema author can specify a wildcard which is notionally inserted everywhere in the content model, or allowed only at the end of the model. The result is that it is much easier using XSD 1.1 to specify vocabularies which allow arbitrary extension by others, and which can accept new material in new versions of the vocabulary without breaking existing infrastructure keyed to earlier versions of the vocabulary. If version 1.0 of the definition of a vocabulary specifies open content everywhere, then any new elements added in later versions will be accepted by 1.0 processors without difficulty (albeit also without any knowledge of their meaning).

For those charged with developing or maintaining language resources, schema languages offer numerous opportunities for finding errors in the XML transcription of material, or distinguishing material with standard, straight-forward structure from material with anomous structures, and for describing explicitly the class of documents which certain processes are allowed to produce and other processes are required to consume without error.

3. Multi-document XML Validation: SML 1.1

SML, the so-called ‘service modeling language’, is an additional validation technology built on and layered on top of XSD. See (Pandit et al., 2008).

SML was originally developed for checking the validity of models intended to describe complex sets of information-technology services; the name “Service Modeling Language” thus reflects the historical origin of the technology, but in the meantime the name has become a misnomer: SML is a generic mechanism for validation across document boundaries, and has nothing in particular to do with services or their modeling.

An SML model is a set of XML documents, some of them are *model instance documents*, which contain representations of the information being modeled, and others are *definition documents* which define schemas to be used when validating the model instance documents.

In addition to requiring XSD validation of the instance documents, SML provides several additional mechanisms for specifying constraints which can be expressed only awkwardly in XSD, or not at all.

Schematron assertions can be associated with element declarations and type definitions; the asertions are checked for each instance of the element declaration or of the type definition.

The central innovation of SML, however, is its definition of a way to validate references from one document to another. This has a number of aspects.

First, the set of validatable links is not assumed identical to the set of links in the documents: the instance documents may well contain hyperlinks which are not constrained by the SML model and need not be validated. Those inter-document links which *are* to be validated are “SML references”, indicated by the presence of the attribute-value pair `sml:ref='true'` (or its equivalent) on the element which constitutes the reference.[4]

The actual form of the link is not constrained: any systematically defined method of pointing from one XML document to an element in another XML document (all targets of SML references must be elements) may be used. Indeed, a single reference may refer to the the target element in multiple ways, each suitable for a different deployment scenario. Of course, SML processors will understand only a particular set of reference schemes; in the interests of interoperability, SML defines one schema, the SML URI reference scheme, which uses URIs to address the target of the link. XPath 1.0 expressions are as fragment identifiers, and XPath 1.0 is augmented by a `deref()` function to allow SML references to be followed. SML processors may support any reference schemes they choose, but all are required to support the SML URI reference scheme.

Having ensured that the set of links to be validated can be reliably and easily determined, SML then allows various constraints to be imposed on links.

- The `targetRequired` constraint requires that the reference resolve to an element in a document within the SML model.
- The `targetElement` constraint requires that if the reference resolves, then it must resolve to an element which bears a particular element name or “generic identifier”. (A figure reference in a conventional textual hyperlink might be required, for example, to point to a *figure* element). Other elements declared in the schema as substitutable for the named element may, of course, be substituted.

[3] XSD requires that content models (i.e. the right-hand sides of production rules in the document grammar) be ‘deterministic’, i.e. that they not require lookahead. Whenever a wildcard is placed between elements *a* and *b* in the content model, if that wildcard also matches elements named *b*, the result is likely to be a violation of the determinism rule. The content model can normally be rewritten to avoid the problem, but this often proves tedious.

[4] As of this writing, SML references are required to be elements; this can make the current version of SML unsuitable for describing existing vocabularies in which a single element has multiple hyperlinks to other locations, expressed for example by different attributes.

- The `targetType` constraint requires that if the reference resolves, then it must resolve to an element governed by a particular XSD type, or by some type substitutable for it.

- The `acyclic` constraint specifies that the references bound to a particular declaration must not form a cycle. If a catalog of university courses, for example, uses a particular form of hyperlink to refer from a course to its pre-requisite courses, then the pre-requisite link should be checked to make sure that no course it (transitively) among its own pre-requisites.

Language resources sometimes are stored in single large documents, and sometimes in many smaller documents, and the choice between monolithic or fragmented representations often depends heavily on external factors rather than upon any logic intrinsic to the material. It is convenient, in such situations, to allow the material to be realized in either form, without losing the ability to validate it. SML's ability to validate across XML document boundaries is a useful way to ensure that relations within the data can be validated whether in a single document or in many.

4. XML Analysis: "XQuery 1.0 and XPath 2.0 Full-Text 1.0"

"XQuery 1.0 and XPath 2.0 Full-Text 1.0" (Amer-Yahia et al., 2007) is a specification that defines full-text search capabilities. It provides various full-text expressions and options to be used from within XQuery 1.0 or XPath 2.0 expressions. The options relate to stemming, thesauri, the use of stop words, and so forth, as well as to distances (for example "within five words") and units (words, sentences, paragraphs). They also support ranking and relative weighting of sub-expressions.

The specification does not dictate specific algorithms for full-text search implementations, but instead describes only the results of operations. As a consequence, one must expect some variation between implementations. From the point of view of linguistic research, this variation means that it is important to determine, whether through documentation or experimentation, the exact facilities provided by any given implementation.

The tokenization algorithm splits the input data into a sequence of tokens, which, conceptually, are then indexed; the same tokenizer is used to parse queries at run-time into sequences of tokens to be matched against the index. The tokenizer is expected to recognise *xml:lang* attributes and to perform multilingual matching as necessary.

At the time of writing, this specification is a Last Call Working Draft; a formal call for implementations is expected in May of 2008, and so although there are already some implementations, there may be changes in the final specification as a result of implementation experience.

Probably the biggest limitation of the current Full Text draft for language research is the lack of introspection: one cannot find out exactly which token or tokens matched the query, and one cannot directly implement match highlighting in the way one might want for a concordance or keyword-in-context index. Some implementations do provide a way to do this, and a future version of the specification may well standardize it, but for now it represents a severe limitation.

The limitation is greatly ameliorated when one considers that XQuery (like XPath 2.0 itself) operates not only on XML files, but on any data that can be represented as XPath and XQuery Data Model (XDM) instances. This includes for example geospatial data, relational data, RDF, and more. As a result, one can perform joins across different types of database, correlating them with efficient text searching. In summary, "XQuery 1.0 and XPath 2.0 Full-Text 1.0" (Amer-Yahia et al., 2007) will be a valuable tool to researchers, including the language resources community.

5. XML Processing: "XProc: An XML Pipeline Language"

Using XML to represent language resources has become the norm. Actually processing language resources for some purpose often consists of a sequence of processing steps which split, merge, restructure, and transform XML. "XProc: An XML Pipeline Language" (Walsh et al., 2007) is a specification which provides an XML vocabulary for specifying just such sequences, together with an inventory of both simple structural manipulations such as renaming, wrapping, deleting, and extracting items in an XML data stream, as well as larger-scale standards-based operations such as validation, transformation, and querying (see above).

Many kinds of XML technology and standards, including XSLT, XML Schema, XInclude, XQuery, and even SOAP and WSDL, can be understood as mapping from one kind of infoset (Cowan and Tobin, 2004) to another. Today most implementations of XML-based language processing applications process XML directly using programming languages and an API such as SAX or DOM. But many XML processing tasks don't *need* to be done at this low level. There are a number of XML Pipeline languages already available which allow you to specify sequences of standards-based XML operations. It is often possible to replace programming-language-based XML processing with short and simple XML Pipeline descriptions, for example

- schema-validate
- then do XInclude
- then transform with a stylesheet
- then send to a server via SOAP
- then validate the result
- then transform with another stylesheet

The W3C's XML Processing Model WG is working to produce an interoperable XML Pipeline language based on existing technology. The work has nearly finished, and the result should provide a language with wide application to language processing tasks.

The XProc language is itself expressed in XML, and has two main parts:

1. A language for specifying the sequence and configuration of processing steps;
2. A collection of built-in step types, including both low-level structure-manipulation and high-level standards-based operations

There is support for more than just straight-through pipelining of operations, with data-flow equivalents of conditions, loops, and exception-handlers.
The low-level manipulations available include

- sub-tree deletion
- element and/or attribute renaming
- sub-tree wrapping and unwrapping

The higher-level operations available include

- XInclude
- XSLT
- http-request

The pipeline paradigm for producing NLP systems has been heavily exploited by the Language Technology Group at the University of Edinburgh. One example, described in (B. Alex, C. Grover *et al.*, 2008), uses a multi-step pipeline to extract named entities, in particular proteins, from biomedical text, classify the extracted terms, and detect relations between terms. Steps in the pipeline range for generic, low-level tasks such as tokenisation and sentence-boundary detection to high-level processes such as relation extraction which involve not only entity-tagged data but also precomputed statistical models.
The availability of a standardised XML pipeline language offers a real opportunity to improve the principled comparison of alternative approaches, as the modular nature of the pipeline architecture, together with the well-defined interfaces between modules which the XML document structures represent, will make it possible to do properly controlled comparisons of alternative approaches to the key stages in a complex process.

6. Internationalized Formatting: "XSL Formatting Objects (XSL-FO)"

The XSL 1.1 specification (Berglund, 2006) includes facilities for formatting XML, for example into PDF. XSL-FO 2.0 is currently being designed, with increased sophistication and also with increased support for Japanese formatting. The XSL-FO 2.0 requirements document (Bals, 2008) provides more information. XSL-FO is currently the most powerful and most completely internationalized of any widely-used standard for text formatting, with strong support for mixed-language work.
XSL-FO is a fixed XML vocabulary for formatting. In normal use one transforms input XML into the XSL-FO vocabulary using XSLT, and this transformed XML document is then rendered. It is also possible to produce XSL-FO directly, for example using XQuery.
XSl-FO copes with an arbitrary mix of text directions, both in what it calls the inline progression direction (e.g. right-to-left for Hebrew) and in what it calls the block progression direction (e.g. top to bottom for English, or right-to-left for vertical Japanese). It also defines how baselines should be mixed, for example when combining Devanagari and Arabic on the same line. Since language information is available to the formatter, language-specific hyphenation and line-breaking is also generally used.
Currently, XSL-FO is primarily aimed at automatic formatting in a content-driven environment: text flows into page areas, and new pages are created on demand from templates. XSL-FO 2.0 is expected to add support for format-driven processing, in which page areas fetch content as needed, but the 2.0 work is still in the early stages.
Readers interested in the future of XSL-FO are strongly encouraged to inspect the requirements document previously cited and to send comments to the Working Group as instructed in the Status section of that document.

7. XML Internationalization and Localization: "ITS 1.0"

The Internationalization Tag Set (ITS) 1.0 (Lieske and Sasaki, 2007)[5] is a specification which provides an XML vocabulary related to Internationalization and Localization of XML. A prototypical use is specifying which parts of an XML document should be translated or not translated during the localization of XML data. Such information can be expressed with two approaches, which can be used alternatively or complementary. First, *locally*, by adding in an XML document a *translate* attribute to the targeted element node, with the values *yes* or *no*. Second, by describing ITS 1.0 *global rules* which are independent of a specific location and can be applied to several XML documents. Such rules make use of XPath to specify the nodes to which the ITS information should pertain to.
ITS 1.0 specifies for 7 so-called "data categories" a way to express global and local information, defaults, and inheritance behavior (that is, does the information pertain to attributes and / or child elements):

- "Translate" to separate translatable from non-translatable content;
- "Localization Note" to communicate notes to localizers;
- "Terminology" to identify terms and optionally associate them with information, such as definitions;
- "Directionality" to allow the user to specify the base writing direction of blocks, embeddings, and overrides for the Unicode bidirectional algorithm;
- "Ruby" as a run of text which is associated with another run of text (the base text) and used to provide e.g. reading (pronunciation) guidance;

[5] The companion document (Savourel et al., 2008a) describes among others how to apply ITS 1.0 to new and existing XML formats.

- "Language information" to express the language of a piece of content;

- "Elements within Text" to specify the flow characteristics of an element, that is e.g. whether it is part of the flow of its parent, or is nested in a parent element and constitutes an independent flow.

Such information can be applied in many scenarios, for example within localization tools, for the extraction of translatable text, or as a preparation for the localization process. Below are two examples of using ITS 1.0 together with two important standards for XML localization: XLIFF and TBX.

XLIFF (Savourel et al., 2008b) is the "XML Localization Interchange File Format", a interchange file format for localizable content. A prototypical usage scenario is that out of some input data (e.g. an XML document) an XLIFF document is being generated, containing several translation units which wrap markup for the input "source" data and output "target" translations. Translators then create these translations, which are finally integrated into the source data.

The ITS 1.0 data category "Translate" can be used to generate XLIFF documents out of XML input data. Below is an XProc (see sec. 5.) pipeline informally described with the purpose of comparing results of translation tools. It consists of the following XProc steps:

- a step "extract translatable text" for the creation of an XLIFF document out of XML input and ITS 1.0 "Translate" information;

- an automatic translation step, using the XLIFF document and executed with a variety of tools;

- a step for comparing the results of the previous step for the different translation tools.

The use of ITS 1.0 as the first step in the XProc pipeline description helps to "hide" the specifics of input XML data. The automatic translation tools only have to understand the XLIFF format. In this way, the same processing chain can easily be re-used for new translation tools.

The TermBase eXchange (TBX) format is another example where ITS 1.0 helps to generalize processing chains. TBX is used for the representation of terminological information for human consumption or in NLP lexicons. The ITS 1.0 "Terminology" data category helps to identify terms locally or with global rules. In combination with the "Translate" data category, the following processing chain can be envisaged:

- All content which is described as a term via ITS 1.0 information is used to generate a terminological entry in a TBX file.

- The "Terminology" data category allows for adding information to selected terms, e.g. definitions. If such information is given, the information is also added to the terminological entry.

- An optional step relies on the "Translate" data category: for whose terms which are described as translatable, additional, language-specific markup in the terminological entry is generated. The content of this markup has to be filled by the localizer / translators, depending on the target language(s).

8. Outlook: The need for the Integration of Language Resources

The language resources community has struggled for years with the challenge of combining separately developed resources: how to combine your lexicon with mine, or your grammar, corpus etc. This problem is sometimes termed data integration. There are several areas of difficulty in the field of data integration. Some of these have been solved, at least partially, by some of the technologies that have been described in this paper:

- Accessing data from various sources in a uniform manner, so that they can be processed together; XQuery, XPath 2 and XSLT can combine data from (for example) XML documents, relational databases, geospatial databases and more. Current database management systems often allow data to be viewed either as relations or as XML, without requiring any particular effort beyond requesting the XML view. Data from other sources can often be accessed as if it were XML by interposing 'XML lenses' between the data and the consumer. In some cases, URI resolvers are modified to interpose the lens between the data source and the user, so that no active intervention by the user is required.

- Obtaining data (once accessed) in a uniform format, so that it can be processed uniformly; W3C XML Schema can be used to describe XML data, including embedded Internationalization information, and once everything is in XML, XML tools can be used.

- Mapping relationships between hierarchies; this is an unsolved research problem in general, and can be described as the difficulty of combining arguments made from differing and incompatible viewponts. W3C has an Ontology Language, OWL, but this in itself describes any single ontology, not relationships between ontologies. However, when XML documents are marked up with two different vocabularies, it is often possible in practice to write a declarative mapping function as an XSLT styleheet to transform between them.

- Processing; one is sometimes dealing with large amounts of data when handling language corpora. However, what may be voluminous to one observer might be miniscule to another: there are relational databases with petabytes of data that are processed on a daily basis. High volume processing with XQuery is in its infancy, but it is progressing fast. As a practical matter, this means that different XQuery implementations may offer very different performance characteristics, and it will often be useful to experiment with more than one, to find one that fits a particular deployment scenario.

- Storing results; when the results must be stored to disparate databases, the underlying technology may have to translate formats automatically; the XQuery Update Facility is currently (May 2008) in a call for implementations, but promises to offer this functionality, saving changes both to files and to databases or other sources of structured information.

- Presenting results; here XSL-FO is a strong contender, with a number of implementations.

In all cases, the fact that every XML tool can process any XML document is a major benefit that greatly simplifies work.

9. References

S. Amer-Yahia, C. Botev, S. Buxton, P. Case, J. Doerre, M. Holstege, J. Melton, M. Rys, and J. Shanmugasundaram (eds.). 2007. XQuery 1.0 and XPath 2.0 Full-Text 1.0. Technical report, W3C. See <http://www.w3.org/TR/2007/WD-xpath-full-text-10-20070518/>.

B. Alex, C. Grover *et al.* 2008. Assisted curation: Does text mining really help? In *Pacific Symposium on Biocomputing 13*, pages 556–567.

K. Bals (ed.). 2008. Extensible Stylesheet Language (XSL) Requirements Version 2.0. Technical report, W3C. See <http://www.w3.org/TR/xslfo20-req/>.

A. Berglund (ed.). 2006. Extensible Stylesheet Language (XSL) Version 1.1. Technical report, W3C. See <http://www.w3.org/TR/2006/REC-xsl11-20061205/>.

D. J. Birnbaum and D. A. Mundie (eds.). 1999. The problem of anomalous data. *Markup Languages: Theory and Practice*, 1(4):1–19.

Tim Bray, Dave Hollander, Andrew Layman, and Richard Tobin (eds.). 2006. Namespaces in xml 1.1 (second edition). Technical report, W3C.

J. Cowan and R. Tobin (eds.). 2004. Xml information set (second edition). Technical report, W3C. See <http://www.w3.org/TR/2004/REC-xml-infoset-20040204/>.

S. Gao, C. M. Sperberg-McQueen, and H. S. Thompson (eds.). 2007. W3C XML Schema Definition Language (XSDL) 1.1 Part 1: Structures. Technical report, W3C. See <http://www.w3.org/TR/2007/WD-xmlschema11-1-20070830/>.

I. Jacobs and N. Walsh (eds.). 2004. Architecture of the World Wide Web, Volume One. Technical report, W3C. See <http://www.w3.org/TR/2004/REC-webarch-20041215/>.

I. Jacobs (eds.). 2005. World Wide Web Consortium Process Document. Technical report, W3C. See <http://www.w3.org/2005/10/Process-20051014/>.

C. Lieske and F. Sasak (eds.). 2007. Internationalization Tag Set (ITS) 1.0. Technical report, W3C. See <http://www.w3.org/TR/2007/REC-its-20070403/>.

Paolo Marinelli, Claudio Sacerdoti Coen, and Fabio Vitali. 2004. Schemapath, a minimal extension to xml schema for conditional constraints. In *Proceedings of the Thirteenth International World Wide Web Conference*, pages 164–174, New York. ACM Press. Available on the Web in the ACM Digital Library; citation at <http://portal.acm.org/citation.cfm?doid=988672.988695>.

Bhalchandra Pandit, Valentina Popescu, and Virginia Smith (eds.). 2008. Service Modeling Language, Version 1.1. Technical report, W3C. See <http://www.w3.org/TR/sml/>.

D. Peterson, P. V. Biron, A. Malhotra, and C. M. Sperberg-McQueen (eds.). 2006. XML Schema 1.1 Part 2: Datatypes. Technical report, W3C. See <http://www.w3.org/TR/2006/WD-xmlschema11-2-20060217/>.

Y. Savourel, J. Kosek, and R. Ishida (eds.). 2008a. Best Practices for XML Internationalization. Technical report, W3C. See <http://www.w3.org/TR/2008/NOTE-xml-i18n-bp-20080213/>.

Y. Savourel, J. Reid, T. Jewtushenko, and R. M. Raya (eds.). 2008b. XLIFF Version 1.2. Technical report, OASIS. See <http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html>.

N. Walsh, A. Milowski, and H. S. Thompson (eds.). 2007. XProc: An XML Pipeline Language. Technical report, W3C. See <http://www.w3.org/TR/2007/WD-xproc-20071129/>.

D. J. Weitzner (ed.). 2004. W3C Patent Policy. Technical report, W3C. See <http://www.w3.org/Consortium/Patent-Policy-20040205/>.

A Case Study Related to Standardization – Codes for the Representation of Names of Languages

Jennifer DeCamp
MITRE
7515 Colshire Drive, McLean, VA 22102 USA
E-mail: jdecamp@mitre.org

Abstract

This article provides a case study of the development of the International Organization for Standardisation (ISO) *Codes for the representation of names of languages -- Part 3: Alpha-3 code for comprehensive coverage of languages* and on related work for metalanguages and dialects. It reviews the requirements for and uses of this standard, and issues encountered with its development and implementation. .Development issues included lack of consensus on what constituted a language. Implementation issues included legacy codes and technology constraints, leading to continued use of older standards. Issues also included making the standard broadly available and coordinating with multiple organizations, including the World Wide Web Consortium (W3C) and the International Engineering Task Force (IETF), as well as industry.

1. Introduction

Language codes are trigraphs, digraphs, or other character sequences used to designate languages, such as “de” or “deu” (ISO 639-2, 1988; ISO 639-2; 1998) for “German”. This article is a case study of the development of the ISO 639 *Codes for the representation of names of languages -- Part 3: Alpha-3 code for comprehensive coverage of languages* and of related work for metalanguages and dialects. It reviews the requirements for and uses of the standard, and issues encountered with its development, coordination, implementation, and dissemination.

2. Requirements

Language codes were originally required due to limitations of database structure, where it was necessary to have identifiers of consistent length. They were also needed in order to provide unique identifiers for a language (at least, a unique identifier within a single standard and/or set of standards). Language codes have also been useful in providing a designator to cover the many spellings, translations, and other variations of a language name. For instance, “Deutsch,” “Allemande,” or “German” can all be designated by “de”.

A survey of the use of language codes (DeCamp, 2001) showed an increasing interest by industry, governments, and academia in many additional languages and a need to have unique identifiers for those languages. There was also interest in a system where additional unique language identifiers could be easily accessed to cover other languages.

There was considerable interest in an ontology of metalanguage, language, and dialect and/or in related languages, language, and dialect. Such an ontology could be useful in identifying alternative resources (tools, people, etc.) when the requested resource was not available. For instance, if no resources were available in Egyptian Arabic, it would be helpful to identify a more generic form of Arabic (e.g., the metalanguage or language family “Arabic”) and/or a related language (e.g., “Standard Arabic”). The survey showed needs for language identifiers to designate content (e.g., collections, documents, abstracts, musical notation), software (e.g., Microsoft keyboards; spell checkers), and people (e.g., first language; training). The identifiers were needed for the application of tools to text (e.g., spell checkers, machine translation). They were also needed in order to track skills (e.g., to track their education in specific languages). Similar findings were obtained by Phillips and Davis (2006) for needs to designate user preferences for a particular language or a prioritized set of languages when receiving of information on computers, cell phones, etc.

These applications required language and language-related information that could include:

- Metalanguage, usually language family
  (e.g., Arabic)
- Language
  (e.g., Arabic, Egyptian Spoken)
- Script
  (e.g., Arabic script)
- Font
  (e.g., Tahoma)
- Transliteration system
  (e.g., Board of Geographic Names)
- Phonetic representation
  (e.g., International Phonetic Alphabet)
- Historical time period
  (e.g., prior to 1900)
- Modality
  (e.g., speech)
- Genre
  (e.g., news broadcast)

3. Existing Standards

When ISO undertook the development of ISO 639-3, numerous standards existed (and still exist) for language identifiers, but none that met the above requirements. In each system, there were also inconsistencies of level, with a mixture of metalanguages, geographic groups of languages, languages, and dialects. Discussion of issues with ISO 639-1 and ISO 639-2 codes is provided in Constable and Simons, 2000.

3.1 MARC and ANSI/NISO Codes

Library of Congress MARC language codes and the American National Standards Institute (ANSI)/National International Standards Organization (NISO) language codes are oriented towards bibliographic communities. These communities have since adopted the ISO 639-B standard described below or ISO 639-3, discussed later in this paper.

