[*Translating and the Computer 24*: proceedings of the…International Conference…21-22 November 2002, London (Aslib, 2002)]

Knowledge Exchange and Terminology Interchange: The role of standards

Lee Gillam[α], Khurshid Ahmad[α], David Dalby[β], Christopher Cox[χ]
[α]Dept of Computing, University of Surrey: l.gillam@surrey.ac.uk
[β]Linguasphere Observatory: dalby@linguasphere.org
[χ]BSI, UK: Christopher.Cox@bsi-global.com

Abstract

The emergence of standards for storing and retrieving language resources, including terminology and lexicographical data, documents and text corpora, will benefit system developers and users of a range of language and knowledge engineering systems. The developers will be able to cope better with the vagaries of natural language since standardised entries in term databases, or structured documents in a text corpus, reduce the variations encountered in any language resource. The users will benefit because they will be able to take advantage of data produced from a greater number of data producers since standardisation filters out minor and arbitrary variations in which the data is stored which can confound many a current language engineering system. We are concerned with standards that specify markup for terms, words, and documents, and how data can be marked-up and decoded automatically. Much of the discussion is based on the results of the recently completed EU-sponsored project Standards-based Access to Lexicographical and Terminological multilingual resources (SALT - IST-1999-10951), and the recent deliberations within the International Organization for Standardization's Technical Committee 37 (TC37) on standards for terminology and other language resources, specifically ISO 12620 and ISO 16642.

Introduction

It can be argued that the *Semantic Web* (Berners-Lee, 1999), *WordNet* (Miller et al, 1993; Fellbaum, 1998) and its successor *EuroWordNet,* have benefited in part from the original work of the International Federation of the National Standardizing Associations (ISA), specifically the Technical Committee ISA/TC 37 "Terminology", founded in 1936. The ISA committee, championed by Austrian engineer Eugen Wüster, was amongst the first international projects to collect, validate, store and disseminate terminology. This was a concept-oriented prescriptive approach to terminology. Two recently completed International Standards from the International Organization for Standardization (ISO[1]) in computer-oriented terminology owe much to Eugen Wüster and his followers. The first of these standards, ISO 12620, contains an inventory of consensually defined categories of information and their possible values, associated with a given term or cluster of terms. The second, the forthcoming ISO 16642, defines a framework within which these data categories can be used more effectively.

[1] ISO is not an acronym, it is the Greek word for "equal" adopted by the body that succeeded ISA; the capitalization tends to add to the confusion.

The data categories, over 170 in the case of ISO 12620, are an exhaustive list of attributes and possible data associated with a term. ISO 12620 provides the basis - names and definitions - for understanding aspects of the content. Typically a term has associated semantic information e.g. *gender, number,* semantic relations of *hyponymy* and *meronymy, definitions* and in specifications of *concepts* to which the term is associated. Additionally, a term has associated pragmatic information including *context* and *register.* The semantic and pragmatic information is usually encoded as an attribute-value pair embodied in a *table* of attributes and values. An attribute could be labelled as the *name* of a term and the value would be the term itself; an attribute could be 'grammatical category' and take one of the values 'noun/verb/adjective'. These attribute-value pairs are generally organised in a tabular structure, which is common to many a terminology management system (TMS). However, it is not always a simple task to transfer such data from one TMS to another because there are no controls on names of attributes and on the values these attributes can take. If the attributes and values were clearly identified and the labels of attributes specified in accordance with a common reference system, such as ISO 12620, the interchange of data between two TMS's will be relatively easier. Related standards such as ISO 639 - *Codes for the representation of the names of languages* - can be referred to for the systematic naming of languages ('en' for English; 'fr' for French), that may occur as values of the term attribute *language,* in which the term originates or is used.

The tabular organisation of terminology databases depends on the underlying data model employed. The cardinality of these relations - how many times a certain attribute occurs in relation to a specific item of data, for example, gender occurring once, but many contextual examples - determines the table structure required to represent a specific terminology collection. In relational databases, one-to-many relations are generally modelled using keys from one table to another. In the worst case, every attribute could be associated many times to a single piece of data, which would necessitate a table for each set of attribute-value pairs. ISO 16642 outlines an implementation independent structure of terminology collections, presenting as it does a terminological metamodel. A specific implementation of this metamodel is called a *terminological mark up language* (TML). The specification of a TML can be used to define a variety of existing industry and international terminology standards such as MARTIF (ISO 12200) and GENETER[2]. Collections sharing this metamodel can be shown to be interoperable, that is, they have the potential for data interchange to be possible on a structural basis. Such discussions are central to debates on knowledge and ontology in the literature on semantic web and other knowledge-based enterprises.

ISO 16642 and ISO 12620, when combined, can be used to specify and document the creation of high-quality terminological resources for enterprise and governmental uses alike. By relating existing terminology collections to this pair of standards, essentially by conforming to the specification of a TML, employing common sets of attributes and values, and referring to other existing standards such as ISO 639 within this specification, one has a framework for the potential re-use of existing terminology data collections. Conforming to these standards can prove potentially beneficial in terms of communications throughout the enterprise, especially in cases

[2] see for example http://www.uhb.fr/Langues/Craie/balneo/demo geneter.pl?langue=2 last visited 2 October 2002

where different elements of the enterprise make use of the same information in different ways. In the longer term, this combination can be used to underpin knowledge management activities where a terminology collection acts as a repository of knowledge in current use. This knowledge, expressed through commonly used terms, can be measured for signs of growth and decay of the used concepts, for example in less frequent systematic use of once popular terms, and coinage of neologisms may indicate new knowledge, requiring the formation of new concepts.