3.2 ISO 639-1

The ISO standard for two-character codes (ISO 639-1) dates from 1988 but was updated in 2004. There have been some additions to ISO 639-1, and the standard is still in wide use. The two character codes allow for 676 designators (26 x 26), which could at best cover only 9 percent of the 6912 languages listed in *The Ethnologue* without metalanguages or dialects. In 2001, there were 168 actual language codes, although a small number have been added since.

3.3 ISO 639-2

ISO 639-2 for trigraphs or three-letter codes dates from 1998, providing 17,547 code points (26 x 26 x 26). In 2001, there were 444 codes for 437 languages, with some repetitive cross-referenced entries (e.g., “Low German”, and “German, Low”.

ISO 639-2 had two variants: one for the bibliographic or library community and one for the terminological community (i.e., everyone else). For instance, for German, the Bibliographic or “B” code is “deu” and the Terminological or “T” code is “ger”. There are 23 such variants.

The United States Library of Congress is the registration authority for both versions of ISO 639-2 and functionally for ISO 639-1, as it is harmonized with 639-2. In 2001, there were strict requirements for adding languages, including providing fifty documents in the proposed language. In some oral languages (e.g., Pashto), it was difficult and time consuming (particularly in the United States) to meet this requirement. These rules have since been relaxed.

3.4 SIL Ethnologue Codes

The Summer Institute of Linguistics (SIL) *Ethnologue* (*The Ethnologue v. 14*) is well established in the academic community as a source of language identifiers. *The Ethnologue* has extensive coverage of languages. It also has at least the beginnings of an ontology with its language lineages (charts of languages and their language families) and with lists of dialects per language.

While there have been disputes about designations of languages, metalanguages, and dialects, there is an established process for anyone to submit concerns and for the codes to be changed. In addition, the detailed documentation of *The Ethnologue* book (*The Ethnologue* v. 15, 2005) and website provide a working terminology, from which change can be managed.

The Ethnologue also provides supplemental information useful to people working with language codes. Such information included alternative names for the language or metalanguage. A user who did not know the official reference name could thus search on an alternative name and find the relevant code. In addition, *The Ethnologue* provided useful information about the language, such as number of speakers.

3.5 LinguaSphere and GeoLang Codes

An effort was started by David Dalby to provide a four-character system of dialect codes, based on geographic rather than linguistic distribution. A spin-off effort was GeoLang, which is administered by the World Language Documentation Center (WLDC). The four-character format enables the encoding of substantially more information, with 456,976 (26 x 26 x 26 x 26) codes.

3.6 BCP 47 and Internet Assigned Number Authority (IANA) Language Tags

In parallel with the ISO efforts, W3C has been working on the syntax for combining information about language (BCP 47), based on language tags. These language tags are intended:

> “To help identify languages, whether spoken, written, signed, or otherwise signaled, for the purpose of communication. This includes constructed and artificial languages but excludes languages not intended primarily for human communication, such as programming languages” (BCP 47).

A language tag “consists of a ‘primary language’ subtag and a (possibly empty) series of subsequent subtags, each of which refines or narrows the range of languages identified by the overall tag.” The primary language tags are ISO 639-1 codes where they exist, and then ISO 639-2T codes where they exist. An update to BCP47 is in progress to add ISO 639-3 codes to supplement the ISO 639-1 and 639-2T codes. The ISO 639-1 and relevant 639-2T codes are entered in the IANA Language Tag registry, which is cited by BCP 47 as the official source of language identifiers.

The syntax for designating preferences in languages (e.g., with which to view email) would be the ISO 639-2 language codes in a series separated by commas. “English before French before Chinese written in the Traditional script” would be written as "en, fr, zh-Hant" *(*BCP 47*)*.

Information about a language could be designated with codes for language, script, and country or region and up to three subtags for other items. Information not used or not necessary would be deleted (BCP 47).

4. The Development of ISO 639-3

In 2001, work began in ISO Technical Committee (TC) 37 on Terminology and Language Resources to provide a more extended and homogeneous code set. The U.S. delegation proposed using *The Ethnologue v. 14* as the basis for this new code sets.

Concerns were raised that The *Ethnologue v. 14* was not harmonized with ISO 639-1 and ISO 639-2. SIL agreed to make changes to its coding system to use the ISO 639-2 Terminology (three-letter) codes. Thus the ISO 639-2T codes and the ISO 639-3 codes would be identical. There was already a mapping between ISO 639-2T and ISO 639-2B and between these codes and ISO 639-1. The changes in the SIL codes are now reflected in *The Ethnologue* v. 15, which has replaced Version 14 on *The Ethnologue* website. ISO 639-3 became a full ISO standard in 2007.

Codes in *The Ethnologue* v. 15 designating languages were included in the draft ISO 639-3. Codes in ISO 639-3T that were really metalanguages (i.e., designating two or more languages) were moved to a draft ISO 639-5 for *Codes for the Representation of Names of Metalanguages*. However, all metalanguages described in *The Ethnologue* (particularly in the lineage descriptions) had not been assigned Ethnologue codes. These metalanguages are not yet included in ISO 639-5. Their inclusion would be helpful in developing a language ontology.

ISO 639-5 is in Final Draft International Status (FDIS), having passed final voting, and is now in the one year usage period before being declared a full standard. The registration authority (i.e., the group to manage requests for changes or additions) is SIL with overview by ISO TC37. Further discussion will be provided of registration authorities.

The U.S. delegation to TC 37 submitted a proposal to use the different layers of metalanguage outlined in *The Ethnologue v. 15* for a standard on metalanguage (i.e., ISO 639-5), and to use the dialects provided underneath each language for a standard on dialects (i.e., ISO 639-6). However, considerable work was needed to add codes to much of the metalanguage and dialect data. Moreover, there was controversy in the international language community about *The Ethnologue* concerning languages vs. metalanguages vs. dialects.

Meanwhile the British Institute of Standards (BIS) submitted a proposal to use a four-letter code system for ISO 639-6 that was being developed by LinguaSphere. This task was later transferred from LinguaSphere to GeoLang. There have also been discussions of providing an eight-character format with a potential 208,827,064,576 code points (Huilsted, 2006).

GeoLang is preparing a detailed cross-walk between ISO 639-3 languages and ISO GeoLang dialects. A detailed third-party review is need of the cross-walk to ensure seamless access to codes on metalanguage, language, and dialects. This standard is currently in FDIS status.

In the meantime, ISO administration approved the use of registries or databases in lieu of paper to represent the official standards. There is thus no longer a need to provide the codes in lists in standards documents. This change facilitates the more rapid updating of the codes. It also facilitates the coordination of the standards, as all the codes can be in the same database. The database currently being used is the ISO TC 37 Data Category Registry, which will be hosted by the Max Planck Institute.

Dr. Håvard Huilsted, the chair of the ISO 639 Working Group, proposed combining the registration authorities into a single organization that would include the current registration authorities and also other parties with interests and work in language identifiers (e.g., the Unicode Consortium, W3C, IETF, etc.) This meta-organization is the World Language Document Center (WLDC).

5. Development Issues

The following issues are some of those encountered in the development of the new ISO 639 standards.

5.1 Incomplete Code Sets for Language-Related Information

There are international standards codes for scripts (ISO 15924), countries (ISO 3166), currencies (ISO 15924) and transliterations (e.g., ISO 9:1995). However, some of these standards only cover a small set of languages/cultures. In particular, ISO transliteration standards need to be expanded and updated to better reflect current usage and/or additional languages of interest.

5.2 Conflicting Standards for Some Requirements

In some cases, there are conflicting standards for particular requirements. This problem is most evident in transliteration standards, where the ISO standards are only one of many sets of transliteration values. Designation of transliteration is particularly difficult because of the proliferation of standards, sometimes applying only to a specific database. .

5.3 Lack of Consensus on Language vs Dialect

There has frequently been a lack of consensus on what is a language vs. metalanguage vs. a dialect.

Improvements have been made to subsequent versions of ISO 639 to provide more linguistic consistency and definition. For instance, the code for Amharric was raised from a language code to a metalanguage code, as the past code been used to cover five languages (one of which was Amharic). As per ISO 639-3 draft standards on procedures for managing language codes (draft ISO 639-4), codes are not reused when there are changes. There is also a set of user codes, whereby the user or a specific user community can provide identifiers for the metalanguage, language, or—in some cases—the dialect.

5.4 Need for Additional Review

Several organizations have adopted ISO 639-3 in its draft form, which has enabled them to provide feedback. Moreover, the ISO 639-3 and 639-5 codes were added to the TC 37 Data Category Registry, which helped to ensure the lack of redundancy and the possibility of confusion from the same three letters being used for different purposes, at least within ISO TC37 standards. However, additional review is needed of the relationship of metalanguage, language, and dialect codes within ISO 639 and in relationship to other standards in ISO and other organizations.

6. Implementation Issues

The following issues are some of those encountered in the development of the new ISO 639 standards.

6.1 Technology Constraints

A key problem in the early adoption of the codes by the U.S. Government concerned legacy software. Databases were set for ISO 639-2 with two character spaces. This limitation of two character spaces could not be changed until the database was replaced, at which time the three letter codes could be implemented.

There is still an issue in some databases of no expansion capability to the four-character codes for dialects or the strings of codes to designate a language in a region with a certain dialect. For instance, according to BCP 47, a language such as Swiss French could be identified using a LANGUAGE plus a REGION, such as “fr-CH”. Canadian French would similarly be “fr-CA”.

6.2 Lack of Specification in Other Standards

RFC 4646 does not yet reference ISO 639-3, but probably will by the time this article is published.

6.3 Mapping to Legacy Codes

Numerous issues came up with mapping legacy codes to existing codes, particularly given the lack of documentation regarding the original intent of many of the language identifiers. There were extensive problems with geographic rather than linguistic designations and with historic/antiquated names. For instance, one legacy term was “Formosan,” which is not only a geographic area rather than a language (and in fact a geographic area with many languages), but is also a historical term for the country that became “Taiwan”. Given no documentation of the meaning of the term, it was impossible to tell if “Formosan” designated Mandarin Chinese on the island now known as Taiwan, or designated all the languages of Taiwan, or designated all the languages of Taiwan during the time when the name of the country was “Formosa”. Even the most logical answer was suspect, since it may not have been what was originally intended or what was later interpreted and used. The only way to obtain authoritative answers was to review the materials designated by the identifiers.

6.4 Erroneous Uses of “LANG”

A 2004 survey (DeCamp, 2004) of the values for the metadata “LANG” on the internet demonstrated examples of non-standard use, including:

- Use of “LANG” to designate programming language (e.g., “LANG=Visual Basic”)
- Use of “LANG” to indicate encoding (e.g., “LANG=UTF8”)
- Use of “LANG” with non-standard values (e.g., “LANG=Arabic”)

Such uses indicate needs for better training and tools.

7. Conclusion

The development and implementation of ISO 639-3 resulted in extraordinary communication and collaboration across multiple standards organizations, industry, and academia. Such efforts have resulted in liaisons to ISO from W3C, IETF, and the Unicode Consortium and vice versa, with representation at many development meetings.

However, continuing coordination is needed, including within ISO 639, as the dialect codes are completed. Coordination is needed with other organizations involved with language identifiers, including NISO, MARC (the Library of Congress), and IANA. It is also needed with W3C, IETF, and the Unicode Consortium, and with industry, academia, and governments worldwide.

Even with this coordination, there is now a situation of having multiple full or draft standards with close but not exact correlations. Examination is required to see if the dialects listed in *The Ethnologue* have precise correlations with GeoLang’s efforts on ISO 639-6, or if additional codes could be added to ISO 639-6 to provide this correlation.

Issues need to be examined on whether use of *The Ethnologue version. 15* as a consistent standard and beginning ontology for metalanguages, languages, and dialects would be more effective than combining systems (e.g., *The Ethnologue* and GeoLang). Alternatively, could a consistent set of metalanguages, languages, and dialects be provided by GeoLang as a single standard (i.e., replacing ISO 639-3 which is currently based on *The Ethnologue*)?

These issues raise larger questions about standards development, including whether it is possible to achieve single international standards or whether our diversity of cultures and organizations will result in many standards. How will technology cope with multiple standards (e.g., returning to the system of namespaces)?

Will one standard emerge from many as in the case of the Web Ontology Language (OWL), and if so, what causes a standard to emerge (e.g., citation in other standards; implementation in popular products)? Finally, there is a need to look at whether there ways to avoid or reduce the enormous amount of labor involved in having redundant efforts.

8. Acknowledgements

Many people have devoted extensive time and expertise to the ongoing development of ISO 639 codes for languages, metalanguages, and dialects. Among the people in ISO are: Håvard Huilsted, Gerhard Budin, Peter Constable, Gary Simons, Sue Ellen Wright, Christopher Cox, Debbie Garside, and Lee Gillam. There have also been extensive efforts on related standards, including by W3C, IETF, IANA, ANSI, NISO, and other organizations. Special thanks go to Addison Phillips, who provided extensive information on W3C standards.

9. References

Best Current Practice (BCP) 47 Currently a concatenation of RFC 4646 (2006) and 4647.(2006) Internet Society. http://www.rfc-editor.org/rfc/bcp/bcp47.txt

Constable, P. (2002). Toward a model for language identification: defining an ontology of language-related categories. *Ethnologue Monograph*, http://www.sil.org/silewp/abstract.asp?ref=2002-004

Constable, P. and Simons, G. (1999). An analysis of ISO 639: preparing the way for advancements in language identification standards. *Ethnologue Monograph*, http://www.sil.org/silewp/abstract.asp?ref=2002-004

Constable, P. and Simons, G. (2002). Mapping Between ISO 639 and the SIL Ethnologue. *Ethnologue Monograph,* http://www.sil.org/silewp/abstract.asp?ref=2002-004

DeCamp, J.A. (2001). Issues and Proposals for Language Tags. *In Proceedings of the Nineteenth International Unicode Conference (IUC19)* San Jose, CA: Unicode Consortium.

DeCamp, J.A. (2004). A Survey of User Practices Regarding Values for "LANG"; In *Proceedings of the Twenty-Second International Unicode Conference (IUC22),* San Jose, CA: Unicode Consortium.

DeCamp, J.A. (2001). "Towards Common Language Codes." ISO Document N78, presented to ISO TC 37 SC2, http://www.iso.org

The Ethnologue, version 14, http://www.ethnologue.org

The Ethnologue, version 15 (2005). http://www.ethnologue.org

Huilsted, H.H. (2006). Language Codes. Presentation to ISO TC 37 SC2, http://www.iso.org

ISO 639-1: 2002 *Codes for the representation of names of languages -- Part 1: Alpha-2 code for the representation of names of languages*. http://www.loc.gov/standards/iso639-2/php/code_list.php

ISO 639-2:1998 *Codes for the representation of names of languages - Part 2: Alpha-3 code.* http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/index.html

ISO 639-3:2007: *Codes for the representation of names of languages - Part 3: Alpha-3 code for comprehensive coverage of languages* http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39534

ISO/DIS (Draft International Standard) 639-4 (2007) *Code for the representation of names of languages - Part 4: Implementation guidelines and general principles for language coding* http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_tc_browse.htm?commid=48124

ISO/FDIS (Final Draft International Standard) 639-5, 2007. *Codes for the representation of names of languages -- Part 5: Alpha-3 code for language families and groups* http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39536

ISO/FDIS 639-6:2007. *Codes for the representation of names of languages - Part 6: Alpha-4 code for comprehensive coverage of language variants* *http://en.wikipedia.org/wiki/ISO_639-6*

ISO 3166. *Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes.* http://userpage.chemie.fu-berlin.de/diverse/doc/ISO_3166.html

ISO 4217. *Codes for the representation of currencies and funds.* http://www.bsi-global.com/Technical+Information/Publications/_Publications/tig90x.doc

ISO 8601. *Data elements and interchange formats -- Information interchange -- Representation of dates and times* http://en.wikipedia.org/wiki/ISO_8601

ISO 12620-2:2003. *Computer applications in terminology - Data categories for terminology collections* http://en.wikipedia.org/wiki/ISO/TC37

ISO 12620-3:2007. *Computer applications in terminology - Data categories for electronic lexical resources* http://en.wikipedia.org/wiki/ISO/TC37

ISO 15924: *Country Names and Code Elements*. http://www.iso.org/iso/en/prods-services/iso316 ma/02iso-3166-code-lists/index.html

RFC 4646: (2006) *Tags for the Identification of Languages*, Phillips, A., Ed., and M. Davis, Ed. http://www.ietf.org/rfc/rfc4646.txt

RFC 4647: (2006) *Matching of Language Tags*, Phillips, A., Ed., and M. Davis, Ed. http://www.ietf.org/rfc/rfc4647.txt

W3C *Understanding the New Language Tags*, 2006 http://www.w3.org/International/articles/bcp47/

Web Ontology Language http://en.wikipedia.org/wiki/Web_Ontology_Language

Measuring Standards in Lexical Resources

Núria Bel, Santiago Bel
IULA, Universitat Pompeu Fabra
Plaça de la Mercè, 10-12
08002 Barcelona
E-mail: nuria.bel@upf.edu, santiago.bel@upf.edu

Abstract

In this paper we want to express our opinion that the next, necessary step in the standardization of language resources is the development of instruments designed to avoid a subjective interpretation of the proposed standards, particularly in the encoding of lexical resources. The history of science and technology shows that a full deployment of a standard, that is, when standards turned out to be really used and useful, started with the creation of instruments that measured uncontroversial, trusted and standardized values.

1. Introduction

In this paper we want to express our opinion that the next, necessary step in the standardization of language resources is the development of instruments designed to avoid a subjective interpretation of the standards for language resources description, particularly in the encoding of lexical resources. The history of science and technology shows that a full deployment of a standard, that is, when standards turned out to be really used and useful, started with the creation of instruments that delivered uncontroversial, trusted and standardized values. For instance, looking at the example of cartography, the sociologist Bernard Latour (1987) identifies the steps that disciplines had to pass through to become a mature scientific or technological domain. This author argues that the creation of standards for measuring geographical positions (latitude and longitude measured, by standard, in degrees and representing angular distances from the centre of the Earth), and the creation of especially devoted instruments for the measurement of these standards (compass, astrolabe, chronometer) were a series of necessary factors for the stability of a domain that, similarly to language resources, feeds sophisticated technological applications. Thus the availability of instruments to measure geographical positions is considered the key for a "Copernican revolution" as it turned cartography into a stable domain: ready to gather as much data as possible, and data of a reliable quality because its measurements were standardized. The data gathered with these instruments could be accumulated without revisions or changes, although global changes could be applied when an error was detected. And the registers themselves could be interpreted and re-interpreted for drawing maps with an increasing level of detail. Moreover, Latour observes that although standard measurements could have usually involved a loss of information, they made it possible the integration of data collected for different purposes and at varying scales.

Although simplifying Latour's argumentation, we can say that the moral behind the history of cartography is that the collection of large amounts of uncontroversial data was the triggering factor for making possible its interpretation and representation in different levels of abstraction, more exactly in the "cascade of abstractions" that seems to be behind of a truth fruitful scientific and engineering progress. If we trust the comparison between cartography and our field, the creation of standards for language resources in general, and for lexical resources in a particular, is a necessary step, but not a sufficient one, for the full development of the disciplines and techniques that use them. To motivate the comparison just proposed, let us signal at some parallelisms between the world of language resources, specifically in the area of lexica, and the world of cartography.

Lexical resources available today cover a small portion of some knowledge domains, in the same way that maps of middle age only described the known part of the world. For example, while the terminological database of the EU, IATE, contains 1.400.000 terms for English, but an English-Spanish commercial MT system can only contain 60.000 words fully described (i.e. with syntactic and semantic information). Moreover, the information encoded does not cover all the uses of a word. Authors such as Briscoe and Carroll (1993) observed that half of parse failures on unseen test data were caused by inaccurate lexical information, and Baldwin et al. (2004) identified that in parsing 20,000 strings from British National Corpus a 40% of grammar failures were due to missing lexical entries, with a grammar dictionary of about 10,500 lexical entries. And the resources themselves, they somehow lack of proportionality, like in a ptolomeic map. For instance, WordNets for most languages offer very fine grained semantic nuances, but very little information to distinguish among them. Finally, it is also significant the similarity of the cartographical exercise with the one of lexical resources in what concerns the magnitude of the task. For both domains, it is necessary the collection and accumulation of very large quantities of data.

The question now is how to measure linguistic properties such as the ones proposed in the Lexical Markup Framework (Francopoulo et al. 2006) or in the Data Category Registry (Ide et al. 2003), the ISO standard proposals for lexical encoding, in a reliable, replicable and uncontroversial way. One does not need too much thinking to suspect that to reach uncontroversial measurements has not been easy for most of the scientific and technical domains. For instance, electrical units and standards were continuously contested through the last quarter of the 19th century. It took some time to reach the consensus and still more time to achieve that practitioners trusted the measurements obtained by instruments developed by others (Hong, 2004). But finally, the benefits of the quantitative data, the support of theoretical considerations and crucially the increasing quality of instruments specially developed for this task made the standardization of electricity possible. Therefore, we have to create instruments that can measure linguistic properties, to fix how these instruments could work as uncontroversial standard measurements and how to use them when building lexica. What follows is a first thinking of such instruments. Probably an impressionistic view of what could be an approach. We hope that this first proposal raises, at least, a period of discussion about the feasibility and the interest of this approach.

2. Standards for Lexical Resources

The use of standards for lexical resources was proposed first as the only way to the re-use of lexical data. Lexical resources have always been very costly to produce, as they are handcrafted and tied to the particular applications they were aimed to feed. It was easy to see that the impossibility of using it for more than one application was a waste of its potentiality. Large dictionaries in commercial Machine Translation systems were a clear example of the interest of reusing lexica and thus large MT companies reacted by creating OLIF (Thurmair, 2002). OLIF: Open Lexicon Interchange Format is a pioneer on how a list of companies decided to agree on a common format for interchanging lexica.

Under the influence of the precursor GENELEX (Normier and Nossin, 1990), EAGLES: Expert Advisory Group on Language Engineering Standards (and its related projects PAROLE and SIMPLE , Lenci et al. 1999) and the work done in ISLE: International Standards for Language Engineering (Calzolari et al. 2001) have been a more academic oriented proposal for the creation of general purpose lexica that can be reused for different applications with different needs and constraints.

The situation has improved with the recent proposal for an ISO standard for the encoding of lexical resources, the Lexical Markup Framework (LMF, Francopoulo et al. 2006), together with the idea of Standard Data Categories (Ide et al. 2003) both within the TC37 Committee of the ISO, and partially as a follow up of the work done for standardizing terminology. LMF has been a careful exercise for identifying lexical information that is encoded in different styles of lexica, including NLP lexica. But now that the definition for the standard is there, it seems a good moment to reflect on what is next to make its wide use the real key to the fostering of standards in the world of lexical resources in particular and language resources in general.

3. Measuring linguistic properties

The first issue should be to define what we mean by “measuring standards”. First, consider the following definition of measuring: “to give a representation of facts, by abstracting into a set of properties that can be observed”. Currently, abstracting is done by introspection, that is, some properties are assessed by a human to be enough representatives as to decide on the assignment of a conventional abstract representation. Measuring standard properties can then be to define what observations and how many observations are required for assigning a given label. The task defined in this way will enable the development of instruments that count the observations and calibrate to what extent a label can be assigned.

The key point is what properties can be observed and how they relate to the abstraction exercise done already by linguistics based on interpretation rather than in evidence in text. Obviously there are difficulties to “observe” properties of words, but this was also true for most physic phenomena for a long time. Light, colours, forces could not be observed and hence were not measurable. The development of special purpose instruments that rely on some observable information to measure properties that were not directly observable was the breakthrough. Is to focus on instruments that its only goal is the measuring of properties worthwhile in our field? We are not saying that the topic is an easy one, but neither was to measure colours, or weaves. As in other domains, we shall use our knowledge of the phenomena based on linguistic studies to determine how to do it. LMF has identified a number of properties and trusted common linguistic background to define its application. We should try to use these definitions to “measure standards”. In fact, there are properties that can be directly observed in texts and gathered with tools we already have. We can already gather observations about linguistic properties and use them in order to have stable and replicable measures.

What observations have to be done for measuring a particular linguistic property is partially a language dependent problem, but it is something that has to be done anyway as lexical resources are language dependent. Our instruments have to gather information and to recollect data according to standardized specifications parametrized for each language. For instance, for Spanish we can define morphological suffixes that have to be observed for considering a noun feminine or masculine. We can use a vector representation to register the data observed. For instance, that each component of the vector corresponds to a particular suffix. It is then a matter of calibrating the instrument to decide how many observations of these suffixes are meant to be the value *feminine*, for instance. We can make further measures by stipulating the ratio between the observed data and the value we want to define, taking also into account a scale of detail in the description of the lexical entry. We can read this relation as "for the number of observations taken into account, and the amount of information gathered, the noun is feminine". It is obvious that we must accept certain uncertainty in the measures, maybe certain error. But errors will be known and consistent, what does not happen when manually encoding. And it is also true that instruments can become more and more precise after certain cycles of accumulation, in Latour terms.

For exemplifying our approach, we have tried to define tools for measuring nominal gender and verbal transitivity in Spanish. This is just a sample of how to approach the task for further increasing the use of uncontroversial tools for building standardized resources that can be easily used by other tools. Now we will try to turn into the more technical details.

4. Morphosyntactic information

Part-of-speech taggers can already be considered instruments to measure the category of a word in context. Probably such an instrument can be used also for predicting gender and number for nouns and adjectives. But let's suppose that we want to gather data to measure these two morphosyntactic properties.

We have defined the conditions that allow us to recognize the forms of feminine and plural in a particular language, i.e. Spanish. Thus, for most of the words in Spanish, feminine gender means to have an inflection –a/-as, while masculine suffixes are –o/-os and we also have a suffix –es that can be for both genders. Our instrument will then look for occurrences of a word to decide whether it has such an inflection. It is true that there is a small proportion of words that are feminine and does not have this inflection in Spanish, but for the large majority, it will go.

Take for instance the Spanish word *casa* ('house'). We want to measure its gender, i.e. whether it is masculine of feminine, and we start by gathering data. Our tool looks for gender suffixes, as observations that can be taken from the data. One of the problems will be the ambiguity of the word endings. The '–a' suffix can belong to two Spanish paradigms, f1 and f3 as in the table below.

	singular		plural	
	masc	fem	masc	fem
f1 (funcionario, funcionaria, funcionarios, funcionarias)	o	a	s	s
f3 (fábrica, fábricas)		a		s

Table 1. Spanish inflection paradigms

Our 'casa' belongs to the f3 paradigm, i.e. *casa*, *casas*. The problem lies in that there is another word that looks like the missing cases for 'casa' being of the f1 paradigm, i.e. *caso, casos* ('case') that are in fact forms of another word. One possible solution is to accept that there is no enough certainty to 'measure' it, and to leave the decision to a human, or to start thinking on how we can make an instrument that gathers more information to be able to reach a level of acceptable uncertainty. Maybe a simple 'bag of words' technique will help in deciding whether it is a sole word, or they are two words.

5. Syntactic information

We can start with subcategorization information, which is used by a large number of NLP applications. In EAGLES the notion of subcategorisation is interpreted as "the lexical specification of a predicate's local phrasal context" (Sanfilippo et al. 1996, p.1). This definition has been also used in LMF. Briefly, subcategorization corresponds to a set of possible syntactic structures (the head and its syntactic arguments, with their phrasal realization) associated with an entry (typically a verb, but also a so-called predicative noun, an adjective or an adverb).

The specifications of lexica that include such information describe subcategorization information also as a closed list of possible phrasal combination. We can take this list also as the necessary point of reference to build up our measures. For instance, there is the class of "transitive verbs", which are those that appear with a direct complement, being this complement realized as a pronoun, a noun phrase or a sentence. Merlo and Stevenson (2001), for instance, use several strategies to assess verbal transitivity in texts. These authors use the coocurrence with accusative pronouns as a certain observation to be taken into account for measuring verbal transitivity. In Spanish, this strategy is also to deliver good results as all direct objects can be substituted by an accusative clitic pronoun. For other categories, such as noun and adjectives, there are also syntactic features that can be automatically measured with a certain degree of certainty, as we will show below with the case of copulative adjectives in Spanish.

6. Components of an instrument for measuring standards

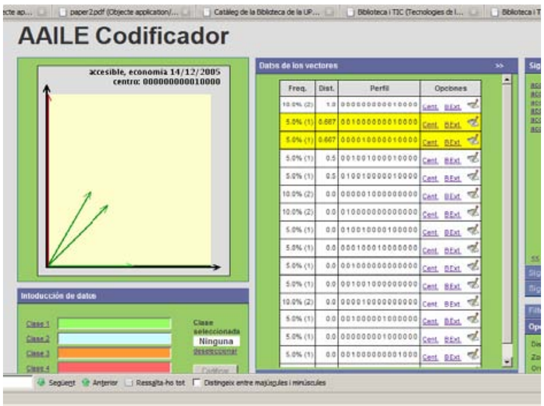
Thanks to web services, to linguistic technologies, and more crucially to existing standards for the annotation of language resources, we can approach our goal in a modular way. Let's examine how the workflow for our instrument is defined and how elements that we already know very well fit in this workflow. Firstly, we will define how to get information from texts. Using some web services, this information is available easily. We can use a web service that interacts with a collection of texts by means of the CQP system (Evert, 2005), which will retrieve data for being measured.

It is easy to imagine that, instead of using a corpus, some agents can be sent to 'explore' domains of knowledge, and to measure the words found in that domain. These data, the measures, are sent to a registry that is later interpreted when encoding lexica. The construction of the lexicon is then like making the map for an area of knowledge.

Keeping this example easy, we will use the widely used CQP web service as was mentioned before whose tokenization code should also be considered a standard. For instance, we could be interested in the lemma "accesible" ('accesible') for a specific domain "economy" in our corpus. We start by retrieving occurrences of this word in texts of this particular domain. The second step is to use a pos tagger to tag morphosyntactically the text retrieved in the first step. POS tags can be handled generically because of the use of standards in that step too. As a third step we reach the moment to measure the information gathered.

In order to represent the information contained in the retrieved word occurrences, we can map linguistic contexts into vectors. We did it already in the project AAILE (Bel et al. 2006) where we used Regular Expressions (RE's) that search for local syntactic information –sequences of tags—in a part of speech tagged corpus. Different RE's check whether a number *n* of particular contexts are found in each occurrence of a word in a corpus (personal pronouns for verb transitivity or co-occurrence with a copulative verb for adjective predicative value). The positive or negative results of each RE checking are stored as values of *n* dimension vectors. In our case, we have defined binary values. The RE's component is embedded into a component that is prepared to read text looking for a particular word, the one the lexicographer is encoding, under certain conditions. In the current implementation, it has access by means of a web service to a pos tagged corpus, but it could be made so that it has access to any repository of texts, and that it extracts the occurrences and send them to an independent pos tagger. As a last step we want to save this information into a database for later use, for replicability and for recalculation if an error occurs. At this moment, we reached our goal. We have an objective collection of data of the lemma "accesible" in the domain "economy".

This data processing can search large resources and take a long time but it is important to notice that this task can be performed completely in background and without human interaction. Fortunately, to keep the results of our process doesn't require large resources and for sure it can be reused for further data treatment, calculation of other properties, etc. We will present below two screen shots of one of our tools that can help to figure out how these measurements could look like. In Figure 1 we see how vectors can be used to measure distances between different occurrences, variance, etc.

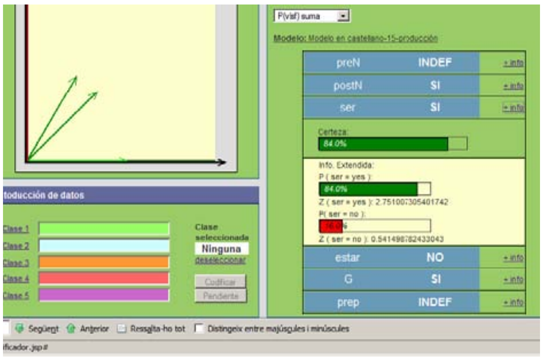


accesible, economía 14/12/2005
centro: 000000000010000

Freq.	Dist.	Perfil	Opciones
10.0% (2)	1.0	000000000010000	Cent. BExt.
5.0% (1)	0.667	001000000010000	Cent. BExt.
5.0% (1)	0.667	000010000010000	Cent. BExt.
5.0% (1)	0.5	001001000010000	Cent. BExt.
5.0% (1)	0.5	010010000010000	Cent. BExt.
10.0% (2)	0.0	000001000000000	Cent. BExt.
10.0% (2)	0.0	010000000000000	Cent. BExt.
5.0% (1)	0.0	010010000100000	Cent. BExt.
5.0% (1)	0.0	000100010000000	Cent. BExt.
5.0% (1)	0.0	001000000000000	Cent. BExt.
5.0% (1)	0.0	001001000000000	Cent. BExt.
10.0% (2)	0.0	000010000000000	Cent. BExt.
5.0% (1)	0.0	001000001000000	Cent. BExt.
5.0% (1)	0.0	000000001000000	Cent. BExt.
5.0% (1)	0.0	001000000001000	Cent. BExt.

Figure 1: Vectors for representing and measuring syntactic properties in AAILE

This information can be also used to make predictions about the values, and deliver also associated uncertainty measures. In figure 2, it shows to what extend the measurement about this adjective being predicative is certain: a 84% after having seen the occurrences shown in Figure 1.



P(v|sf) suma

Modelo: Modelo en castellano-15-producción

preN	INDEF	+ info
postN	SI	+ info
ser	SI	+ info

Certeza: 84.0%

Info. Extendida:
P (ser = yes): 84.0%
Z (ser = yes): 2.751007305401742
P(ser = no): 16.0%
Z (ser = no): 0.541498782433043

estar	NO	+ info
G	SI	+ info
prep	INDEF	+ info

Figure2: Measuring adjective syntactic information

7. Conclusion

We have tried to present here an idea on how to improve and promote the full use of standards. We have based our reflections on the experiences of other scientific and technological domains that we think have some parallelism with the field of Language Resources.

What we have presented here is simply a first exercise. There are many questions open that need to be studied and solved with a dose of imagination. Nevertheless, it seems to us that the field of language resources must reflect on the current situation and to look for a breakthrough that overcomes the problem of lack of trust in available resources, impossibility of integrating different resources or different levels of encoding, and the most crucial issue, the lack of large enough lexical databases.

8. Acknowledgments

This work has been supported by the Spanish Ministerio de Educación y Ciencia: project AAILE, HUM2004-05111 and the Ramón y Cajal, Program.

9. References

Bel, Núria; Espeja, Sergio; Marimon, Montserrat (2006). "New tools for the encoding of lexical data extracted from corpus" in Proceedings of the 5th International Conference on Language Resources and Evaluation. Paris: European Language Resources Association. Pp.. 1362-1367

Calzolari, N., Lenci, A., Zampolli, A,. Bel, N., Villegas, M., & Thurmair, G. (2001). 'The ISLE in the Ocean Transatlantic Standards for Multilingual Lexicons (with an Eye to Machine Translation)'. In Proceedings of the MT Summit VIII. Santiago de Compostela, 2001.

Evert, Stefan(2005): The CQP query language tutorial. Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 Lexical Markup Framework (LMF), Proceedings of the 5th International Conference on Language Resources and Evaluation. Paris: European Language Resources Association.

Hong, S. (2004) Building circuits of trust. History of Science. Science 305.

Ide, N., Romary, L., & de la Clergerie, E. (2003). “International Standard for a Linguistic Annotation Framework”. *Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology,* Edmunton.

ISO 24613 Language resource management - Lexical markup framework. ISO Geneva 2005.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A. (2000). 'SIMPLE: A General Framework for the Development of Multilingual Lexicons'. En International Journal of Lexicography, XIII(4), 249-263.

Latour, B. (1987) Science in Action: How to Follow Scientists and Engineers Through Society. Cambridge, MA: Harvard University Press.

Merlo P and S. Stevenson (2001) "Automatic Verb Classification based on Statistical Distribution of Argument Structure", *Computational Linguistics*, 27:3, pp. 373--408

Normier, B., M. Nossin (1990). "GENELEX Project: Eureka for Linguistic Engineering", Proc. of the International Workshop on Electronic Dictionaries, OISA, Kanagawa, Japan, pp. 63-70.

Quasthoff, U. (1998). "Tools for Automatic Lexicon Maintenance: Acquisition, Error Correction, and the Generation of Missing Values". In Proc. of LREC, pp. 853-856.

Sanfilippo et al. (1996) Subcategorization Standards. Report of the EAGLES Lexicon/Syntax Group. SHARP Laboratories of Europe, Oxford Science Park, Oxford, UK.

.

Crosswalking from Terminology to Terminology:
Leveraging Semantic Information across Communities of Practice

Sue Ellen Wright and Dave Summers
Kent State University Institute for Applied Linguistics
109 Satterfield Hall, Kent, Ohio 44242 USA
E-mail: sellenwright@gmail.com, dgsummers@gmail.com

Abstract

Both subject language terminologies (controlled vocabularies) as represented by the Simple Knowledge Organization Scheme of the W3C (SKOS) and discourse-related terminologies as elaborated by ISO TC 37 (TBX) contain terms, definitions, and semantic relations, which could under ideal circumstances be leveraged across applications and platforms in order to provide semantic anchors for conceptual references when creating knowledge resources in web environments. This paper proposes a cross-mapping between comparable elements in the two environments, taking into account both close similarities and significant differences in the semantic content of important elements on either side of the equation. The final evaluation proposes that the richer concept ordering environment of TC 37 concept systems requires the use of OWL-Lite in order to accommodate all levels of granularity.

1. Terminologies and communities of practice

Since the Terminology and Knowledge Engineering (TKE) conference in 2005, a working group that grew out of the SALT project (SALT, 2001; Budin and Melby, 2000) has envisioned a crosswalk between the Simple Knowledge Organization System (SKOS) (SKOS Reference and SKOS Primer, 2008)[1] and the ISO 12620:1999 / LISA TBX[2] (ISO DIS 32042:2008) standards. In the emerging environment of the Semantic and Pragmatic Web, there is keen interest in options for leveraging existing knowledge ordering schemes from a variety of knowledge representation resources (KRRs). Differences in communities of practice going back deep into the previous century have resulted in incompatible approaches to the representation of terms, definitions, and relationships between concepts, often based on different intentional aims. Although these differences may be known on a superficial level, it is not always obvious where specific discontinuities lie. As a consequence, despite similarities in surface-level vocabulary usage and apparent goals, cross-leveraging of linguistic and semantic information among different types of resources that lay claim to the "terminological" approach is not trivial. Hence the apparent tautology in the title for this paper references the incompatibility of usage between different communities that use the term *terminology*. We will distinguish between *subject-language terminologies* (SLTs; Svenonius, 2000), which reflect practice in the SKOS community, and *discourse-oriented terminologies*, also called *language-purposed vocabulary* (Tudhope et al., 2006, 26), or terminological databases (termbases, TDBs), which are the province of the TBX environment.

ISO Technical Committee (TC) 37, *Terminology and other content and language resources* defines *terminology* in ISO 1087-1:2000 as a "set of designations belonging to a special language", whereby these designations reflect the concepts used in that special language and are represented by *terms*: "verbal designations of general concepts in a specific subject field." Although not specifically stated here, *terminologies* in the sense of TC 37 document these designations for use in discourse, primarily in written texts. The range of approaches commonly covered by such terminologies comprises a continuum of resources reflecting increasing degrees of control and prescriptivity. On the far left side of such a cline, truly *uncontrolled vocabulary* native to a domain is used for both oral and written communication. Although this usage is specialized, it is also a subset of natural language and is embedded in general language. Still within the realm of natural language, but gradually moving further away from general language, there are standardized terminologies created as pre-negotiated, consensus-based resources for vocabulary used in official documents of various kinds, as well as rigorously constrained terminology intended for use in controlled English or other controlled languages (O'Brien, 2003). Here the prescription of grammatical usage, syntax and style moves discourse-oriented terminology further away from general language, but without crossing totally over into the status of artificial language.

The scope statement for SKOS has undergone a number of iterations, whereby the latest definition states, "SKOS – Simple Knowledge Organization System – provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other types of controlled vocabulary" (SKOS Primer, 2008). Earlier references to terminologies and terminological concept systems have wisely been removed from the definition. Nevertheless, when we examine the various standards and literature of the documentation community, *terms* and *terminology* appear in conjunction with controlled vocabularies. The formal definition of *term* in NISO Z39.19 (2005) is deceivingly similar to that in ISO

[1] Given the time frame involved with the submission of this abstract, it does not reflect an in-depth study of the latest SKOS draft specifications listed in the references.
[2] Check *References* for missing acronym full forms.

1087: “One or more words designating a concept.” Here the official designation for a set of these particular terms is *controlled vocabulary*: *A list of terms that have been enumerated explicitly.* As an almost obvious consequence, such collections of terms are also widely called *terminologies,* even in the standard. This standard, as well as most resources in the field, does not distinguish between terminologies that represent the language of relatively uncontrolled discourse in special subject fields on the one hand and the notion of controlled vocabularies, such as thesauri, on the other. SKOS-type resources are primarily used to label and retrieve information or resources and are made available in the form of *terminology services* (Tudhope et al., 2007).

In an effort to distinguish between terminologies and natural language lexicons, Svenonius attempts to sort out this ambiguity by designating controlled vocabularies as *subject languages*, which she contrasts to natural language. “Subject languages are artificial languages, designed for the special purpose of retrieving information. As such they differ in certain essential respects from natural languages… In a natural language the extension, or extensional meaning, of a word is the class of entities denoted by that word, such as the class consisting of all butterflies. In a subject language the extension of a term is the class of all documents about what the term denotes, such as all documents about butterflies.” (Svenonius, 2000, 129-132.)

Svenonius also stresses the need for univocality (one concept, one term) in subject language terminologies, a requirement that echoes the Z39.19 requirement that terms in controlled vocabularies be clarified as *preferred,* with synonyms mapped to preferred terms for search and retrieval purposes. This distinction contrasts practice in discourse-oriented terminology management, where one has the option of assigning a preferred attribute to a term, especially for prescriptive or controlled resources, but where many synonyms are admitted, and even deprecated terms are fully documented. The assumption here is that such non-preferred terms are not just pointers to descriptor terms, but actually occur in texts, depending on register, usage community, and text purpose.

In this context, Tudhope et al. (2006) cite *lexical databases* used for linguistic purposes as resources that are compatible with controlled vocabularies, but it should be noted that they cite *Wordnet,* with its “synset” notations, as their primary example of such a resource, which significantly narrows the understanding of what is meant by *lexical database*.

2. Fundamental distinctions

Despite Svenonius insightful distinction, she does not account for the fact that in addition to subject language “terminologies” (our SLTs) and general language lexical resources, there is also a class of resources commonly called *terminologies* that:

- are concept-based;
- document special languages;
- are only in some cases prescriptive by establishing univocality (one term, one concept);
- frequently document levels of usage and register, thus designating multiple preferred terms subject to multiple pragmatic constraints;
- are designed for use by authors and translators creating text, but not necessarily for information, document, or other types of object retrieval.

Distinctions between TDBs and SLTs are critical for efforts to map TBX data categories to SKOS elements because the *terms* included in the two systems do not necessarily represent the same conceptual content, although valuable referential relations between the two approaches exist.

TDBs run the gamut from multilingual, strongly text-oriented bilingual glossaries created on the fly for project-related purposes in the localization industry to extensive enterprise-wide or government-sponsored term banks. Some feature little explicit systematic content in the form of potentially ontological information, while others elaborate concept systems and semantic networks. In attempting to categorize knowledge representation resources with respect to relative systematicity, Wright (2007, 159) identifies TDBs as a continuum including both systematic and non-systematic approaches. Insofar as some terminologies elaborate concept fields and networks, they are akin in some degree to other concept schemes used for knowledge representation. As shown in Table 1, systematic TDBs explicate a variety of relations, most prominently generic hierarchies, meronymy and metonymy, sequentiality, and a variety of other freer types. In addition to the fairly constrained set cited in ISO 12620:1999, Nuopponen documents over 40 terminal nodes in her classification of ontological concept relations used in TDBs (Nuopponen, 2005, 213).

Although some smaller terminological collections may be ordered into a single cohesive concept system, many large dynamic systems are more likely to identify shallow concept fields, or simply designate parent concepts or simple relations. Indeed, it is a feature of so-called “ad hoc terminology management,” which is strongly text or corpus-based, that elaborators of TDBs frequently have difficulty recognizing precise concept relations until a certain critical mass of concept entries has been collected for a given subject field. Current thinking postulates the possibility of moving the generation of concept systems outside the core structure of terminology management systems, using persistent identifiers to reference individual concept entries within the system in order to provide authoritative definitions and other information for conceptual references in relational or knowledge resources (ISO WD 24618, 2008).

3. What might be leveraged

Tudhope et al. document current efforts to create terminology web services that are in some degree interactive. Given the challenges involved in solving accessibility issues for different controlled vocabulary networks, any effort to link TDBs and SLTs needs to define potential mappable contact points between the two approaches. As illustrated in Table 1, there are enough differences between SKOS and TDB elements to require a conscious, potentially complex mapping strategy. Although it is not the purpose of this paper to outline an implementation of RDF representation for TDBs, some sort of RDF representation must be the first step towards realizing the general availability of such content within the terminology services environment. Before moving to this step, however, those data elements (data categories) that are of mutual interest must be identified, and methodologies for reliable, persistent access must be resolved. Table 1 represents an initial approach to this sort of mapping. Rather than present elements in the familiar alphabetical order found in the two standards, we have chosen to group similar elements together in order to facilitate discussion. We do not, however, propose any doctrinaire attachment to this presentation as an ordering scheme per se.

The core data categories contained in most terminological resources, and the ones that are most critical for any kind of semantic cross-usability, are *terms* and *definitions*. The issue arises as to whether terms map to terms (labels) and definitions map to definitions, which will be examined below. It should also be noted in this context that metadata registries (MDRs) also contain rigorous definitions prepared by subject-field experts which are also made available in web environments as anchors for the disambiguation of references in semantic resources (see Windhouwer et al., 2008), but this consideration goes beyond the scope of the current paper.

For its part, TBX "is designed to support the analysis, representation, dissemination, and exchange of information from terminological databases (termbases)" (ISO DIS 30045:2008). The data categories listed in the TBX standard are taken in part from ISO 12620:1999 and have been modified to some extent for purposes of simplification. In future they will reside as a defined data category specification (DCS) in the TC 37 **ISOcat** metadata registry. It is not our intention to replicate functionalities across communities of practice, but rather to leverage specific items of information or to reference authoritative documentation residing in diverse resources. Data categories can be grouped as follows:

- *Specific references to controlled vocabulary*

Whereas ISO 12620:1999 contained a list of items related to controlled vocabularies (*thesaurus name, thesaurus descriptor, top term, broader term, narrower term*), the current TBX specification includes only *thesaurus descriptor.* The elimination of other items reflects in part an effort to avoid replicating the functionality of a controlled vocabulary, something that is seldom done and is actually seldom desirable. A simple cross-reference to a label that matches the term or concept in question is a more efficient approach in that it leaves positional specification within the controlled vocabulary environment. In our view, however, the elimination of *thesaurus name* (for which we would suggest the substitution of the SKOS element *inScheme*) is unfortunate. The reason for this proposal is that concepts and terms documented in TDBs can easily be associated with multiple concept schemes expressed in multiple controlled vocabularies. It would be highly useful, even for diverse concept systems within the local environment of a TDB, to be able to indicate the specific scheme referenced by a relation, although it can be assumed that for external references, the URI used to link to the targeted item will provide unequivocal identification for the resource involved.

Within the framework of the discussion surrounding TDBs and SKOS, one suggestion that has been made is to incorporate SKOS elements into TBX. Another, which has been reflected in earlier stages of our own work, is to identify missing elements and selectively add them to TBX. At this juncture, we would propose instead that the entire SKOS element set be added to the **ISOcat** registry, where it would be available for all TC 37 language resources. There is a precedent for this kind of insertion in that other standards, such as Dublin Core (2008) and the Language Codes (ISO 639 family of standards) have also been incorporated into the DCR.

It is also not highly likely that the fairly direct link between the thesaurus-related elements cited ISO 12620:1999 and the SKOS environment will be particularly fruitful as an avenue for mapping termbase elements to SKOS elements because very few TDBs utilize this option, with perhaps the possible exception of some national term banks that may map to authoritative national thesauri or classification schemes. Thus we have dispensed with the old *documentation language* elements from 12620:1999 and focus on those data categories that are more widely used and that consequently may be present for use in any future mapping exercise.

- *Labels and terms*

The fundamental difference between SLTs and TDBs is nowhere more critical than with respect to the relationship between SKOS labels and termbase terms. Although overtly prescriptive termbase environments specify the term status attribute as a required value, many do not, which can mean, for instance, that concept entries that only contain one term in a language will not be differentiated as *preferred.* Furthermore, TBX offers a variety of ways to distinguish synonyms in a concept entry, for instance by indicating sortable subset information (i.e., *customer, project, product, application, environment, businessUnit, security* (A.10.3)), any of which could be used instead of a status attribute for purposes of indicating the preferential status of a term in a particular context. Furthermore, and perhaps more importantly, the status attribute or any of these sorting attributes, can be used in TBX terminologies to indicate appropriate usage in discourse, whereas the *preferred, alternate,* and *hidden* elements in SKOS reflect knowledge retrieval strategies rather than discourse usage. Well-structured label sets in the SKOS environment should ideally reflect synonym sets such as those presented in concept-oriented termbase entries, but this has not always been the case in the past, where subordinate concepts or closely related concepts are grouped together with respect to a label in order to increase the concision of controlled vocabularies.

Another factor that plays a role here is that TBX and **ISOcat** attributes provide a much finer granularity for distinguishing term types than is afforded by SKOS labels, reflecting the degree of specificity required in text-oriented situations. Hence, terms in TBX are further specified according to twenty some term types (e.g., abbreviation, full form, and symbol, to name a few). Another factor, of course, is that the assignment of the *preferred* or *admitted* attributes in either case is a function of usage or system design, so data integrity in exchanging information between the two system types is compromised by the need to distinguish two properties (*termType* and *status*) on the TBX side of the equation for each of the individual elements on the SKOS side. One missing element on the TBX side, interestingly, is the inclusion of misspellings in the SKOS *hidden label* element. Although this could conceivably be documented in TBX as a *deprecated term* that is also a *variant*, perhaps the addition of *misspelling* as an optional value of *termType* might be interesting even outside the interoperability environment. Given the considerations cited here, it seems most expedient to ignore the option of identifying the status of a label from the TBX side and only try to map to a SKOS label (of whatever class) that matches the TBX term at the string + subject field level.

- *Definitions*

Although SKOS classifies definitions as one of several notes, definitions play a central role in TDBs, so we have pulled them out for special discussion. We have rather arbitrarily left *scopeNote* for the time being together with *notes* in Table 1, but it is impossible to discuss rigorous terminological definitions without attempting to distinguish them from (or indeed, equate them to) scope notes. Unfortunately, the various recommendations and examples provided in the literature do not make this a simple task. *Definition* in SKOS is defined as: *A statement or formal explanation of the meaning of a concept*, and *ScopeNote* is: *A note that helps to clarify the meaning of a concept*. (Miles and Brickley, 2005). This distinction supports our tentative mapping of *skos:scopeNote* to TBX *example*. The examples provided in the context of SKOS Core, however, project a confusing image:

```
<skos:Concept
rdf:about="http://my.example.org/GCL/702#s
copeNote">
  <skos:prefLabel xml:lang="en">
Competitiveness</skos:prefLabel>
  <skos:scopeNote xml:lang="en">The ability
of businesses to compete in local, national
or international markets.</skos:scopeNote>
</skos:Concept>

<skos:Concept
rdf:about="http://www.example.com/concepts
#banana">
  <skos:prefLabel xml:lang="en">
banana</skos:prefLabel>
  <skos:definition xml:lang="en">A long
curved fruit with a yellow skin and soft,
sweet white flesh inside.</skos:definition>
</skos:Concept>
```

In terms of definition theory, for instance as specified in ISO 704 (2000), both the definition and the scope note shown here are more or less rigorous definitions that define the discourse-oriented concepts represented by *competitiveness* and *banana*, respectively. Further efforts to clarify this distinction lead to O'Reilly's xml.com webpage, where the hope for an authoritative explanation remains, alas, unfulfilled:

To clarify the difference between skos:definition and skos:scopeNote, a definition should be an attempt to completely explain the meaning of a concept, whereas a scope note may consist of partial information about what is or is not included within the meaning (or scope) of a concept.

> **<skos:definition>A feature type category for places such as the Erie Canal </skos:definition>**
>
> **<skos:scopeNote>Manmade waterway used by watercraft or for drainage, irrigation, mining, or water power</skos:scopeNote>** (Mikhalenko 2005)

Interestingly, from the standpoint of ISO 704, this second *skos:scopeNote* is actually the more appropriate definition with respect to discourse-oriented TDB approach, while the *skos:definition* provides an SLT-style definition of the feature as a label in a controlled vocabulary. This distinction looks very useful to us, but it would lead us to map the TBX *definition* to a SKOS *scopeNote*. Needless to say, this is an issue that needs to be resolved at some point, particularly if it reflects differences in practice within the SKOS community.

- *Note*

As indicated in Table 1, SKOS specifies a variety of notes, most of which can be expressed in the TBX framework by positioning the *note* element inside various containers in the TBX metamodel. (We do not have the space to illustrate this feature in the context of this paper.) *Example* in TBX matches well to the *skos:example*, but is considered to be a core descriptive element used to delimit the concept treated by a terminological entry, so it is not classified as *note* material in **ISOcat**. This distinction is not problematic, however, with respect to mapping from one system to the other.

- *ConceptScheme / Concept System*

All TBX *concept systems* (defined as: *The structured set of concepts established according to the relations between them, each concept being determined by its position in the set.* (ISO 12620:1999)) can be classified as *skos:concept Schemes* (*A set of concepts, optionally including statements about semantic relationships between those concepts.* (SKOS Core)), but the *optional* component of the SKOS definition means that not all concept schemes are concept systems in the sense of ISO 12620 or ISO 1087-1. Since each concept in a termbase is treated in its own entry, the non-mnemonic entry identifier, which could ideally be configured as a persistent identifier, serves as the concept identifier and the means whereby the entry is accessed either internally within the native termbase structure or from external resources via some sort of URI-type link. As noted above, the *inScheme* element provided by SKOS has, in our view, tremendous potential for use in TBX quite apart from any crosswalk, because the concept entries in a TDB generally exist independently of any given concept system and can participate in multiple schemes even within their native environments, which renders position notations within an unnamed scheme or simple indications of parent-child relationships potentially ambiguous.

- *Subject / SubjectField*

SKOS subject-related elements are designed to account for point of view, pointing both from a concept entry to a subject marker with which it is associated or enabling the assembly of a set of entries associated with a given subject. Here, however, the distinction between SLTs and TDBs resurfaces as a significant issue: in SKOS, where the label represents a resource, the *subject* is the subject or one of the subjects treated by that resource. In TBX, the *subject Field* is one or more of the specialized disciplines with which the concept treated in an entry is associated. This distinction is critical to the specification of the concept via the definition and other descriptive elements, because these items are only considered valid within the framework of the subject or subjects declared for a given concept entry. Although some TDBs allow for multiple subject field assignments in cases where term-concept pairs are shared by more than one declared subject field, it is not common practice in TDB management to differentiate primary or preferred subject fields. Despite these differences, the use of *subject/subjectField* attributes in determining search or mapping criteria is likely to be useful for purposes of semantic retrieval. Nevertheless, this is a problematic area even when matching controlled vocabulary to controlled vocabulary or termbase to termbase because of the frequent ambiguities and discontinuities involved in the specification of subject field categories across application boundaries. The most valuable environment for such matching would require a shared subject classification model.

- *SemanticRelation / Concept relations*

On the one hand, concept relations would appear to be a critical focus for this discussion, which is essentially centered around leveraging semantic information across methodological boundaries. But by the same token, the subtle differences between the systems pose the real risk that the equivalences proposed in Table 1 may be highly deceiving. At the most apparent level, the equation of *has TopTerm* with *broaderConceptGeneric* is potentially dangerous because a broader concept is not necessarily a top term in a scheme or system: *A concept two or more levels of abstraction higher than subject concept in a hierarchical concept system* (ISO 12620:1999). This designation is frequently used when formulating rigorous definitions if the immediate superordinate concept is deemed to be too specific or unfamiliar for use as a transparent genus element. Broader and narrower concepts map fairly safely to superordinate and subordinate generic concepts (*isA* relations), as long as one bears in mind the distinctions already cited between the conceptual extensions of labels and terms.

- *Collections, collectable properties, and members*

Currently there is a serious discussion in the SKOS community attempting to differentiate collections and concept schemes, a concern which to a certain extent alarms the authors, because we had thought we understood the distinction and felt it fit comfortably into TBX structures. In our way of viewing things, a *skos:collection* is any set of coordinate concepts in a generic system assembled to represent a set of concepts that comprise siblings dependent on a superordinate node. Typically termbases do not necessarily explicate such sets, but they are nevertheless often generatable at the user interface level based on the specification of the parent node as a search criterion for the data category *superordinateConcept Generic*, as evidenced by our mapping in Table 1. Any such subordinate concept is automatically a member of such a set. ISO 12620/TBX has no facility for creating ordered collections, i.e., for imposing ordering rules on such a dynamically assembled set, except in those instances where a *conceptPosition* (A.7.2) number is specified in the data model, in which case this number can be used to impose order in the set. This kind of ordering, however, can be very difficult to achieve in large, dynamically changing TDBs because it would probably entail human intervention, together with an overview of the various siblings involved..

- *Related concepts*

Although simple *skos:related* would appear to map comfortably to *relatedConcept*, from this point on in Table 1 it becomes apparent that TBX and termbase solutions on the whole provide a richer set of relation types than is necessary for *simple* knowledge organization within the framework of controlled vocabularies. Currently we visualize using OWL-Lite in order to facilitate the expression of more diverse relation types. Indeed, we propose *relationType* as an addition to the DCR, although one can also make an argument for moving more complex relations outside the core of a termbase, in which case relation resources using OWL technology could be built external to a termbase, using persistent identifiers to anchor RDF relation specifications via the concept entry IDs embedded in terminological entries. This scenario has also been proposed for linking relation resources to metadata registries (Wittenberg, 2007).

4. Who would want to do this?

The idea of information interchange has always inspired a certain population of nay-sayers. In the 1980s, it was not unusual to encounter resistance to the elaboration and deployment of data interchange standards. Today we see a wide variety of such standards in an equally broad number of application areas: DITA, XLIFF, LMF, TBX, TMX, SBVR. The advantage of these kinds of prenegotiated markup formats is that they facilitate automatic processing and the ability to leverage information from diverse resources residing in distributed networks and on different platforms. Nevertheless, the very variety of formats presents issues with respect to interoperability and exchangeability, particularly in dynamic environments. XML provides the ability to reference material expressed using one XML application inside resources that are formatted in some other XML application. This capability suggests the advisability of utilizing existing formats for specified functionalities rather than building those functionalities into other formats.

Although this powerful option enables leveraging to a certain extent, it does not necessarily ensure true interoperability, given the fact that even very closely related environments can nonetheless employ dramatically different approaches to conceptual ordering systems. Within communities of practice, mapping, crosswalk, and even mashup strategies are designed to attain more or less lossless interchange – either on a dynamic or a snapshot basis. This paper is proposing an approach that takes the principle of interchange to a higher meta-level by suggesting crosswalks between different interchange and interoperability formats that were not necessarily designed with the same metamodels and the same functional applications in mind. The goal here would be to be able to retrieve and utilize some, but not necessarily all, of the information stored in diverse resources. Potential implementations might include:

NLP researchers developing yet another new toy, which is interesting on an intellectual level, but which has little practical application in pragmatic organization environments – This “application” reflects a criticism that “no one” in industry is doing this now, so it is not really essential to think about doing it in the future.

Publically available resources, such as metadata registries and standardized terminologies (e.g., a planned ISO concepts directory) contain parsably accessible terms, labels, and rigorous definitions that can be harvested for use in other similar resources and as semantic anchors for values used in ontological resources of various kinds. This is a significant factor because one of the issues involved in ontology design is that ontology developers frequently are not themselves subject area specialists, coupled with the fact that the latter are not necessarily knowledge engineers. Especially as the creators of standardized resources expand their skills in developing nonambiguous knowledge resources, the presence of authoritative, persistent, accessible semantic information can contribute to the Trust Layer in the framework of the Semantic Web.

Private enterprises, particularly those with large distributed infrastructures, are relying on formal ISO-level standards as well as markup languages, exchange formats, and other consensus-based approaches that are being developed by business-to-business consortia such as OMG, OASIS, LISA, etc. These organizations might use this approach to access and exploit the information that resides within in-house corpora and knowledge bases for internal use. Despite concerns for protecting proprietary information, some companies (Microsoft, IBM) maintain extensive public information resources and have or are launching ambitious “community computing” services that are base on or will be building linguistic resources.

Governmental and other public service entities frequently possess at least theoretically extensive resources that contain terminological and semantic content, but that have been developed in widely diverse environments over time using approaches and formats that may or may not be mutually compatible. One approach to this is to harvest so-called “snapshots” of resources, convert such snapshots into a common annotated format, and make them available as combined content accessed through a common interface. The concern here is that many of these resources may continue to evolve in their native environments, with the result that the combined resource runs the risk of being obsoleted if there is no on-going, potentially costly, maintenance strategy. Procedures for establishing interoperability in real time for dynamically evolving resources (coupled with the implementation of viable persistent identifiers) might make it possible for real-time interaction of aggregated knowledge representation resources, even if they have been originally configured for use by different communities of practice.

5. Theoretical approaches and task definitions

Obviously, this paper represents only a statement of the problem at hand and an initial sketch of potential mapping paths between the systems involved. A number of individual items require further clarification in dialogue between the communities of practice, and some are subject to on-going decision making on both sides of the question. One future task involves the further refinement of the mapping tables, but it is nonetheless important to note that the premise upon which the mapping exercise rests is the assumption that semantic mapping can or should take place between resources developed in environments that are playing by significantly different rules. Any decisions along these lines, as noted above, must be accompanied by the generation of RDF notation for TDB data. The question arises whether existing methodologies, at least viewed from the perspective of discourse-oriented terminology management, make sense in an environment where tools exist for creating terminological concept systems external to, but linked to, concept-oriented TDBs.

6. References & Abbreviations

ANSI/NISO – American National Standards Institute/ National Information Standards Organization

___. Z39.19-2005. Guidelines for the construction, format, and management of monolingual controlled vocabularies.
http://www.mt-archive.info/CLT-2003-Obrien.pdf.

Budin, G. and Melby, A. (2000) Accessibility of Multilingual Terminological Resources – Current Problems and Prospects for the Future. http://www.ttt.org/salt/.

DCR – Data Category Registry.

DCS – Data Category Selection.

DIS – Draft International Standard.

DITA – Darwin Information Typing Architecture.

Dublin Core Metadata Element Set, Version 1.1 2008 .http://dublincore.org/documents/dces/.

ISO – International Organization for Standardization

ISO standards: Geneva, International Organization for Standardization:

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

ISO 639-2:1998. Code for the representation of languages – Part 2: Alpha-3 Code.

ISO 639-3:2008. Codes for the representation of languages – Part 3: Alpha-3 Code for comprehensive coverage of languages.
http://www.sil.org/iso639-3/default.asp

ISO 704:2000. Terminology work – Principles and methods.

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

ISO 12620:1999. Computer applications in Terminology — Data categories

ISO 12620 DIS 2007. Terminology and other language and content resources – Data Categories – Specification of data categories and management of a Data Category Registry for language resources

ISO 16642:2003. Computer applications in terminology – TMF (Terminological Markup Framework)

ISO DIS 32042:2008. TermBase eXchange (TBX) Format Specification.

ISO WD 24618. Citation of Electronic Resources.

ISO/IEC 11179 :2007. Information Technology – Metadata registries (MDR) : ISO/IEC JTC1 SC32 WG2 Development/Maintenance, http://metadata-standards.org/11179/.

ISOcat. (2008). Data Category Registry: Defining widely accepted linguistic concepts. http://www.isocat.org/.

KRR – Knowledge Representation Resource

LISA – Localisation Industry Strandards Association.

LMF – Lexical Markup Framework.

Microsoft. (2008). Microsoft Terminology Live Community. Windows Live (German). (Periodic accessibility) https://members.microsoft.com/wincg/de-de/mtcf_home.aspx?s=1&langid=1210.

MDR – Metadata Registry.

Mikhalenko, P. 2005 Introducing SKOS. O'Reilly xml.com. http://www.xml.com/pub/a/2005/06/22/skos.html

Nuopponen, A. (2005). Concept Relations v2. An update of a concept relation classification. In *Terminology and Content Development. Proceedings of the 7th International Conference on Terminology and Knowledge Engineering*, B. Nistrop-Madsen and Thomsen, H.E. (eds.), Copenhagen, Litera, pp. 128-138.

OASIS – Organization for the Advancement of Structured Information Standards.

O'Brien, S. (2003). Controlling Controlled English: An Analysis of Several Controlled English Rule Sets. http://www.eamt.org/archive/dublin/OBRIEN.PDF.

OWL Web Ontology Language Oveview. http://www.w3.org/TR/owl-features/.

OMG – Object Management Group.

RDF – Resource Description Framework.

SALT: (2001). Standards-based Access service to multilingual Lexicons and Terminologies http://www.ttt.org/salt/

SBVR – Semantics of Business Vocabulary and Rules.

SKOS Simple Knowledge Organization System Primer: W3C Working Draft 21 February 2008. Antoine Isaac, A. and Summers, E., Eds. http://www.w3.org/TR/2008/WD-skos-primer-20080221/

SKOS Simple Knowledge Organization System Reference: W3C Working Draft 25 January 2008. Miles, A. and Bechhofer, S., Eds. http://www.w3.org/TR/skos-reference/

SKOS Core Vocabulary Specification 2005. Miles, A. and Brickly, D., Eds. http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102/.

Svenonius, E. (2000). The Intellectual Foundation of Information Organization: Digital Libraries and Electronic Publishing. Cambridge, Mass. MIT Press.

TBX – Termbase eXchange, *see* ISO 32042.

TDB – Terminology Database

SLT – Subject Language Terminology

TKE – Terminology and Knowledge Engineering

Tudhope, D., Koch, T., Heery, R. (2006). Terminology Services and Technology: JISC state of the art review. http://www.ukoln.ac.uk/terminology/TSreview-jisc-final-Sept.html

URI – Uniform Resource Identifier.

W3C – Worldwide Web Consortium.

WD – Working Draft.

Windhouwer, M.; Kemp-Snijders, M.; Wittenberg, P., and Wright, S.E. (2008). ISOcat: Coralling Data Categories in the Wild. Poster article for LREC 2008.

Wittenberg, 2007. ISO position paper on persistent identifiers.

WordNet: A lexical database for the English language. http://wordnet.princeton.edu/. Accessed 2008-03-02.

Wright. S.E. (2007). Coping with Indeterminacy: Knowledge Organization Systems in Digital Environments. In. Indeterminacy in Terminology and LSP: Studies in honour of Heribert Picht. Amsterdam and Philadelphia: John Benjamins Publishing Company, pp. 137-179.

XLIFF – XML Localization Interchange File Format.

Table 1 : SKOS-TBX Cross-mapping

SKOS classes and properties	12620 data category → SKOS	12620 data category ≠ Not in SKOS
Labels and Terms		
label	N/A	term (A.1)[1]
prefLabel	term (A.1), status = preferred term	
altLabel	variant, abbreviation, full form, etc.	term, status = admitted term (A.2.9.1)
hiddenLabel	term termType = deprecated term (A.2.9.1)	term, status = deprecated term (not recommended) (A.2.9.1)
	Misspelling	(not included in TBX)
symbol	term termType = symbol (A.2.1)	
altSymbol	Ditto above, status = admitted term (A.2.9.1)	
prefSymbol	Ditto above, status = preferred term (A.2.9.1)	
Definitions		
definition [2]	definition (A.5.1)	(See discussion above)
Notes		
note	note (A.8)	
changeNote	note inside transacGrp, type="modification"	
editorialNote	note (A.8) within terminology management transactions (A.10.1)	
example	example (A.5.4)	
historyNote	history note	(not included in TBX)
scopeNote	explanation (A.5.2)	(See discussion above)
Concept System/Scheme		
ConceptScheme	concept system (A.7.1)	
Concept	entry identifier (A.10.15)	
inScheme	[pointer to A.7.1]	(suggested for TBX)
Subject references		subject field (A.4.5)
subject		(not broken down in TBX; see discussion above)
primarySubject		
subjectIndicator		
isSubjectOf		
isPrimarySubjectOf		
Relations		
semanticRelation	N/A	Concept relations
hasTopConcept	Potentially: broader concept generic (A.7.2.1)	
broader	superordinate concept generic (A.7.2.2)	
narrower	subordinate concept generic (A.7.2.3)	
CollectableProperty	N/A	[Embedded as genus element in rigorous definitions]
Collection		[Any superordinate concept that could become the subject of a collection]

[1] Non-mnemonic TBX data category ID number. There is no way in SKOS to deal with multiple preferred terms classified by other categories, although this could be finessed by treating these categories as scheme identifiers, such as inScheme=clientSet1, businessUnit1, etc.

[2] In SKOS, definitions appear under note, whereas they are fundamental elements in termbases. See also *scopeNote.*

SKOS classes and properties	**12620 data category → SKOS**	**12620 data category ≠ Not in SKOS**
member	coordinate concept (A.7.2.4)	[Any subordinate concept that could become a member of a collection]
OrderedCollection	Ordered thesaurus (A.9.6) ??	
memberList	ordered coordinate concept (A.7.2.4.1)	
related	related concept (A.7.2.5)	
		broader concept partitive
		superordinate concept partitive
		subordinate concept partitive
		coordinate concept partitive
		temporally related concept
		spatially related concept
		associated concept
		(Nuopponen categories)
		antonym (is disjoint with) (10.18.6.1)

Data Category Registry: Morpho-syntactic and Syntactic Profiles

Gil Francopoulo, Thierry Declerck, Virach Sornlertlamvanich,

Eric de la Clergerie, Monica Monachini

affiliation of first author: Tagmatica, 126 rue de Picpus, 75012 Paris, France

gil.francopoulo@wanadoo.fr, declerck@dfki.de, virach@tcllab.org,
Eric.Clergerie@inria.fr, monica.monachini@ilc.cnr.it

Abstract

After a brief presentation of the data model, we describe a work in progress to define an initial set of morpho-syntactic and syntactic data categories dedicated to NLP applications. The aim is to improve interoperability among language resources and to optimize the process leading to their integration in applications. The main point is to be sure that when a language resource makes use of a value, the other language resources and programs have the same interpretation for this given value. From a practical point of view, these values are collected from existing lists, discussed, extended, and then recorded within a freely accessible data base: the ISO Data Category Registry.

1. Introduction

Data associated with language resources are identified and stored in a wide variety of environments like terminological data collections and NLP resources. With this respect, we believe that the production of a family of consensual ISO specifications and data can be a useful aid for the NLP actors.

In this paper, after a brief presentation of the data model, we describe a work in progress within ISO-TC37 whose aim is to gather and record data categories (Ide et al, 2004; Wright, 2004).

2. Context

The TC37 standards are currently elaborated as high level specifications and deal with word segmentation (ISO 24614), annotations (ISO 24611, 24612 and 24615), feature structures (ISO 24610), and lexicons (ISO 24613). These standards rely on low level specifications dedicated to constants, namely data categories (revision of ISO 12620), language codes (ISO 639), scripts codes (ISO 15924), country codes (ISO 3166) and Unicode (ISO 10646).

This bi-level approach will form a coherent family of standards with the following common and simple rules:

1) The high level specifications provide structural elements that are **decorated by the standardized constants**;
2) The low level specifications provide these standardized constants.

This decoupling is offered in order to provide a fine flexibility with regard to language and practice diversity. To be more concrete, for instance, in a high level structure such as a lexicon, different elements like a Lexical Entry and a Sense will be defined and linked together in order to allow the definition of different senses for a word, as follows:

```
<LexicalEntry>
  <feat att="partOfSpeech" val="noun"/>
    <Lemma>
      <feat att="writtenForm" val="bank"/>
    </Lemma>
  <Sense id="bank1">
    <Definition>
      <feat att="text" val="Business that keeps and lends money"/>
    </Definition>
  </Sense>
  <Sense id="bank2">
    <Definition>
      <feat att="text" val="Land along the side of a river"/>
    </Definition>
  </Sense>
</LexicalEntry>
```

In this example, LexicalEntry, Lemma, Sense, and Definition belong to high level specifications, more precisely: LMF. In contrast, partOfSpeech, noun, writtenForm, and text belong to low level specifications, more precisely: the Data Category Registry.

The usage of each of these high level elements is specified, together with their cardinality. The precise combination of high level elements and low level ones is not specified: this is left to the user. In other terms, the user selects the structural elements he needs, and provided that a suitable set of data categories is available, the user is able to decorate the structural elements for a given language.

3. Variations

For the high level specifications, a consensus must be found among what is to be considered as "the best practices" of our field. Implicitly, a mixed strategy based on "coherent union" of structures and a meta-model approach is often taken, depending on the agreement among the community.

The main criteria are:

- the various theoretical approaches;
- the languages covered;
- the type of resources (syntax, semantics …)

These three criteria apply on the data category side as well.

4. General objectives

The main objective of TC37 is interoperability and our work is done in the context of the revision of ISO-12620. The most frequently encountered problem is "how to merge data?" whereby the hardest sub-challenge is "how to compare data?".

To address these issues, first, the use of a uniform policy should contribute to system coherence and functionality. And secondly, each data category (DC) must be well defined in order to allow elementary operations like: "is DC-A the same notion as DC-B ?" "is DC-C more general (or more specific) than DC-D ?", or "is DC-E related somehow to DC-F ?".

5. Specific objectives

With this respect, we have two distinct objectives:

1) Test the current specification of the revision of ISO-12620 as a proof of concept ;

2) Concretely record an initial set of data for morpho-syntax and syntax.

The goal is not to create a rich network of links between data categories.

6. History of ISO-12620

The ISO standard 12620 was published in 1999. The document specifies the content of data categories and presents a long list of values, whose primary aim was be used in terminological data collections.

The revision of ISO-12620 is somehow different. The work started in 2003. The document is currently in Final Draft for International Standard (FDIS) stage[1], and the schedule is to reach International Standard (IS) publication in 2009. The development is twofold. The revised version specifies how the data categories will be described and managed, but in contrast to the initial version, the values will not be presented in the ISO document. The values will be managed within a database endorsed by ISO that is called the Data Category Registry (DCR).

Another point to mention is the type of high level structure that is addressed by the new set of data categories. The old version targeted only terminological data collections but the new version target is much broader. The coverage is all TC37 activities, which means that NLP applications are concerned, hence largely increasing the number of values. For instance, the old ISO-12620 had only three values for part of speech, namely: **noun**, **adjective** and **verb**, but now because of NLP data structures, values like **preposition** and **punctuation** are needed. So, instead of only three values, the list contains now one hundred values.

7. Current registry

As cited earlier, the 12620 revision work started in 2003, and a lot of energy has been spent along the years in various meetings and document writings, in order to find an operational consensus. The two tasks (DC specification and DC recording) were conducted in parallel with frequent interactions.

This model has been implemented in a system called "Syntax[2]" which is currently running and is located at http://syntax.inist.fr where about a dozen people have entered values, mainly in the domain of terminology, morpho-syntax, and syntax. The list of the current values is presented in Annex-B, with an indentation for the broader link information.

8. Data model

The current model allows a lot of options but we limit ourselves to a subset of features, as presented in the UML class diagram in Annex-A.

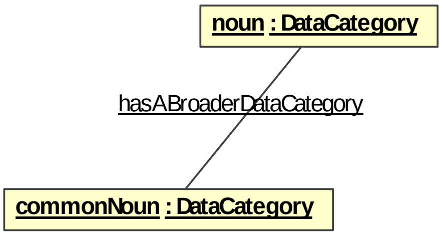
The registry is divided into profiles. A profile is a set of data categories. Each profile is associated with a team of experts with a convenor, who collectively represent a community of practice in the area of language resources. There are currently about ten profiles and as many or more sub-activities, such as terminology, metadata etc, covering all activities of ISO-TC-37. The current paper focuses on two profiles dedicated to NLP, namely the morpho-syntactic and syntactic profiles.

Many times, a data category belongs to only one profile, but a small number of them belongs to several profiles (e.g. part of speech).

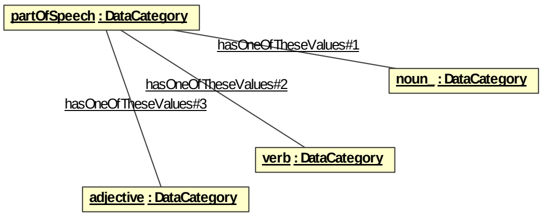
We differentiate between the notion of broader relation and the notion of value domain. The broader link allows a hierarchy of constants that forms an ontology. Example: a **common noun** is a more specialized value than **noun**.

[1] For a reader who is interested in reading the FDIS document, it may be accessed through the National Body channel: ASCII for US, DIN for Germany etc.

[2] The name is not very well chosen and does not mean that the system deals only with syntactic descriptions.



The notion of value domain is different. A value domain allows a set of valid values to be identified. In other terms, a value domain that is attached to a data category X provides a set of potential values for X and these values are themselves data categories. Example: **noun** is a value for **partOfSpeech**.



9. Data: methodology

We proceeded in three phases:

Phase-1: collating of candidates data categories
Phase-2: grouping, structuring, and redaction of a first draft of the definitions
Phase-3: revision

For the morpho-syntactic profile, a long initial list of data categories has been collected from:

- Current ISO-12620:1999
- Eagles and Multext-East
- Some values for Semitic languages coming from Sfax University

For the syntactic profile, an initial list was collected based on:

- Eagles
- Tiger (German project)
- Technolangue/Easy (French project)

Let us add that some values needed from TC37 standards like MAF (ISO-24611), SynAF (ISO-24615) (Declerck et al, 2006) and LMF (ISO-24613) (Francopoulo et al, 2006) have been added to the two profiles.

Each data category has an identifier that is English based. The name does not contain any spaces, and if more than one word is needed, it is expressed in so-called *camel case* (e.g. **commonNoun**) as specified in the revision of ISO-12620.

Currently each DC has a definition in English and French. Let us note that a lot of time has been devoted to write rigorous definitions, taking into account the various stable sources in our field. A definition may be complemented by a note.

A DC may be linked through a broader link to another DC. A DC may have a value domain.

Each DC has, at least, a name in English and one in French, which may be used directly for display without any transformation (e.g. **common noun**).

Currently, the ontology of values (through the broader link) is rather flat and does not exceed three levels. There are no constraints between DCs.

There is currently no indication concerning the use of a given DC for a specific language, but the new version will include a linguistic section that will enable some further constraints on value domains that may reflect specific usage in different object languages.

Thus, to reply to the question: "Is DC-A, the same notion as DC-B?", the user needs to compare identifier of DC-A to identifier of DC-B. If an explanation is needed to understand why two DCs are different, each DC has a precise definition for this purpose.

10. Data: organization

The number of values is rather huge, so in order to facilitate management, a series of directories[3] has been created within the two following profiles.

[3] A directory is equivalent to a sub-profile.

Morpho-syntactic profile:

Basics	These are general purpose linguistic constants, like: **comment**, **derivation**, **elision**, **foreignText**, and **label.**	61	items
Cases	Examples of values: **ablativeCase** or **dativeCase**.	33	
FormRelated	These are constants for the specifications of forms like: **spokenForm**, **writtenForm**, **abbreviation**, **expansionVariation**, **transliteration**, **romanization**, **transcription**, **script.**	36	
Morphological Features excluding cases	Attributes include for instance **grammaticalGender**, **mood** and **tense**. Values include, for instance, **feminine, indicative**, **present**.	82	
Operations	Constants include for instance, **addAffix**, **addLemma**.	29	
Part of speech	Part of speech values are structured with a top level set composed of 10 values like **noun** or **verb**. A very precise ontology is specified for grammatical words. Most of parts of speech are common to lexicons and annotations but two set of values (i.e. **punctuation** and **residual**) are specific to annotation and are not usually used in lexical descriptions[4].	120	
Register, dating and frequency	Constants include, for instance, **slangRegister** or **rarelyUsed**.	19	
Total		380	items

In contrast to the values of the morpho-syntactic profile, which mainly concern the lexicon, most values in the syntactic profile deal with annotation.

Syntactic profile:

Basics	These are general purpose annotation constants, like: **tagging**, **standoffNotation**, **embeddedNotation**. A few of them like **negation** or **contiguous** concern lexicons.	29	items
Constituency	These comprise constants used to annotate constituency elements. Examples of values are: **chunk**, **declarativeClause**, **verbNucleus**, **nounPhrase**. Usual abbreviations like NP for **nounPhrase** are declared in the name section of the data category.	27	
Dependency	These comprise constants used to annotate relation between syntactic elements. Examples of values are: **verbModifier**, **modifier**, **syntacticHead**, **subject, introducer, directObject**, **coordination**, **adjunct**. Let us note that a certain freedom is left to the user concerning the level of detail and the type of target: for instance, both **verbModifier** and **modifier** are proposed.	32	
Total		88	items

11. Problems encountered

As said earlier, we started from existing lists that are rather stable like those for Eagles or Multext-East. The problems that we encountered were that we had to write definitions. We searched in various sources and found some definitions that looked fine in isolation for some data categories, but they did not constitute a coherent set of definitions.

Linguistics is not a field with a common agreement on basic terms. As a matter of example, the entry "morphology" in Wikipedia, gives us a good view of these divergences. In linguistics, terms like "paradigm", "collocation", "morpheme", "ergative" have so many interpretations in the different theories that they are almost impossible to use in a normative context where a precise meaning is required.

Another problem we faced was that we had to write definitions that are valid for lexicons and annotation, and an important term like "word" does not have the same meaning in both contexts. A word in a lexicon is lexical entry that is associated with a lemma. A word in an annotation is an occurrence of an inflected form (in

[4] For the people working in terminology and lexicons, punctuation is usually not considered as a part of speech. The situation is rather different when the objective is to represent text specific structures like coordination in the context of syntactic annotation, in this case, a punctuation mark is usually considered as a plain word, and as such, needs a part of speech tagging.

an inflected language). Theses notions are rather different.

To deal with this problem, we carefully avoided dangerous terms and we delimited a secure set of terms. When needed, we formed multi-word expressions from secure components. This is the strategy that has been adopted in the DCR and in general within the ISO-TC37 family of standards.

12. Forthcoming data

The current database records values for West/East European languages and, to a certain extent, for Semitic languages. The rationale for such a strategy is that, first, it was easier for us to begin by these values because stable lists already existed for these languages. Secondly, we faced a "chicken and egg" situation: we rely on ISO voluntaries and no one will describe minority languages if the well-known languages were not covered.

We know that it is clearly not enough

Two other parallel tasks are currently being conducted. One task deals with Asian values within the NEDO project (Takenobu et al, 2006; Charoenporn et al, 2007; Shirai et al, 2008). A small set of values has been entered in the database. The other task deals with African values, and a study is being conducted by the ISO South African delegation, but the values have not been entered yet in the database.

Each value is associated with a version number to allow a stable compliance in case of modification. The rules for management and usage are defined in the ISO-12620 revision.

13. Forthcoming registry

The current system is rather simple. It permits to make simple interactive queries, to download the result of a query, to download a data category, a directory or a profile. The available formats are XML and HTML.

The registry has been populated with numerous data categories, but different users (including ourselves) asked for an upgrade with improved interface features and fully developed functionalities.

An improved model is currently being designed (2007-2008) in order to address two important issues namely the distinction between the language section (working language) and linguistic section (object language) and the ability to record constraints and richer relations. Another difference is that the relation "broader" has been renamed into "IsA".

The new model will be implemented in a system called "ISOcat" at http://www.isocat.org. This new system is currently in beta version and will be presented during LREC-2008 and described in (Kemps-Snijders et al, 2008; Wittenburg et al, 2007).

Instead of being based on traditional synchronized PHP programs, the new software is based on Java/Ajax technologies and promises to be more user friendly. The operational switch from Syntax to ISOcat is scheduled for the end of 2008.

14. Conclusion

The registry is far from being complete but it begins to be used within different ISO-TC37 based standard applications in order to be tested. The idea is to progressively increase the number and coverage of these data categories. The ambition is that the registry will become the reference point when using linguistic terms and data elements in lexicons and annotations in NLP context.

15. Acknowledgements

The work presented here is partially funded by the EU eContent-22236 LIRICS project and in part by the French ANR-Passage project (Action ANR-06 MDCA-013).

16. References

Charoenporn T., Thoongsup S., Sornlertlamvanish V., Isahara H. (2007) Thai Lexicon. SEALS Conference, Univ of Maryland, College Park. US

Declerck T. (2006) SynAF: Towards a standard for syntactic annotation. LREC Genoa.

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. (2006) Lexical Markup Framework (LMF). LREC Genoa.

Ide N., Romary L (2004) A Registry of Standard Data Categories for linguistic Annotation. LREC Lisboa.

ISO-12620:1999, Computer application in terminology - Data categories, ISO Geneva

Kemps-Snijders M., Windhouwer M., Wittenburg P., Wright S.E. (2008, forthcoming) A revised Data Model for the ISO Data Category Registry, submitted to TKE-2008, Copenhagen.

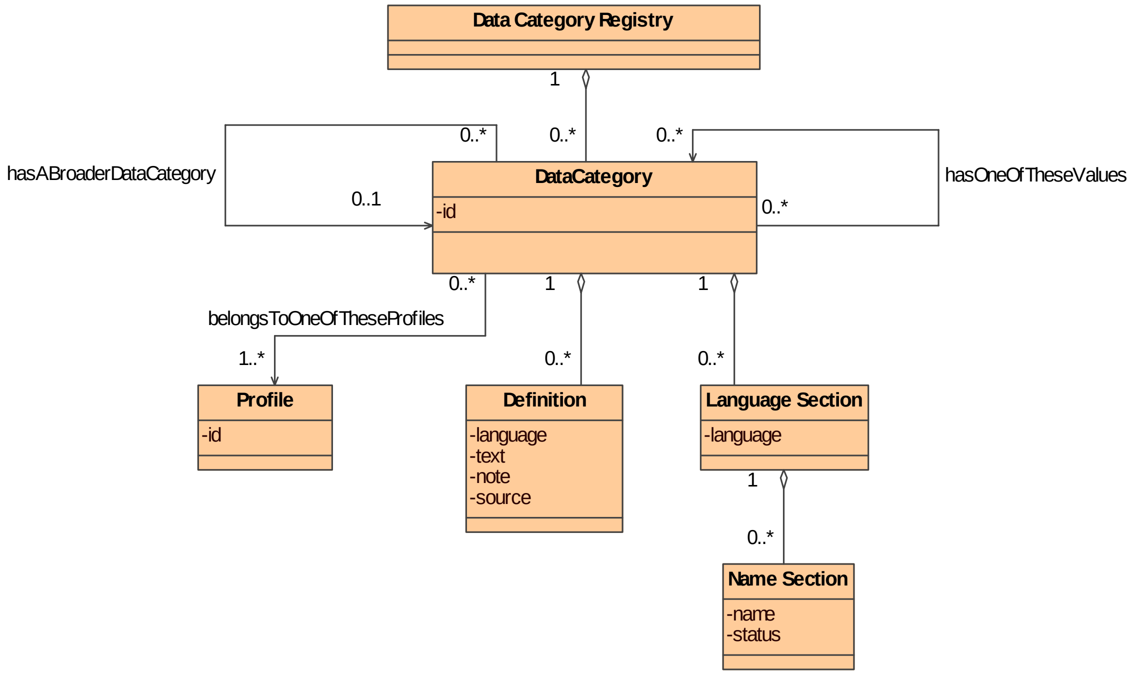
Shirai K., Tokunaga T., Huang CR., Hsieh SK, Kuo TY., Sornlertlamvanich, Charoenporn T. (2008) Constructing Taxonomy of Numerative Classifiers for Asian Languages IJCNLP Hyderabad, India

Takenobu T., Sornlertlamvanich V., Charoenporn T., Calzolari N., Monachini M., Soria C., Huang CR., Hao Y., Prevot L., Kiyoaki S. (2006) Infrastructure for standardization of Asian language resources COLING/ACL Sydney, Australia

Wittenburg P., Wright S.E. (2007) Infrastructure note on registry databases: technical note at http://www.tc37sc4.org/new_doc/iso_tc37_sc4_N436_ontology_memo_peter_Sue_busan2007.pdf

Wright S.E. (2004) A global data category registry for interoperable language resources: technical note at http://www.tc37sc4.org/new_doc/ISO_TC_37-4_N175_SEW-A_Global_Data_Category_Registry.pdf

Annex-A: UML class diagram of the portions of the current registry that we use



Annex-B: current set of values

Morpho-syntax: Basics

- agreement
- any
- approximate
- be
- coding
  - characterCoding
  - countryCoding
  - dateCoding
  - languageCoding
  - scriptCoding
- comment
- creationDate
- definition
- direction
- domain
- exact
- example
- expletive
- externalReference
- externalSystem
- have
- id
- image
- impossible
- label
- language
- leftEnvironment
- lexeme
- logicalOperator
  - logicalAnd
  - logicalNot
  - logicalOr
- logicalValue
  - no
  - yes
- macron
- namedEntity
- numValue
- pluralType
- position
- possible
- quotative
- rank
- reduplicationFunction
- reduplicationType
- required
- restriction
- rightEnvironment
- scope
- sound
- source
- space
- stringValue
- text
- type
- unspecified
- utterance
- value
- variation
- view
- word

Morpho-syntax: Cases

- case
  - abessiveCase
  - ablativeCase
  - absolutiveCase
  - accusativeCase
  - adessiveCase
  - aditiveCase
  - allativeCase
  - benefactiveCase
  - causativeCase
  - comitativeCase
  - dativeCase
  - delativeCase
  - elativeCase
  - equativeCase
  - ergativeCase
  - essiveCase
  - genitiveCase
  - illativeCase
  - inessiveCase
  - instrumentalCase
  - lativeCase
  - locativeCase
  - nominativeCase
  - obliqueCase
  - partitiveCase
  - prolativeCase
  - sociativeCase
  - sublativeCase
  - superessiveCase
  - terminativeCase
  - translativeCase
  - vocativeCase

Morpho-syntax: Form Related

- affix
  - infix
  - prefix
  - suffix
- affixRank
- allomorph
- apocope
- componentRank
- conjugated
- contextualVariation
- expansionVariation
- geographicalVariant
- graphicalSeparator
- homograph
- homonym
- homophone
- lemma
- lexicalType
- morpheme
  - etymologicalRoot
- native
- orthographyName
- patternType
- phoneticForm
- phoneticSeparator
- pinyin
  - nonSpacedPinyin
  - spacedPinyinAndTone
- reduplication
- root
- script
- stem
- stemRank
- symbol
- token
- writtenForm

Morpho-syntax: Morphological Features Excluding Cases

- activeVoice
- animate
- aorist
- bound
- cessative
- collective
- commonGender
- comparative
- conditional
- definite
- dual
- elInclusion
- elative
- feminine
- finite
- firstPerson
- fullArticle
- future
- gerundive
- honorific
- imperative
- imperfect
- imperfective
- inanimate
- inchoative
- indefinite
- indicative
- indifferent
- infinitive
- intensity
- masculine
- masdar
- middleVoice
- morphologicalFeature
  - animacy
  - aspect
  - cliticness
  - definiteness
  - degree
  - finiteness
  - grammaticalGender

- grammaticalNumber
- grammaticalTense
- modificationType
- negative
- ownedNumber
- ownerGender
- ownerNumber
- ownerPerson
- person
    - objectPerson
    - subjectPerson
- syntacticType
- verbFormMood
- voice
- zuInclusion

neuter
nonFinite
otherAnimacy
participle
passiveVoice
past
paucal
perfective
personal
plural
- brokenPlural

positive
possessive
postModifier
preModifier
present
quadrial
referentType
secondPerson
shortArticle
singular
subjunctive
superlative
thirdPerson
trial
unaccomplished

Morpho-syntax: Operations

- abbreviation
- elision
- location
- operation
    - add
    - addAffix
    - addAfter
    - addBefore
    - addComponentLemma
    - addComponentStem
    - addFirstConsonant
    - addFirstVowel
    - addLemma
    - addLowerCaseComponentLemma
    - copy
    - derivation
    - remove
    - removeAfter
    - removeBefore
    - substitute
- operator
    - graphicalOperator
    - phoneticOperator
- romanization
- rule
- scheme
- transcription
- transformType
- transliteration

Morpho-syntax: Part of speech

- adjective
    - ordinalAdjective
    - participleAdjective
        - pastParticipleAdjective
        - presentParticipleAdjective
    - qualifierAdjective
- adposition
    - circumposition
    - postposition
    - preposition
        - compoundPreposition
        - fusedPreposition
        - simplePreposition
- adverb
    - generalAdverb
    - particleAdverb
- classifier
- conjunction
    - coordinatingConjunction
    - subordinatingConjunction
- determiner
    - article
        - definiteArticle
        - indefiniteArticle
        - partitiveArticle
    - demonstrativeDeterminer
    - exclamativeDeterminer
    - indefiniteDeterminer
    - interrogativeDeterminer
    - possessiveDeterminer
    - reflexiveDeterminer
    - relativeDeterminer
- interjection
- noun
    - commonNoun
    - countableNoun
    - diminutiveNoun
    - massNoun
    - properNoun
- numeral
    - numeralApprox
    - numeralBoth
    - numeralDigit
    - numeralLetter
    - numeralMForm
    - numeralRoman
- partOfSpeech
- particle
    - affirmativeParticle
    - comparativeParticle
    - conditionalParticle
    - coordinationParticle
    - distinctiveParticle
    - futureParticle
    - infinitiveParticle
    - interrogativeParticle
    - modalParticle
    - negativeParticle
    - possessiveParticle
    - relativeParticle
    - superlativeParticle
    - unclassifiedParticle
- pronoun
    - affixedPersonalPronoun
    - allusivePronoun
    - conditionalPronoun
    - demonstrativePronoun
    - emphaticPronoun
    - exclamativePronoun
    - impersonalPronoun
    - indefinitePronoun
    - interrogativePronoun
    - negativePronoun
    - personalPronoun
        - strongPersonalPronoun
        - weakPersonalPronoun
    - possessivePronoun
    - reciprocalPronoun
    - reflexivePronoun
    - relativePronoun
- punctuation
    - closePunctuation
        - closeBracket
        - closeCurlyBracket
        - closeParenthesis
    - mainPunctuation
        - declarativePunctuation
            - exclamativePoint
            - point
            - semiColon
            - suspensionPoints
        - interrogativePunctuation
            - questionMark
            - invertedQuestionMark
    - openPunctuation
        - openBracket
        - openCurlyBracket
        - openParenthesis
    - secondaryPunctuation
        - bullet
        - colon
        - comma
        - hyphen
        - invertedComma
        - quote

-
  -
    - slash
  - unclassifiedPunctuation
- relationNoun
- residual
  - foreignText
  - foreignWord
  - formula
  - letter
  - unclassifiedResidual
- verb
  - auxiliary
  - copula
  - mainVerb
  - modal
- voiceNoun

Morpho-syntax: Register Dating Frequency

- benchLevelRegister
- commonlyUsed
- dating
- dialectRegister
- facetiousRegister
- formalRegister
- frequency
- inHouseRegister
- infrequentlyUsed
- ironicRegister
- modern
- neutralRegister
- old
- rarelyUsed
- register
- slangRegister
- tabooRegister
- technicalRegister
- vulgarRegister

Syntax: Basics

- annotation
  - morphosyntacticAnnotation
  - syntacticAnnotation
- annotationDeepness
- annotationStyle
- annotationType
- clitic
  - enclitic
  - proclitic
- constituency
- constituencyAndDependency
- contiguous
- deepParsing
- dependency
- doubleNegation
- embeddedNotation
- first
- mixedNotation
- negation
- next
- predicate
- previous
- propagation
- shallowParsing
- standoffNotation
- syntacticFeature
- tagging
- whType
- yesNoType

Syntax: Constituency

- grammaticalUnit
  - chunk
    - adjectiveChunk
    - adpositionChunk
    - adverbChunk
    - nounChunk
    - postpositionChunk
    - prepositionChunk
    - verbNucleus
  - clause
    - declarativeClause
    - imperativeClause
    - interrogativeClause
    - relativeClause
  - phrase
    - adjectivePhrase
    - adpositionPhrase
    - adverbPhrase
    - comparativePhrase
    - coordinatedPhrase
    - nounPhrase
    - postpositionPhrase
    - prepositionPhrase
    - prepositionVerbPhrase
    - superlativePhrase
    - verbPhrase
  - sentence

Syntax: Dependency

- adjunct
- apposed
- apposition
- attribute
- auxiliary
- complementizer
- coordination
- coordinator
- directObject
- function
- head
- introducer
- juxtaposition
- leftCoordinated
- modifier
  - adverbModifier
  - nounModifier
  - postnominalModifier
  - prenominalModifier
  - prepositionModifier
  - verbModifier
- relation
  - comparativeRelation
  - genitive
  - relativeRelation
  - superlativeRelation
- rightCoordinated
- subject
- syntacticArgument
- syntacticHead
- verbComplement

On the Relevance, Standards and Usage of Metadata for electronic language resources

Peter Wittenburg, Daan Broeder
MPI for Psycholinguistics.
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
E-mail: peter.wittenburg@mpi.nl

Abstract

This paper gives an overview on past, current, and possible future of descr iptive metadata standards for language resources. The description of typical usages and the state of standardization provides general input for deciding about which standard to build on. The short outline of the future sets an agenda how to accommodate necessary diversity without completely sacrificing the possible benefits of standardization.

1. Introduction

We will name metadata for electronically accessible resources that are comparable with library cards describing publications with a number of keywords "Descriptive Metadata" (DMD). This sort of metadata has a very old tradition in the library world, since it functioned as fingerprints for publications for all sorts of retrieval, referencing and management activities. In particular in the After-Gutenberg area when the number of books and copies increased dramatically the usage of library cards became indispensable. People needed to make a difference between a book as described by a few of its typical attributes such as author, title and year of publication and the various physical copies that were distributed. Library cards became abstract incarnations of publications. Needless to say that also museums and archives use the concept of DMD to describe and manage their large holdings of physical objects. In parallel large content classification systems such as Iconclass [1] were developed to allow experts to classify the content of paintings and sculptures and to add such classifications to the DMD.

Consequently, librarians were the first to introduce electronic library card systems and use computers for management and retrieval functions. Without being comprehensive we can refer to the MARC [2] schema(s) as an early and prominent representative of such electronic library classification scheme which remained in the classical domain of describing publications. Around 1995 a number of experts mainly with a library background started to work on the Dublin Core standard [3], since they felt the need to come up with a unified method to describe all sorts of content that became available. The Dublin Core initiative helped to raise discussions about many aspects and it certainly had a great impact on spreading the awareness of the need for DMD in the various disciplines, since they all were faced with an extremely fast growing number and diversity of digital resources.

While the Dublin Core experts worked on a discipline crossing framework with a common and by nature unspecific vocabulary, an increasing number of experts in various disciplines started to work on schemes for DMD with more discipline or topic oriented foci. Here we can refer for example to LOM [4] for teaching materials and ISO 19115 [5] for Geographic Information.

In the language resource domain we can refer to a number of early attempts to introduce formally described DMD with limited scope. The TEI initiative [6] worked out a set of header tags to describe in large detail various aspects of different types of resources such as lexica. We can also refer to initiatives such as CHILDES [7] that introduced header records as part of their resources to help in retrieving and in formulating scientific questions. At the MPI a consequent digitization policy from 1995 lead to an enormous increase in digital resources requiring a first institute wide DMD solution in 1998 also triggering thoughts about resource management aspects.

At LREC 2000 the first workshop about metadata in our domain [8] was organized where amongst others a summary of the ISLE White Paper on metadata [9] was presented. A first ISLE working group meeting resulted in the IMDI initiative which presented its framework existing of a schema and tools for editing, browsing and searching and an extension to describe lexica at LREC 2002 [10, 11] and which was continuously upgraded until now. At the LREC 2000 workshop Simons and Constable [12] also argued for the need for a proper language identification schema which now resulted in ISO 639-3 - a standard largely based on SIL's Ethnologue work [13]. In December 2000 Bird and Simons organized the first OLAC workshop [14] where they presented the OLAC initiative [15] that also included a proposal for building a metadata framework. In contrast to the IMDI work that started with bottom-up discussion processes to determine the wishes of the discipline experts and that decided to use discipline terminology, OLAC made the decision to start from the Dublin Core vocabulary and to add a few important categories to satisfy special needs of the community.

Both IMDI and OLAC harvest metadata records from a number of institutions and agree on using the OAI PMH (protocol for metadata harvesting) [16] as a standard for

DMD exchange. In the case of IMDI a gateway is provided that maps IMDI concepts to DC terms where possible and that offers all resources for harvesting by service providers such as OLAC or library initiatives. Despite these early attempts the total amount of resources visible via this kind of open, structured and harvestable DMD is not at all satisfying. As overviews such as ENABLER [17] have shown too few institutions and projects until now took the effort to create DMD and even fewer offer their DMD so that it can be harvested. Many are at the level of offering their holding via typical web-sites.

However, the long and broad experience in the field will allow us to draw conclusions about the usage and limitations of our current DMD frameworks. Also by borrowing new methods and standards from other initiatives such as W3C [18], TEI and ISO TC37/SC4 [19] we will be able to derive a new more flexible standard for DMD in the field of language resources and tools.

2. Usage of Descriptive Metadata

During the recent 8 years it became much more obvious what the possibilities for DMD in future scientific information management and usage will be. First of all DMD in general adds valuable additional information to resources, information that is mostly not encoded in the resource itself. This information can have different character such as administrative, describing the content, describing technical and usage details, etc. Basis for such added value are classification steps since resources need to be associated with a certain language, a certain genre, a certain group of creators, a certain subject etc.

2.1 Query/View/Browsing

Of primary relevance is DMD to help researchers and other users to find useful data to answer a given research question. Dependent on the researcher's background the queries will be more or less detailed. Linguists will be interested in combining metadata and content queries such as for example for a longitudinal study: *Give me the frequencies of correct usage of the 3. person plural inflectional forms for 3, 4 and 5 years old children and allow me to compare between boys and girls.* Such queries have a direct research impact. Non expert users may ask more general queries such as *where can I find resources about the Kuikuru language* to then have a quick look whether they are relevant for the work. In both cases the user expects that the DMD will provide a direct link to the resources themselves to be able to utilize them. Increasingly often researchers will want to create their own virtual collection probably by virtually combining resources from different repositories. Actually, the collection building is carried out by re-grouping and linking DMD descriptions, the resources in general will not be moved for many reasons. Such virtual collections will then form the basis for ongoing scientific work and the context needs to be preserved for documentation purposes. In all these cases DMD are research tools.

For other users metadata could be used for general discovery purposes or to advertise resources and facilitate access. DMD could be linked with geographical locations or used to group resources dynamically to exhibition like web presentations and portals. For large online repositories such as Flickr [20] and YouTube [21] the principle of social tagging became very popular where users associate values with certain fields. Also in these cases DMDs are created for later discovery and grouping purpose based on individual and non constrained classification steps.

2.2 Information Management

The other important pillar for DMD creation and usage is information management. In scientific repositories or archives with a long-term preservation intention authenticity, proper classification and grouping are essential. DMD can be the basis of various management operations such as copying, moving, associating access rights, migration of formats, checking consistence, etc. Relating various resources with each other has a completely different function, since for management tasks it is crucial to treat for example a sound recording and a video recording created at the same time and describing the same event and various annotations of them as one unit. Here metadata becomes the function of the glue that allows managers to bundle resources and therefore to facilitate sensitive operations.

So far DMD are used for interpretation by humans either facilitated by a search engine or by visually browsing. In future eScience scenarios increasingly often we will have to support machine driven operations such as the creation of (semi) automatic abstractions, i.e. hierarchically grouping resources in new data driven manners, or as automatic profile matching to suggest alternatives for certain tasks.

3. State of Standardization

First, we need to discuss what we can call a standard. Obviously a wide usage of a specified rule does not mean that this rule can be called a "standard". In addition the rule needs to have passed the process specified by an internationally accepted standardization organization such as ISO which has formed a special sub-committee TC37/SC4 devoted to standards in language resource management. Since these processes in general take much time we also see new widely accepted initiatives such as W3C which speed up processes in new areas as for example the Internet. Also Dublin Core and TEI can be called widely respected initiatives. Finally, we have quasi standards defined for example by great monopolies such as Microsoft. Important for the acceptance of a certain rule is the belief that investments will not be wasted and that respected organizations can guarantee some stability over a longer period in time.

First, we will discuss the state of Dublin Core (DC). DC finally specified its 15 elements. It is widely understood that the semantics of the DC elements are not specific enough to be satisfying for scientific purposes, but that they are meant for general web discovery support. The DC initiative added the so called qualified DC elements which can partly be seen as semantic refinements of the general terms. While DC:Date can be any date, qualified DC offers for example the "creation date" of which the semantics is more precise. The DC elements are offered under stable and persistent identifiers, thus allowing projects to include them in so-called application profiles which are schemas specifically tailored for projects etc. Re-using DC elements introduces semantic interoperability in particular when the element semantics are specific enough. We can assume that the DC elements will be maintained, nevertheless an official recognition as a standard would be welcome.

With P5 TEI [22] has launched a new version stabilizing its header descriptors. In some projects such as national corpus projects relevant TEI header descriptors have been used. Yet for typical DMD TEI terminology is not that frequently used. Another very important achievement of TEI is the ODD framework [23] allowing users to establish complete schemas from various pre-registered component schemas. This framework was already used when implementing LMF [24] the flexible component based lexicon framework standardized recently by ISO.

The IMDI element set [25] has been stabilized over the years and has been integrated into the Data Category Registry maintained by ISO [26]. Therefore it is ensured that the elements will be maintained over the next decades and that they can be identified by unique and persistent identifiers. The structured IMDI schema is stable as well, however, new requirements such as the necessity of storing persistent and unique identifiers to refer to the resources require minor extensions were added in a downwards compatible manner. Also in the IMDI case it was understood that it does not make sense to standardize the schema. The IMDI schema offers possibilities for extensions. Users can add key-value pairs at various parts specific for their needs, however, these keys and values are not defined and therefore might only have relevance for an individual instance. IMDI can also be extended by special profiles that are predefined sets of such extra key-value pairs. Such as it was for example necessary to add a whole set of descriptive elements to satisfy the specific needs of the sign language community. IMDI profiles can be seen as an intermediate step in so far as such profiles are meant to support relevant sub-disciplines. However, the added elements are not part of the IMDI set and schema, but they are made available via editor and search engine. Extensions are not yet registered in recognized concept registries and there is not yet a suitable process for adding relevant concepts to a registry.

The OLAC set [27] is making use of the DC set of 15 elements and defines additional qualifiers and vocabularies specific for the linguistic domain. Such as qualifiers for DC:Subject (olac:language, olac:linguistic-field) DC:Type (olac:linguistic-type, olac:discourse-type) and DC:Creator/DC:Contributor (olac:role) Also OLAC has possibilities to add extensions and refinements, these are supported in the schema. Similar as for IMDI the OLAC initiative is probably too small to guarantee long-term survival of the defined concepts. A registration at a large and well-supported concept registry will become necessary in the long-term.

Widely accepted is the OAI PMH standard as a format to exchange metadata between data providers and service providers. It is a light-weight protocol that allows data providers to offer schema-based formats of their own choice, however, each provider also needs to deliver records making use of the Dublin Core semantics. For data providers such as IMDI that are using other vocabularies the conversion in general means a reduction of the information that is offered. It is left to the service providers to determine how they will offer services, i.e. what kind of vocabulary they are offering to the users, whether they want to reconstruct browsable hierarchies and what the granularity of the services are. While for specialists it may be of relevance to have detailed access to specific records of a certain collection, non-specialists may be satisfied with one single hit representing the whole collection. Since metadata should be open and at least in archive environments XML schema based files are accessible, often a much more simple harvesting method is applied. Service providers simply read the web-accessible descriptions, parse them and generate the type of service they are aiming at. This method simply requires the accessibility of the DMD via the HTTP protocol.

4. Experience

Since in particular the IMDI and OLAC infrastructures are around for about 7 years now, we can draw a number of conclusions from the experience so far.

- Both sets, IMDI and OLAC, have stabilized over the years and offer solid infrastructures. OLAC is focusing on cross-repository services. Although it offers an editor to create descriptions, its main focus is on acting as service provider, i.e. harvest DMD via the OAI PMH, requiring a low granularity[1] and offering a search engine to look for interesting resources applying the OLAC/DC vocabulary. IMDI offers a structured schema allowing describing resources and resource bundles in greater detail. This enables users to formulate queries that are directly relevant for

[1] From the IMDI domain only DMD are accepted that represent language-oriented collections.

research questions. It comes with an editor, allowing to create DMD and to embed them in browsable hierarchies, it allows depositors to create canonical hierarchies that can be used for management purposes and users to create their own private hierarchies, it offers native XML and HTML browsing facilities by applying on the fly XSLT transformations, it offers a gateway to act as full OAI PMH data provider, it harvests data from other registered data providers and it offers structured and unstructured search options. REST interfaces allow users to embed for example metadata queries in web portals.

- The coverage has grown during the last years and is impressive in spite of the limited funds that could be used. However in total the amount of language resources that have been registered and that are accessible via the portals is very small compared to the amount of language resources that have been created. Therefore we cannot speak about a satisfying solution. The reasons for this are very heterogeneous. Here we can only mention the most important ones: (1) Metadata creation is expensive and extra work, which is in general not budgeted for. (2) Researchers still lack convincing arguments to invest in efforts for the benefit of other users. (3) The available metadata sets could be not useful since their schema and terminology are not appropriate for the resources to be integrated. Users often want to be able to tailor their sets to their needs. (4) Available knowledge about existing language resources is often such that even the responsible researchers don't know exactly how to classify them and where they are exactly physically stored. (5) Some researchers still see their resources as their private capital which they don't want to share. (6) In some cases there are ethical considerations or privacy reasons that forbid even publishing metadata about resources.
- From broad discussions in our discipline we know that terminology and localization issues are crucial for researchers to create DMD. Sub discipline terminology is different from what is used in sets such as IMDI and people hesitate to use non-familiar vocabularies to classify their resources. Missing support for a working language is also a point of uncertainties.
- Even for professional frameworks such as IMDI with ample technical support, we can see that the willingness to create richly filled in metadata descriptions is rather low and that adherence to standards is not guaranteed. Statistics carried out on 27.000 metadata records in the MPI and DOBES archive [28] show that some fields such as for example "genre" are not used since categorization is seen as problematic or since it costs too much time to decide about it and that other fields such as language ID are not used properly since it would cost too much time to look up in the integrated Ethnologue list what the exact ID is.
- Right now we see the first real applications where depositors themselves see a benefit from investing time for metadata creation. Archives such as the one at the MPI with about 60 Mio annotations for resources from many different teams and a large variety of languages are of a size and richness that it makes sense for a researcher to formulate scientific queries that contain metadata constraints to restrict the collection on which content queries can be formulated. Another application where metadata is required is dynamic portals that exhibit the richness according metadata categories such as "genre". On the fly metadata queries can present those resources that contain for example stories about certain subjects etc. A third increasingly accepted argument for the usefulness of metadata is its importance for building virtual collections suitable to work on a certain research question. Without the need for copying the real resources users can simply copy and recombine metadata descriptions for this purpose.
- The general pressure from funding agencies is growing to produce well-organized and well-described collections at the end of funding periods. Also the insight of the disciplines is growing that accessibility and re-usage of digital resources is ultimately dependent on proper metadata descriptions and proper digital archiving.
- The seven years of experience has resulted in a much deeper understanding of a number of "technical" problems that need to be solved such as metadata granularity, resource bundling, mapping between different metadata vocabularies, irrelevance of schemas for semantic interoperability in most cases of discovery, usefulness of registering concepts in open registries as basis for semantic interoperability, granularity of concept descriptions in such registries etc.

5. Future

Based on these experiences we can now come to some conclusions for further standardization work. (1) We will need a flexible component based framework for metadata creation and handling, where each researcher can compose his/her own tailored schema partly based on pre-defined components and where interoperability is guaranteed by the duty to only use concepts that are registered in widely accepted and persistent registries. (2) Each schema should itself be registered to allow search engines to make use of the specifications that are required for doing cross-collection operations. (3) The concept definitions in such registries need to support sub-discipline terminology provided for all relevant languages. (4) Relation registries need to be maintained to allow search engines and mapping algorithms to bridge between different semantic domains identified by different name spaces. (5) All needs to be based on unique and persistent identifiers that refer to the schema components and concepts used. (6) Centers need to offer easy deposit and resource registry services allowing even individual researchers to upload their resources and DMD to make them visible and accessible. (7) Much more evangelization is required to inform researchers about standards, costs and benefits of metadata descriptions.

We are convinced that such a framework as described above will dramatically increase the acceptance of descriptive metadata. The availability of increasingly more DMD referring to accessible resources will help convincing researchers about their usefulness. It also will motivate researchers to not only make use of "their" collection but to extend their research work to related collections dependent on the research question.

To ensure the expensive investments that are necessary for creating DMD we need to use organizations such as ISO TC37/SC4 to take care of a worldwide acceptance and the stability of the chosen solutions.

6. References

[1] 1990 : C. Gordon, "An Introduction to Iconclass", in: A. Roberts (ed.), Terminology for Museums. Proceedings of an International Conference, Cambridge, England, 21-24 September 1988, Cambridge 1990, 233-244

[2] http://www.loc.gov/marc/

[3] http://dublincore.org/

[4] 1484.12.1-2002 IEEE Standard for Learning Object Metadata. 1484.12.3-2005 IEEE Learning Technology Standard - Extensible Markup Language (XML) Schema Definition Language Binding for Learning Object Metadata

[5] ISO 19115:2003 Geographic information -- Metadata

[6] TEI, C.M. Sperberg-McQueen and L. Burnard. Guidelines for Electronic Text Encoding and Interchange, XML-Compatible Edition (TEI P4). Text Encoding Initiative Consortium and University of Virginia Press, 2001. URL: http://www.tei-c.org/P4X/

[7]] CHILDES or CHAT format, Child Language Data Exchange System; http://childes.psy.cmu.edu

[8] Proceedings of LREC 2000 Workshop on Meta-Descriptions and Annotation Schemes for Multimodal/Multimedia Language Resources, LREC 2000, Athens

[9] Wittenburg, P., Broeder, D., Sloman, B.; Meta-Description Standard for Multi-Media Language Resources. LREC 2000 Workshop on Meta-Descriptions and Annotation Schemes for Multimodal/Multimedia Language Resources, LREC 2000, Athens

[10] Broeder, D., Offenga, F., Willems, D.; Metadata Tools Supporting Controlled Vocabulary Services. LREC 2002, Las Palmas

[11] Wittenburg, P., Peters, W., Broeder, D.; Metadata Proposals for Corpora and Lexica. LREC 2002, Las Palmas

[12] Constable, P., Simons, G.; Language identification and IT - Addressing problems of linguistic diversity on a global scale. LREC 2000 Workshop on Meta-Descriptions and Annotation Schemes for Multimodal/Multimedia Language Resources, LREC 2000, Athens

[13] http://www.ethnologue.com/

[14] 2000: Bird, S Simons, G White Paper on Establishing an Infrastructure for Open Language Archiving. Workshop on Web-Based Language Documentation and Description, Philadelphia, PA page 12-15

[15] http://www.language-archives.org/

[16] http://www.openarchives.org/

[17] Calzolari N., Choukri K., Gavrilidou M., Maegaard B., Baroni P., Fersøe H., Lenci A., Mapelli V., Monachini M., Piperidi S. (2004) ENABLER Thematic Network of National Projects: Technical, Strategic and Political Issues of LRs. In LREC 2004 Proceedings, Lisbon.

[18] http://www.w3.org/

[19] http://www.tc37sc4.org/

[20] http://www.flickr.com/

[21] http://www.youtube.com/

[22] TEI P5 Guidelines for Text Encoding and Interchange. TEI Consortium, ACH, ACL, ALLC
CM Sperberg-McQueen, L Burnard - 2005 - Oxford Providence Charlottesville Nancy

[23] Sebastian Rahtz, Converting to schema: the TEI and RelaxNG. Available from http://www.idealliance.org/papers/dx_xmle02/papers/03-03-08/03-03-08.html. Paper presented at XML Europe 2002, Barcelona, May 2002.

[24] G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, C. Soria Lexical Markup Framework (LMF) In: Proc. of the International Conference on Language Resources and Evaluation (LREC), 2006.

[25]Broeder, D. and Wittenburg, P. (2006) The IMDI metadata framework, its current application and future direction. International Journal of Metadata, Semantics and Ontologies. 1(2): 119–132.

[26] http://lux12.mpi.nl/isocat/

[27] http://www.language-archives.org/OLAC/metadata.html

[28] http://www.delaman.org/docs/meeting06/broeder-imdi-usage.ppt

Applying the Corpus Encoding Standard to a quadrilingual corpus of legal texts

Verena Lyding
EURAC Research
Viale Druso1, 39100 Bolzano/Italy
E-mail: verena.lyding@eurac.edu

Abstract

In this article we describe the encoding of a quadrilingual corpus of legal text documents employing the Corpus Encoding Standard (CES). We start describing the purpose and particular characteristics of the corpus to be used in terminology work. In relation to the role of the corpus we discuss considerations concerning the usage and applicability of the standard for our case. After we conclude our discussion with the decision to apply the Corpus Encoding Standard for the encoding of primary and linguistic data, we go on documenting the implementation process. We systematically go through the different dimensions and stages of the encoding process. We describe decisions taken and the motivation behind it. We further point out areas where we had to deviate from the standard to meet our application’s needs. In particular we concentrate on describing the specificities of bibliographic information for legislative texts and on specifying the textual structure of legal documents.

1. Introduction

The creation of a quadrilingual corpus of legal text documents that is examined in this paper was carried out in the context of the LexALP project concerned with the creation and harmonization of multilingual terminology (www.eurac.edu/lexalp). The project aims at harmonizing terminology for the Alpine Convention in its four languages (Italian, German, French, Slovene). Terminology relevant to the fields of spatial planning and sustainable development is considered. Although the prevalent focus of the project clearly is a linguistic one with the main project aim being the creation of a terminology collection, an interdisciplinary working approach and the employment of modern technologies was a strong requisite. From the perspective of incorporating language technologies the creation of a multilingual corpus to serve as a textual basis in terminology work was a defined goal. Next to it the adaptation of an online terminology database (cf.. Sérasset et al., 2006) and the implementation of several tools for the extraction and employment of terminology were foreseen.

2. Point of departure in corpus creation

All considerations concerning corpus planning were controlled by the defined purpose of the corpus: to facilitate terminology work. Further restrictions regarding the ways to go about corpus implementation were imposed by consideration of cost and present infrastructure to reuse.

2.1 Requirements for a corpus for terminology work

The defined purpose of the corpus is to provide a basis for terminology extraction and serve as reference for the elaboration of terms and their usages. Terminology extraction was foreseen to be carried out automatically as well as manually. The elaboration of terms (comparing terms and their usages, searching for definitions and contexts) was to be done only manually. Since no specification of mandatory input formats for term extraction tools was available at the time of corpus creation, requirements were derived from demands of manual corpus investigation[1] only.

The most essential matter for terminology work is the corpus design, the selection of documents to be integrated in the corpus. Detailed information about criteria for document selection as well as the resulting corpus composition is given in section 3.1. Concerning the encoding [2] of texts, terminologists require access to detailed information about the source of every piece of text. Furthermore it is important for them to know about the position of a text fragment within the whole text. For terminology work it is often important to have access to textual context exceeding the sentence level. Sentence-level-alignment of documents that are available in several language versions are helpful for determining correspondences of words and phrases over languages.

2.2 Influencing factors concerning methodological and technical decisions

It was clear from the beginning of the project that due to funding limitations encoding of structural and linguistic information would have to be automatized wherever possible. Structural annotation was set as the prevalent aim over linguistic annotation. Previous corpus projects conducted at our institute suggested to stick to the in-house corpus environment, as implemented as a set of interconnected database tables to be searched over a web interface by employing the regular operators provided by the database query language (cf. Streiter et al., 2004). The corpus was intended for internal use as well as for public access. By providing corpus access to the general public we had to pay attention to copyright issues. Decisions on encoding had to take into account the particularities and differences of the different legal systems and languages

[1] With ‘manual corpus investigation’ we intend the targeted investigation with corpus querying tools.

[2] According to the CES documentation “’Encoding’ is any means of making explicit some interpretation of a text or collection of texts.”

involved.

3. Planning the corpus creation

Before starting with the compilation and processing of the corpus, decisions concerning the corpus design and the corpus encoding were taken. For the corpus design, a set of selection criteria was jointly determined by legal experts and terminologists. For the corpus encoding, the employment of a standard, the Corpus Encoding Standard (CES)[3], was evaluated and a decision was taken to what extent we would adopt this standard.

3.1 Corpus design

As described in Chiocchetti & Lyding (2006), we decided to select the documents that were to be integrated into the corpus by the following six criteria:

- entire documents only
- strong relevance to the subjects of spatial planning and sustainable development
- primary sources of the law for each legal system
- coverage of all official languages of the Alpine Convention
- latest amendments and versions of all legislation
- terminological relevance

Legal experts of the countries participating in the project were put in charge to select legal documents complying with the defined criteria.

The resulting corpus comprises more than 3000 legal documents relevant to the fields of spatial planning and sustainable development. The documents in four languages are taken from six national legal systems and three supranational frameworks (Chiocchetti & Lyding, 2006).

3.2 Employing an encoding standard

When planning the corpus creation, we aimed at adopting a standard for its encoding and markup, to adhere to good scientific practice and to foster sustainability and interchange with other initiatives. Furthermore, we were seeking advice concerning information to collect for documenting language text as well as concerning elements to employ for structural and linguistic annotation.

3.2.1. Examining the applicability of the Corpus Encoding Standard

To our knowledge, the Corpus Encoding Standard (CES and XCES – its XML instantiation) as a part of the EAGLES Guidelines is the most widely adopted standard for corpus encoding. The CES documentation defines the standard as "designed to be optimally suited for use in language engineering research and applications, in order to serve as a widely accepted set of encoding standards for corpus-based work in natural language processing applications" (CES, abstract). Although the primary purpose of our corpus is the employment in manual investigations and not in language engineering, there were still several reasons that made us consider the adoption of the CES.

In the first place, the general acceptance of the Corpus Encoding Standard and thus its resultant authoritativeness make it attractive. Despite the fact that the CES states to be designed for use in language engineering research and application, the general criteria to which it adheres[4] seemed equally sensible and suitable for our purposes.

Further, the detailed documentation –including recommendations concerning the elements to use for the descriptive representation of a corpus as well as guidelines as to how to apply them– promised to provide a helpful starting point for consistently going about corpus creation and documentation.

Given the aforementioned criteria at our institute for a corpus, it fitted our needs that the CES is intended for data interchange more than for local processing. Finally the CES aims at providing a "series of increasingly refined encodings of text" while it "does not require prohibitively large amounts of manual intervention to achieve minimum conformance to the standard" (CES, section 0.5). This feature of the CES is in line with our limited capacities for employing manual work.

3.2.2. Decisions taken concerning the adoption of the CES

The CES documentation (section 0.3) distinguishes the following encoding dimensions for which guidance is provided:

- a set of meta-language level recommendations (particular profile of SGML use, character sets, etc.);
- tagsets and recommendations for documentation of encoded data;
- tagsets and recommendations for encoding primary data, including written texts across all genres, for the purposes of corpus-based work in language engineering.
- tagsets and recommendations for encoding linguistic annotation commonly associated with texts in language engineering, currently including:
    - segmentation of the text into sentences and words (tokens),
    - morpho-syntactic tagging,
    - parallel text alignment.

For the purpose of the addressed corpus project we took the following decisions concerning the different encoding dimensions:

Meta-language level

Due to the fact that by now XML has widely replaced the application of SGML, we decided to adopt the XML

[3] The Corpus Encoding Standard (CES) is part of the EAGLES Guidelines developed by the Expert Advisory Group on Language Engineering Standards (EAGLES), http://www.ilc.cnr.it/EAGLES/home.html.

[4] The CES documentation defines the following encoding principles: adequate coverage, consistency, recoverability, validatability, capturability, processability, extensibility, compactness, readability (CES, section 1.5)

instantiation of the Corpus Encoding Standard (XCES). The XCES is still in the state of a beta release. While implementing the corpus, we constantly had to refer also to the more detailed CES documentation[5].

The CES recommends the use of the ISO 8859-X series[6] for text encoding. For representing the four languages Italian, French, German and Slovene, two different encodings of the ISO 8859-X series would have been needed. ISO-8859-1 (Latin1) does accommodate German, Italian and French. Slovene is covered by the ISO-8859-2 (Latin2). To fit the four languages into a single encoding, and considering that today UTF-8[7] is widely used, we decided to use UTF-8 encoding for all our data.

Documentation of encoded data

In the CES all data concerning the documentation of the corpus (global information about the text, its content and its encoding) and its contained texts are stored in the so called 'header' that is attached to each corpus document. It is an adoption of the TEI header customized "to suit the specific needs of corpus-based research" (CES, section 3).

For documenting encoded texts we decided not to stick to the CES. The data categories provided for the bibliographic description of the source text turned out to be too generic for encoding bibliographic indications of legal documents. Also, the structuring of information in the header seemed unwieldy and impractical for our purposes. The profile-related and encoding-related data categories were partly in accordance with our needs.

We decided to organize and accommodate documentation-related data in database tables as it was done in preceding projects. Providing means for automatically transforming the documentation data into CES header files might be feasible as far as the elements of our documentation scheme can be brought into accordance with the CES header recommendations. In doing so deriving minimal CES headers from our scheme will be less of a problem than integrating all of our documentation details into the CES header format.

Encoding of primary data

Our main aim in encoding primary data was to mark textual divisions down to the sentence level and to explicitly classify functionally meaningful text divisions and distinguish them from exclusively typographic divisions. Concretely we aimed for a markup scheme that would enable us to mark up different parts of legal documents such as preamble, annex, legal paragraph and to distinguish these divisions from typographic paragraphs. For encoding primary data we decided to adopt the CES recommendations which introduce a set of elements for text structuring. For all elements the CES specifies the syntactic and semantic scope and the structure on how to combine and order elements. Applying and partly refining the provided means for encoding primary data the CES was adopted up to level 1. Further, also sub-paragraph level annotations up to the sentence level were adopted. Manual verification of all introduced markup however could not be guaranteed.

Encoding of linguistic data

Encoding of linguistic annotation was restricted to the marking of sentence boundaries and to the alignment of multilingual documents. In both cases the CES was applied.

4. Implementation of corpus encoding

In the following sub-sections we will discuss our experiences with implementing the encoding decisions as described in the previous section. We will point out what difficulties arose and how we dealt with them for the specific case of creating a quadrilingual corpus of legal text for terminology work.

4.1 Meta-language level

We used the DTDs provided for XCES, which is still in beta status. The DTDs are not completely consistent with all encoding options described in the documentation. As a result, in some cases we had to take decisions according to our understanding of the documentation available for XCES.

Although we had decided to store all corpus documents in UTF-8 encoding, during document collection we could not insist on this. The collection of corpus documents was carried out as a collaborative task. Legal experts of each country participating in the project were in charge of selecting and collecting the relevant legislation. The project partners uploaded the corpus documents to a shared space via a web interface. Collaborators carrying out the document upload were often not familiar with characteristics of character encodings or even ways to determine encodings and convert between them. After document uploading was completed, we were faced with a collection of texts with different encodings. To comply with our aim of keeping the corpus in UTF-8 encoding, in a first step we had to determine the encoding for every document and do a conversion to UTF-8 where necessary.

4.2 Documentation of encoded data

As motivated in the previous section, we preferred to use our institute's conventions for encoding documentation data. In contrast to the CES recommendations, we did not associate every type of documentation information to each single corpus document, but distinguished between documentation data that holds for the corpus as a whole and documentation data that concerns the specific document. Concretely, general documentation on the corpus as a whole like information on encoding conventions is stored as text files without any strictly defined format. Documentation data related to single documents is organized in schematically defined database

[5] Although we chose to implement the XML instantiation of the CES and therefore consulted the CES documentation as well as the XCES documentation for reasons of simplicity we will continue referring to the CES documentation only.

[6] For details on ISO-8859 see http://www.iso.org

7 Defined by 'The Unicode Standard', http://www.unicode.org

entries and connected to its respective text documents via a unique identification number. Apart from this systematic difference between the CES recommendation and our approach, a great divergence from the standard also arose in the specification of elements for the encoding of bibliographic data. As mentioned in section 3.2.2., the elements provided by the CES did not suit our need for detailed description of legal documents.

In what follows we will briefly introduce the four main elements of a CES-compliant header and comment on their applicability as well as describe how these information categories were accommodated in our approach. In particular we will take a closer look at the representation of bibliographic data.

The CES header is described by the <cesHeader> element, which is allowed to contain the following four elements: <fileDesc>, <encodingDesc>, <profileDesc> and <revisionDesc>. Only the <fileDesc> element is obligatory. In the CES documentation these four elements are briefly characterized as follows:

<fileDesc>
contains a full bibliographic description of the corpus itself or of a text within it.

<encodingDesc>
documents the relationship between an electronic text and the source or sources from which it was derived.

<profileDesc>
provides further information about various aspects of a text, specifically the language used, the situation and date of its production, the participants and their setting, and a descriptive classification for it.

<revisionDesc>
summarizes the revision history for a file.

Next to including basic administrative categories like date of creation and character encoding, our main focus in collecting documentation data was the recording of extensive bibliographic details. In the CES the elements for describing bibliographic properties are subsumed under the <fileDesc> element. The overall purpose of the corpus together with the sampling rationale and the like are subsumed under the <encodingDesc> element in the CES recommendations. For our corpus this information does form part of the general documentation of the corpus and is not connected to single documents. Elements specifying language and text category are subsumed under the <profileDesc> element. This information is attached to single corpus documents also in our scheme.

The CES header further includes a <revisionDesc> element that holds information about the revision history of a corpus text. For our corpus, the revision status of single texts was done for groups of documents in a less systematic way.

Bibliographic information for legal documents differs greatly from bibliographic information for monographs or journal articles. While monographs or scientific articles are always associated with an author or a responsible institution, this does not hold in the same way for legal documents. On the contrary, for legal documents it is highly important to keep track of the legal system that they belong to, which is not foreseen to be indicated by the CES header.

Table 1 lists and defines the elements that are stipulated in the CES header under the <sourceDesc> element. In addition an indication is provided whether the element is applicable to the documentation of legal documents. The <sourceDesc> element is defined as an element for supplying "bibliographic description of the copy text(s) from which an electronic text was derived or generated" (CES, section 3.3).

Obligatory elements are printed in bold face.

CES elements for encoding bibliographic data	**Description**	**Applicable**
h.title+	Title of the bibliographic item	yes
(h.author | respStmt)?	Author of a work | Person or institution responsible for the intellectual content of a work	no
(edition, respStmt?)*	Details on an edition of the work (plus person or institution responsible for the intellectual content of the edition)	no
imprint+	Information relating to the publication of an item (such as publication place, publisher and publication date)	not relevant
idno*	Standard number for identifying the bibliographic item (like ISBN number)	no
(biblNote | biblScope)*)	Descriptive note supplying additional information on the bibliographic item | Scope of the bibliographic item such as total page number, or named subdivision	yes

Table 1: CES header elements for indicating bibliographic descriptions.

As becomes clear, only few categories provided by the CES are applicable to the bibliographic documentation of legal documents. At the same time the information categories needed for documenting a corpus of legal documents are rather too specific to form part of general guidelines like the CES. To meet the needs of very specific projects, while still providing general recommendations the CES could suggest different bibliographic schemes for different document types. The <biblStruct> element right under the <sourceDesc> element could provide an attribute named ‘type’ for indicating the specific document type. Ideally the set of document types as well as the information categories associated to it would be open to modifications and extensions by the user.

For the remaining of this section we will present the documentation scheme that was adopted for the LexALP corpus instead of the CES header. It was suggested by our legal experts and terminologists, based on existing in-house guidelines and experiences gained in previous projects

To bring into accordance the bibliographic aspects of nine legal systems and frameworks we were dealing with, elements common to all systems had to be defined. In addition also the mutual differences had to be taken into account. For every legal system and framework we provided a separate element set, which in each case subsumed the set of core categories that were defined as applicable to all systems. Further, we defined which elements are optional and which elements are obligatory. For some of the elements we constrained the data types. The schemes are implemented as database tables that are populated by the legal expert/terminologist via a web interface.

Table 2 shows the element set introduced for documenting texts of the German legal system. Elements belonging to the subset of elements applicable to all legal systems are underlined. Obligatory elements are indicated in bold face.

Element	**Description**
Full title	full title of legal document
short title	short title of legal document
abbreviation	abbreviation officially or commonly used to refer to the legislation
legal system	legal system of the legislation
legal hierarchy	legal level to which the legislation belongs (e.g. regional, national or supranational)
legal text type	type of the legislation (e.g. decree, regulation, decision)
date of the first version	date of the first official publication
date of the last version	date of the last amendment
Gesetzblatt year	year of the official publication
Gesetzblatt part	publication in part number X
Gesetzblatt page number	page number in official publication

Table 2: Element set for indicating bibliographic descriptions for German legislation

Next to the bibliographic information we also kept administrative data (as relates to the <profileDesc> element) and classification data (as is stated in the <encodingDesc> element) associated to each document.

4.3 Encoding of primary and linguistic data

For implementing the markup for the distinction of functional and typographic text divisions as described above, we chose to apply the <div> and <p> elements as introduced by the CES documentation. Annotation of linguistic elements was carried out to the level of marking sentence boundaries. For this level of annotation, the <s> and <title> elements were used. Alignment of multilingual documents was carried out automatically by employing the Alinea tool (Kraif, 2001), which provides for CES conformant output format.

4.3.1. Annotation of text divisions down to paragraph level

As the <div> element by the CES is defined to be used “to represent textual divisions of any kind” (CES, section 4.5.6.), it was employed for marking functional meaningful textual divisions. The <p> element was employed to mark any occurrence of a typographic paragraph.

We used the <div> element as a means to distinguish the different functional parts of legal documents. Overall we specified 8 types of functional divisions as defined in table 3:

<div>-type	**Description**
intro	introductory text including the main title
preamble	preamble
content	table of contents
main	main part of the document excluding preamble and annexes
chapter	division larger than legal paragraph that subdivides the main part of a text
section	legal paragraph
note	legal note at the end of the document
annex	annex

Table 3: Specification of functional divisions

Although most of the division types are shared by different legal systems, the composition of the elements can differ among legal systems. Most of the legal documents do not include tables of content, and also subdivision of text larger than legal paragraphs (e.g. ‘chapter’ or ‘part’) are only present in some legal systems. Figure 1 gives an extract of a document of the European law annotated down to the paragraph level (<p> element).

```
<div type="intro">
<p>
Richtlinie 2000/76/EG des Europäischen Parlaments
und des Rates
[...]
```

```
</p>
</div>
<div type="preamble">
<p>
DAS EUROPÄISCHE PARLAMENT UND DER
RAT DER EUROPÄISCHEN UNION -gestützt auf
den Vertrag
[...]
</p>
<p>
(1) Im fünften umweltpolitischen Aktions-
[...]
</p>
</div>
<div type="main">
<div type="section">
<p>
Artikel 1
Ziele
Diese Richtlinie bezweckt die Vermeidung
[...]
</div>
</div>
<div type="annex">
<p>
ANHANG I
Äquivalenzfaktoren für Dibenzo-p-Dioxine und
Dibenzofurane
[...]
</p>
</div>
```

Figure 1: Extract of annotated European directive

We put a stronger restriction on the allowed subparts of the <body> element than defined by the CES. According to our definition it could not contain anything but one or more <div> elements. Further we initially concentrated on an absolutely minimal version of the DTD for encoding primary data to keep the automatic annotation simple, and also since many of the sub elements to the <div> element as defined by the CES seemed not to occur in legal documents. An extract of the restricted DTD is shown in figure 2. The body of a text was only allowed to be composed of one or more <div> elements. Each <div> element should either hold <div> or <p> elements.

```
<!ELEMENT text (body)>
<!ELEMENT body (div)+>
<!ELEMENT div (div|p)+>
```

Figure 2: Extract of restricted DTD for encoding primary data

Eventually we planned to include <table>, <figure> and <list> elements as allowed elements within a <div> element.

4.3.2. Annotation of orthographic sentences

For annotations lower than the paragraph level we also added restrictions to the CES recommendations. We wanted to guarantee that all text is marked down to the sentence level, therefore <p> elements were defined to hold one or more <s> elements marking orthographic sentences. <s> elements could hold plain text or a <title> element to indicate that the marked sentence has the quality of being a title. Figure 3 gives an extract of the definition of paragraph and sub-paragraph elements.

```
<!ELEMENT p (s)+>
<!ELEMENT s (title|#PCDATA)>
<!ELEMENT title (#PCDATA)>
```

Figure 3: Extract of restricted DTD for encoding primary data

For each element the n-attribute to specify “a number or other label for the element” (CES documentation, section 4.5.1) was indicated. The n-value is a path composed of codes for all surrounding elements that the current element is included in. The second <s> element occurring in the third paragraph of a preamble for example would hold the n-value “body.preamble.p3.s2”. This information was included to guarantee easy access to the information about the text position of every text division and is displayed to the user when searching the corpus.

4.3.3. Annotation of alignment of bilingual text

The alignment of multilingual versions of documents as present for documents of the European and International law as well as for documents from Switzerland and the Alpine Convention were carried out automatically employing the Alinea (Kraif, 2001) tool. We decided to do a pairwise alignment for all language combinations. For each quadruple of documents we got six bilingual language pairs: German-French, German-Italian, German-Slovene, French-Italian, French-Slovene, Italian-Slovene.

The alignment was carried out on sentence level. Where major text divisions of documents were corresponding, the <div> elements were used as reference for pre-alignment. Unfortunately this was not always the case[8] and in cases of differing numbers of <div> elements pre-alignment could only be indicated for the entire text.

The Alinea tool provides for the alignment format conformant to the CES. Linking between text segments is indicated by linking ids of the <s> elements.

5. Processing and employing the encoded corpus

In the previous section we described the concrete implementation of encoding of documentation data, primary and linguistic data. In the following two sub-paragraphs we will briefly examine our experiences with processing CES-encoded corpus documents and integrating it in the query environment.

[8] Often multilingual documents differed in the number of annexes included and missing or existent tables of content.

5.1 Validating CES documents

As described in section 4.3.1 the main means for marking up the structural composition of legal documents is the employment of the <div> element distinguished by its different types (e.g. preamble or annex). Relying on the DTD for encoding primary data, it was neither possible to formally specify the allowed order of different <div> elements, nor to validate it. Since we had decided to employ the DTD instead of the XML-scheme for validation we had to temporarily transform <div> elements into more specific elements, according to their type (e.g. <div type=”preamble”> was temporarily transformed into <preamble>).

5.2 Integration into the querying environment

As mentioned before, the XML-encoded corpus documents were not thought to be directly used as basis for querying the corpus. Instead the institute’s environment for storing multilingual corpora should be employed. A detailed description of the employed database structure is given in Lyding et al. (2006).

Textual data from the XML-annotated corpus files had to be transformed and transferred to the database keeping the annotated structuring into functional divisions and paragraphs as well as sentence boundaries. Information about the position of specific segments within the entire document as recorded by the n-attribute (see section 4.3.2.) is also transferred to the database and provided to the user when querying the corpus. Two different user interfaces are provided for project partners and users of the general public. The internal user interface allows for searches on sentence and paragraph level. Searches above the paragraph level, although technically possible, were excluded for the moment due to performance issues. From the search results a direct link to the full text document is provided. Due to copyright considerations, we restricted searches through the public interfaces to the sentence level only. The full text document can not be retrieved directly, but via an automatically started Google search.

6. Conclusion

By systematically describing the use of the Corpus Encoding Standard for encoding a quadrilingual corpus to be employed in terminology work, we showed what questions arise when a concrete project aim meets the recommendations of a standard. We discussed the motivation for referring to a standard and for using it. Further we pointed out decisions that had to be taken prior to and also during the implementation of the encoding. By describing our experiences we wanted to relate the Corpus Encoding Standard to one very specific example of use.

7. References

Chiocchetti, E & Lyding, V. (2006). Multilingual Corpus for Terminology Work. In A. Abel, M. Stuflesser & M. Putz (Eds.), *Multilingualism across Europe: Findings, Needs, Best Practices.* Bolzano/Bozen : EURAC, pp. 505--513.

Corpus Encoding Standard – Documentation http://www.cs.vassar.edu/CES

Kraif, O. (2001). Exploitation des cognates dans les systèmes d’alignement bi-textuel: architecture et evaluation. *TAL,* 4(3), ATALA, Paris, pp. 833--867

Lyding, V., Chiocchetti, E., Sérasset, G. & Brunet-Manquat, F. (2006). The LexALP Information System: Term Bank and Corpus for Multilingual Legal Terminology Consolidated. Paper presented at the *MLRI-Multilingual Language Resources and Interoperability workshop*, *COLING/ACL2006,* Sydney.

Sérasset, G., Brunet-Manquat, F. & Chiocchetti, E. (to be published). Multilingual Legal Terminology on the Jibiki Platform: The LexALP Project. Paper presented at the *COLING/ACL2006,* Sydney

Streiter, O., Stuflesser, M. & Ties, I. (2004). CLE, an aligned, tri-lingual Ladin-Italian-German Corpus. Corpus Design and Interface. Workshop on *"First Steps for Language Documentation of Minority Languages : Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation".* LREC 2004, Lisbon

Application of terminology standards for a multilingual term bank: the EuroTermBank experience

Andrejs Vasiljevs, Signe Rirdance
Tilde, University of Latvia
Vienibas gatve 75a, Riga, LV-1004, Latvia
E-mail: andrejs@tilde.lv, signe.rirdance@tilde.lv

Abstract

This paper describes the experience of implementing applicable international standards in the area of consolidation of multilingual terminology resources. It identifies the specific requirements of the EuroTermBank project in consolidation of diverse terminology resources, merging of terminology entries across multiple resources, and federation of external interlinked term banks. It provides a brief review of applicable standards in the areas of terminology data modelling and resource consolidation, and analyzes how implementation of these standards has helped in reaching the goals of the EuroTermBank project. It also identifies limitations of current standards and provides a number of recommendations for further enhancements, focusing in particular on TBX (TermBase eXchange), as EuroTermBank implements TBX to satisfy a range of its specific requirements.

1. Introduction

Historically, most terminology resources have been developed within a rather narrow setting of an organization, a company, or an industry sector, very often related to translation needs. This has resulted in fragmentation of resources across terminology holders and limited availability of harmonized terminology data on the national and supranational levels (Henriksen, Povlsen, Vasiljevs, 2005). Despite the fact that international standards have been developed, wide proliferation of data models and technical formats, including proprietary ones, is a given, and adoption of existing standards and recommendations has been rather slow.

On the other hand, demand for creation of consolidated multilingual terminology resources is growing, both in the public sector, as governments are required to communicate with their citizens in more and more languages, and in the commercial sector, as companies move to communicate with their customers in multiple languages simultaneously across the globe. In order to respond to this demand, new standards-based models of terminology consolidation and distribution are required.

This paper examines the application of terminology standards in the EuroTermBank project, which deals with consolidation of dispersed multilingual terminology resources in a publicly available online term bank. It analyzes how application of existing standards and new approaches has helped in reaching the goals of the project, and identifies areas for further enhancement of these standards.

2. EuroTermBank project

EuroTermBank project is targeted at facilitation of terminology data accessibility and exchange at several levels. Its goal is collection, consolidation and dissemination of dispersed terminology resources through an online terminology data bank (Auksoriute et al, 2006).

The initial focus of the EuroTermBank was to contribute to the improvement of the terminology infrastructure in selected new European Union member countries (Latvia, Lithuania, Estonia, Poland, Hungary); however, EuroTermBank continues to expand its activities to other countries in EU and beyond. This aim is accomplished by establishing cooperative networks of terminology institutions on various levels and by consolidation and harmonization of existing terminology resources.

Currently, EuroTermBank contains 1 918 886 terms in 30 languages, with additional resources available through interlinked data banks.

EuroTermBank enables the exchange of terminology data with existing national and EU terminology databases by establishing cooperative relationships, aligning methodologies and standards, designing and implementing data exchange mechanisms and procedures.

Through harmonization, collection and dissemination of public terminology resources, EuroTermBank is aimed to facilitate enhancement of public sector information and strengthen the linguistic infrastructure in the new EU member countries.

Development, population and maintenance of a web-based terminology data bank constitute the major tangible outcome of the project. The data bank works on a two-tier principle – as a central database and as an interlink node or a gateway to other national and international terminology banks.

Hence, EuroTermBank project deals with the issue of terminology exchange formats within the terminology collections included in the EuroTermBank database as well as among a federated group of independently maintained online terminology banks that are interlinked to EuroTermBank.

Within EuroTermBank project, significant preparation work was done in assessing the relevant international standards for terminology (Auksoriute et al. 2006). With the multitude of actors involved, including project partners and various types of terminology holders, it was clear that only implementation of applicable international standards can ensure reaching the goals of the project. This includes a standards-based approach to describing terminology collections, defining the data model and ensuring a unified data exchange format. Hence EuroTermBank has in fact been a testing ground for applying a number of related international standards in the realm of terminology.

3. Requirements for EuroTermBank in terminology data modelling and consolidation

Taken the above, it is clear that EuroTermBank project poses a set of requirements towards applicable terminology standards and can therefore serve as a test of how well these standards support real-life scenarios in consolidation of diverse multilingual terminology resources.

3.1. Consolidation of diverse source formats and data structures

In consolidation of diverse terminology collections, a number of problems arise that relate to terminology exchange formats. The most obvious problem relates to the diversity of formats in which terminology collections were provided for inclusion in EuroTermBank. Quite a few of the terminology resources initially identified for inclusion in EuroTermBank were available only in printed form. Other resources were in different electronic formats, including Word documents, Excel spreadsheets, plain text files, HTML pages and PDF files; some were provided in highly structured XML. A few recent collections came in Trados MultiTerm and TBX formats (Liedskalnins, Rirdance, Vasiljevs, 2008).

A further complication is the diverse terminology data structure across terminology collections. A number of collections contain very few basic data categories; others are much more elaborate. However, most difficulties arise from the fact that data categories have not been uniformly named or defined across different collections. Furthermore, huge terminology resources typically contain a certain amount of incorrect data, or some data may be missing and incomplete.

Hence, the main challenges addressed within the EuroTermBank project have been:

- definition of a data model and corresponding data categories to ensure full coverage, flexibility and exchangeability;
- transformation and storage of the initially diverse resources in a single unified format, to enable unified access and representation of terminology data to the user, regardless of the specifics of the source collection and regardless whether the data come from EuroTermBank stored resources or from external interlinked term banks.

Other priorities identified for the project include preservation of full information available in the source data and ensuring optimal performance of the system.

3.2. Consolidation of terminology entries across collections

When consolidating a large number of terminology resources, with several collections belonging to the same subject field classification, one can assume a certain amount of fully or partially overlapping terminology entries in various languages that can be mapped to the same overarching concept. In practice, displaying a long list of identical search results significantly worsens user's experience. Furthermore, there are benefits to be gained by collating term entries that designate the same concept from bilingual and multilingual terminology collections representing diverse languages.

For example, if there is a term pair EN *tree* – LV *koks* coming from a Latvian IT terminology resource and another term pair EN *tree* – LT *medis* from a Lithuanian IT terminology resource, there is an obvious interest to join these two into unified entry EN *tree* – LV *koks* – LT *medis*. Such multilingual entry allows establishing a certain correspondence between language terms that is not directly available in any terminology resource (in our example, the new term pair LV *koks* – LT *medis*).

However, merging entries just on the basis of a matching term in one language that is common for these entries will lead to many erroneous term correspondences, due to the frequent ambiguity of terms among subject fields or rarer cases of ambiguity in the context within one subject field. The cost of manual evaluation by a terminologist whether the given entries do denote the same concept is prohibitive, taken large databases like EuroTermBank that contain about 1.9 million terms. Therefore, EuroTermBank applies a practical solution by introducing terminology entry compounding, which is an automated approach for matching terminology entries based on available data. This mechanism is a considerable aid for the professionals of the terminology field in identifying potentially matching entries across diverse term collections.

3.3. Federation of external interlinked term banks

Even this day, terminology resources on the internet remain fragmented across diverse term banks and terminology projects. A number of user scenarios require consolidation on a multilingual and multinational scale, therefore EuroTermBank not only stores all available terminology content in its database, but also acts as a gateway providing unified access to multiple remote terminology databases. Such federation of term banks is a new concept in linking portals and data repositories, and it should go far beyond the establishment of pointers or links, towards the level of exchangeability and semantic interoperability of data and data structures (Galinski, 2007).

To ensure the viability of the federated system of terminology databases, inclusion of a term bank in this federated model requires it to be independently supported and maintained both institutionally and technically.

Within EuroTermBank, the mechanism that enables federation of external databases is called interlinking. Interlinking an external database to EuroTermBank enables users to query the external database from EuroTermBank web interface.

4. Application of standards for scenario based data modelling

To ensure exchangeability and facilitate recognition and comprehension of data categories for new or outside users terminology data should be modeled based on ISO 12200:1999 and 12620:1999 standards. Principles of these ISO standards require that the term entries are concept oriented, contain a rather broad selection of data categories that permits the necessary level of detail and permit full descriptions of each term.

However, these standards are very extensive and general and there is a strong need for guidelines on how to apply them in particular usage context or applications.

To better understand existing practices and user needs for terminology systems, we carried out an interviews of different institutions involved in terminology work in 7 countries. In addition, 51 questionnaires were filled in, providing an overview of possible usage cases and corresponding user requirements for particular cases.

As a result of this analysis, we distinguished 3 levels or scenarios of terminology work – local, national and international (Henriksen et al. 2005).

4.1. Data modeling in local scenario

Within the local scenario, main conditions and goals that are important for the design of a data structure are: tight time frames, translation-oriented needs, exchangeability, and limitation of terminology work to one or a few domains. These criteria speak in favour of a highly customized and only moderately exhaustive data structure where data categories are consistent with the requirements of the particular application area and have a translation related focus.

A focus on translation requirements implies coverage of more than one language. It must therefore be considered whether such descriptive concept related information as definitions or explanations are necessary for each language or only for one language. If the term collection is multilingual a definition for each language is usually necessary. If the term collection is only bilingual, it may not be necessary.

A focus on translation requirements also indicates inclusion of data categories permitting sufficient information about the use of a term, for example, different types of grammar information, context information and collocation information. Some translation settings may also require grammar information for each word of a term. Furthermore, it is often considered very important to document the degree of equivalence between terms of different languages. Data categories that could be relevant in this respect are, for example, false friend, directionality and transfer comment.

The below data structure containing four levels reflects a multilingual terminology setting permitting, for example, concept descriptive information for each language and grammar information for each word. In multilingual as well as bilingual terminology settings it can however be considered to omit the word level and locate grammar information at the term level instead. In some bilingual terminology settings it can also be considered to have a definition for only one language. Consequently, the data structure in a bilingual framework may include only 2 levels, namely, concept and term levels.

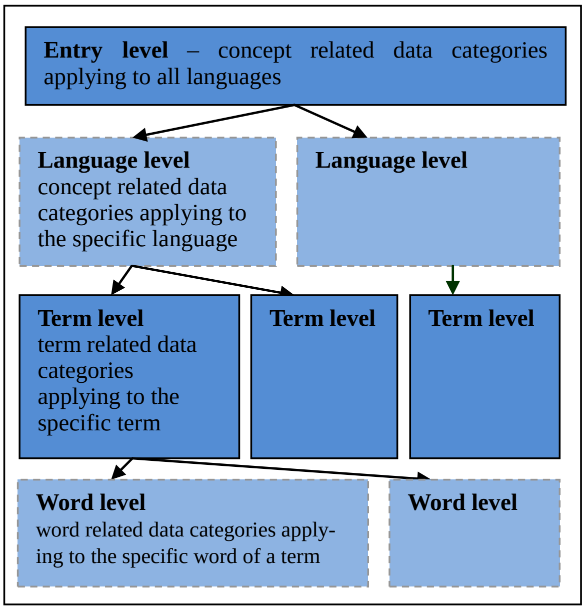

Entry level – concept related data categories applying to all languages
Language level concept related data categories applying to the specific language
Language level
Term level term related data categories applying to the specific term
Term level
Term level
Word level word related data categories applying to the specific word of a term
Word level

Figure 1 Four-level structure for terminology data

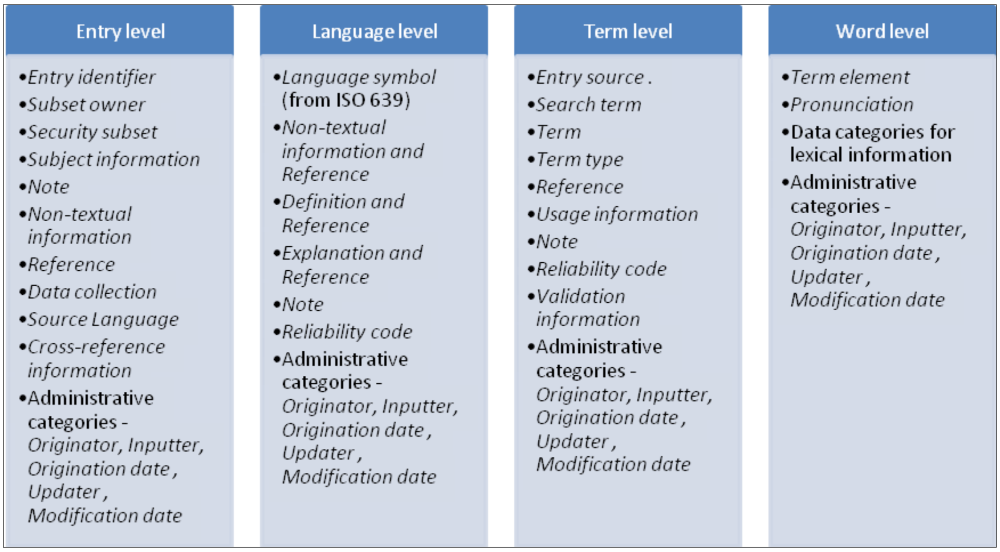
4.2. Data modeling in national scenario

In the national scenario, conditions and goals influencing the design of a data structure are adequate financial support, exchangeability, broad domain coverage and high quality in general terms. Besides, a national term collection is aimed at terminology coordination and regulation rather than at translation. These criteria point towards a data structure that permits an exhaustive selection of data categories covering very different user requirements and enabling users to develop entries for very different purposes and of a very high quality.

This implies that the data structure should often contain 2 levels: concept and term levels (at least when the term collection is monolingual) and that data categories should represent a wide selection of information types and include term status qualifiers reflecting for example acceptability, approval or applicability of a term in a given context. An example of a term status qualifier is normative authorization which is assigned by an authoritative body and includes qualifiers as standardized term, preferred term, admitted term and deprecated term.

4.3. Data modeling in international scenario

Within the international scenario, the criteria considered important are very similar to those important in a national scenario. A crucial difference is however that an international terminology cooperation is multilingual by nature. Therefore it is recommended that the data structure should include four levels permitting concept descriptive information for each language and grammar information for each word of a term.



Entry level
- *Entry identifier*
- *Subset owner*
- *Security subset*
- *Subject information*
- *Note*
- *Non-textual information*
- *Reference*
- *Data collection*
- *Source Language*
- *Cross-reference information*
- Administrative categories - *Originator, Inputter, Origination date , Updater , Modification date*

Language level
- *Language symbol* (from ISO 639)
- *Non-textual information and Reference*
- *Definition and Reference*
- *Explanation and Reference*
- *Note*
- *Reliability code*
- Administrative categories - *Originator, Inputter, Origination date , Updater , Modification date*

Term level
- *Entry source .*
- *Search term*
- *Term*
- *Term type*
- *Reference*
- *Usage information*
- *Note*
- *Reliability code*
- *Validation information*
- Administrative categories - *Originator, Inputter, Origination date , Updater , Modification date*

Word level
- *Term element*
- *Pronunciation*
- Data categories for lexical information
- Administrative categories - *Originator, Inputter, Origination date , Updater , Modification date*

Figure 2 EuroTermBank data model based on ISO 12620 data categories

Since EuroTermBank represents international scenario of terminology work its data structure is modeled in 4 levels. See Figure 3 for ISO 12620 data categories selected for EuroTermBank based on user needs and resource analysis.

5. Application of standards in consolidation of EuroTermBank terminology resources

ISO/TC 37 has developed a number of standards in the field of terminology (Auksoriute et al, 2006). Of these, ISO 12620:1999 specifies data categories for recording terminological information in both computerized and non-computerized environments and for the interchange and retrieval of terminological information. ISO 12200:1999 specifies the MARTIF interchange format. However, these two standards only allow for negotiated interchange and are not strict enough for a specific interchange scenario without additional agreements. ISO 16642:2003 is related to the terminology markup framework TMF enabling to specify interoperable markup languages on the basis a common meta model. Therefore, TMF is not a terminology interchange format in itself, but MARTIF is such a TMF-compatible markup language.

LISA, the Localization Industry Standards Association, has adopted TBX, short for TermBase eXchange (www.lisa.org/standards/tbx), which is an open XML-based standard format for terminological data, which is a practical terminology exchange format that is compliant with the terminology markup framework TMF. TBX is based on the TMF structural meta-model; it specifies a set of data categories from ISO 12620 and adopts an XML style compatible with ISO 12200. TBX assumes a concept-oriented approach, which implies that the terms in a bi- or multilingual entry are synonymous unless otherwise noted.

All acceptable variations on TBX have the same core structure. They differ mainly with respect to the data categories from ISO 12620 that are allowed by a particular user group.

TBX is in a process of adoption as an ISO standard by 2009.

5.1. EuroTermBank implementation of TBX

EuroTermBank system implements the TBX standard to satisfy a number of requirements: enabling data exchange between different ETB modules, interoperability with external databases, data import/export, and data storage in the EuroTermBank internal database.

A list of required terminological data categories was created during the EuroTermBank project based on best practice research. Selected data categories were compared to data categories specified in ISO 12620 to verify their compatibility. As TBX standard defines XML-based format, it was possible to use only the required data categories and still be compatible with TBX standard.

Although TBX standard is mainly devised as an exchange format, in EuroTermBank it is also used for terminological data storage in the database, as terminology has specific characteristics that make it difficult to store such type of data:

- it has many optional data categories;
- data categories frequently have no format restrictions;
- size of some data categories is not predictable.

These problems were solved in EuroTermBank by storing data in the XML-based format defined in the TBX standard. This provided the following benefits:

- storage of all TBX data categories;
- no format and size limitations for data categories;
- simple extensibility.

TBX standard is used also for data import and export to and from EuroTermBank database. All resources to be included in the portal's internal database are converted to TBX format. Source formats vary from printed paper resources to highly structured XML files. As TBX is also the storage format, there are no significant reasons for introducing another format. As TBX allows storage of all standardized data categories, it is possible to convert all resources to TBX format. Even if resources have resource specific data categories that are not included in the standard, it is possible to store these categories as supplemented XML tags without changing the physical data storage model.

TBX format is applied throughout the EuroTermBank system. Since TBX format is used through all the resource life-cycle stages, it also ensures data consistency. Using an open and non-proprietary standard is appropriate not only for EuroTermBank resource interoperability within the internal system, but also for communicating globally with external terminology databases. EuroTermBank system is designed to provide external systems with standardized data in TBX format and receive data from external systems in the very same way. There is no need to define a new framework either for processing every single external data provider or for the data provided by the system.

In EuroTermBank system, the TBX standard enables data storage of all four terminological concept levels – entry level, language level, term level and word level (Schmitz, Vasiljevs, 2006). It also supports all data categories identified during the best practice research. All of 92 resources imported in EuroTermBank have been converted to TBX format without data loss, ensuring not only standard compliance, but also extensibility of the format.

Using the TBX standard throughout the system provides data consistency as data are not converted either in the system's internal modules or in the communications with external systems. From external systems that are already connected to the EuroTermBank system, one is directly providing data in TBX format. Other systems use proprietary exchange formats so conversion to TBX is applied before passing data to EuroTermBank. Furthermore, there are several systems that are on the way to use EuroTermBank system as the data source for terminology and communicate in the TBX standardized format.

5.2. Limitations of TBX

With its strength in terminological data storage and exchange, it is also true that TBX possesses certain limitations and weaknesses. TBX falls short of ensuring blind interchange between any given implementations, since it provides ample freedom, for example, in application of data categories. Thus some data categories may be required in one term bank, while optional or not present in another one, or one and the same data category may appear on different levels of the entry structure. Although TBX is not intended to ensure blind interchange, this limitation hampers its wider implementation.

Therefore an important step forward is development of TBX-Basic, a lightweight version of TBX that identifies a limited set of data categories, including a minimum set of mandatory categories (www.lisa.org/sigs/terminology). It is meant to satisfy the requirements for small or medium sized language industries and will be included in TBX as an appendix demonstrating an example of a TML (Terminology Markup Language) that is compatible with TBX.

TBX is also criticized for its concept-based multi-linguality and non-directionality, stating that TBX does not cover terminology in areas that are subject to societal or cultural influences and where there is no concept with synonymous terms in many languages (Thurmair, 2006). Thurmair concludes that TBX is only suitable for the representation of technical terms where a 1:1 correspondence between participating languages can be assumed.

In response to this criticism we should take into account that TBX does provide for language-specific descriptions of concepts using definition, comment, context or other text field. In cases where 1:1 correspondence is not present, a new concept with either only one or a limited set of languages can be defined. While it is true that TBX is not suited for exchange of machine translation dictionaries that contain a large number of general vocabulary terms, this is not the purpose of TBX. As shown by EuroTermBank experience, TBX does serve as a practical and highly usable exchange format for a number of terminology exchange scenarios.

The concept of terminology exchange becomes relevant and important in scenarios involving merging or exchange between several terminology resources or collections, which involves collating or merging term entries across collections as described in this article. Despite this being a major scenario in terminology exchange, there is no straightforward way in TBX for creating relations between terminological entries from different resources. Although technically it is possible, it is not part of the standard. The situation in creating relations between single resource entries is better, with a few types of relations – broader, generic and related – explicitly defined within the standard. However, these relations are limited and would be insufficient for creating more complex ontology structures.

6. Conclusions

Exchange formats are becoming increasingly important in the area of terminology management, be it for multilingual term banks like EuroTermBank or business scenarios where company mergers necessitate merging of several diverse multilingual terminology databases. At the same time, it has to be recognized that the slow and partial implementation of international terminology standards is a limiting factor for providing sufficient feedback for extensions and enhancements of these standards.

This paper has demonstrated applicability of terminology standards in multilingual terminology resource consolidation. Guidelines for data modelling in different scenarios and application of TBX standard for data exchange can serve for better adaptation of standards in practical applications.

7. Acknowledgements

Many thanks to colleagues in all EuroTermBank project partner organizations: Tilde (Latvia), Institute for Information Management at Cologne University of Applied Science (Germany), Centre for Language Technology at University of Copenhagen (Denmark), Institute of Lithuanian Language (Lithuania), Terminology Commission of Latvian Academy of Science (Latvia), MorphoLogic (Hungary), University of Tartu (Estonia), Information Processing Centre (Poland). EuroTermBank Consortium would also like to acknowledge and thank the European Union eContent program for supporting the EuroTermBank project as well as support from EU Social Fund.

8. References

Auksoriute, A. et al (2006). *Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project.* Riga, Tilde.

Galinski, C. (2007). New ideas on how to support terminology standardization projects. *eDITion*, 1/2007.

Henriksen, L., Povlsen, C., Vasiljevs, A. (2005). EuroTermBank – a Terminology Resource based on Best Practice, in *Proceedings of LREC 2006, the 5th International Conference on Language Resources and Evaluation*, Genoa, on CD-ROM.

Liedskalnins, A., Rirdance, S., Vasiljevs, A. (2008). From Paper to TBX: Processing Diverse Data Formats for Multilingual Term Bank, in *Proceedings of the Third Baltic Conference on Human Language Technologies*, Kaunas, Lithuania.

Schmitz, K.D., Vasiljevs, A. (2006). Collection, harmonization and dissemination of dispersed multilingual terminology resources in online terminology databank, in *Proceedings of TSTT 2006, Third International Conference on Terminology, Standardization and Technology Transfer*, Beijing.

Thurmair, G. (2006). *Exchange Formats: TBX, OLIF and beyond*. LDV-Forum 21,1, pp. 45-56.

Wright, S.E. (2005). A Guide to Terminological Data Categories – Extracting the Essentials from the Maze. In *Proceedings of TKE 2005, the 7th International Conference on Terminology and Knowledge Engineering*. Copenhagen, pp. 63-77.

List of Authors