Standards are voluntary; without legal or technical instruments of systematically encouraging the use of standards, they remain purely voluntary, whether written prescriptively or as recommend codes of best practice. The decision, for example, by a company to become registered through third party certification to the ISO 9001 requirements for quality management systems may often be the result of a demand from a customer that the company cannot afford to lose.. The technical instruments for encouraging the use of standards in, say, terminology management, include programs that can decode the mark-up language in which data related to individual terms in terminology collections are marked up. The technical instruments should also facilitate the encoding process by providing a program, with a good user interface, that will enable a terminologist to add terms to an existing collection or build a terminology collection from scratch, without necessarily being aware of the underlying implementation, while being assured that the terms will be usable elsewhere.

Developing and institutionalising standards for terminology

Standards and terminology relate to each other in three ways: there is the terminology of standards, there are standards of terminology and there are standards for the management of terminology. The purpose of these three relationships is variously the provision of a shared understanding within, and of, text-based resources, in a variety of human languages. Standards appear in many forms, all of which are drafted by committees of experts: Industry standards such as IEE 802.11 for wireless networking, international standards such ISO 8879 for the Standard Generalized Markup Language (SGML, see also Goldfarb, 1990), the precursor to the eXtensible Markup Language (XML, Bray et al, 2000), and ISO 9000 series of International Standards for Quality management systems. There are national standards, for example the withdrawn BS 5750 on Quality assurance systems, the precursor to ISO 9000 which has progressed to quality management systems, and the ANSI standard for the C programming language. The W3C produces specifications that it calls recommendations and makes them available free of charge. The page-based pricing policy currently practiced by some standards organisations is subject to some considerable debate, especially in an increasingly electronic marketplace.

Standards documents tend to contain terminology arrived at by a consensus of the members involved in their development, for use within the standard, possibly, in a series of standards. This terminology may be relevant to other documents also, and can be cross-referenced. In some cases, specific subject fields, such as Aeronautics,

have a need for a standard reference set of terminology. The purpose of these standard terminologies is to provide consistency and coherence of reference by increasing transparency and accuracy, and reducing ambiguity in the description of specific items, events or relationships. Standards documents available from BSI Standards Online[3] include 277 titles containing the word Terminology', 547 containing the word 'Glossary' and 267 results for 'Vocabulary'. Such documents may contain terms in more than one language. The importance of these standards in safety-critical enterprises cannot be understated.

ISO Technical Committee on Terminology and Other Language resources' (ISO TC 37)

ISO TC 37, drafts, maintains and revises standards that provide rules and procedures for terminology work. These standards are appropriate in many standardisation environments to enable compatible tools and systems to be developed for communication purposes. The development of common reference sets and the knowledge that specific resources conform to these sets provides a basis for interoperability amongst tools and data. TC 37 has 4 subcommittees (SC):

- Principles and methods SC 1 (ISO/TC 37/SC 1),
- Terminography and Lexicography SC 2 (ISO/TC 37/SC 2),
- Computer applications for terminology SC 3 (ISO/TC 37/SC 3),
- Language resource management SC 4 (ISO/TC 37/SC 4).

(The British Standards Institution (BSI) has TS/1, UK's shadow group to ISO/TC 37)

The SCs work in a symbiotic manner: they are independent of each other, but there is a substantial scope of using, and influencing, the principal results of the deliberations of the SCs - the *standards.* For example SC 1 and SC 3 have published standards on 'Terminology Work': SC 1 has published three key standards on *Principles and Methods of Terminology Work* (ISO 704, published in 2000), on *Harmonization of concepts and terms* (ISO 860 in 1996) and part one of a multi-part standard on *Vocabulary* (ISO 1087, Part 1 in 2000). The standard on Vocabulary was taken further by SC 3 which published *Part 2* focussing on *Computer applications* (ISO 1087, Part 2 in 2000). SC 3 has published standards the exchange of terminology, focusing on the medium of exchange - the now historically titled *Magnetic tape exchange format for terminological / lexicographical records* (ISO 6156 published in 1987) and the more recent standard on *Machine Readable Terminology Interchange Format (MARTIF) - negotiated interchange* (ISO 12200 published in 1999). The purpose of this set of standards is to provide a systemic and systematic set of specifications, guidelines and so on, for work within terminology and the "other language resources". (Annex A of this paper comprises a list of some of the ISO TC 37's work).

The principal standards we are concerned with in this paper are ISO 12620:1999, currently under revision in line with ISO11179 (ISO/IEC 11179:1994), specifically Part 3 of this standard concerned with *Basic Attributes of Data Elements,* and ISO

[3] see links from http://www.bsi-global.com/index.xalter - last visited 1 October 2002

16642 - Terminological Markup Framework - which will be published in the near future (c. 2003).

Structure of Terminology Collections

A terminology collection is a collection of terms, along with identifiers and resources that relate to those terms. A terminology collection variously comprises strings, values, codes, and associations and relations, as well *coded knowledge fragments,* for example, association to a large-scale classification system. Such fragments may use proprietary or public identifiers, differing data structures and proprietary or public classification systems. Terminological collections contain terms, their synonyms, abbreviations, concepts and numerous other items. The attributes used to present these items are as important as the item being presented as they give purpose to whatever is contained. Synonyms may, for example, by variously identified by an attribute labelled *syn* or *synonym,* or may exist as a term at the same level in a concept entry. Such data modelling variance needs to be understood. A number of data models for terminology have been developed, including the TRANSTERM[4] model (see Figure 1 below) and the model that underlies ISO 12200 (MARTIF). These models share a number of common items. At an abstract level, they contain concepts with terms relating to these concepts, definitions, identifiers for particular languages and so on. The abstraction enables the potential for an interchange language that can make various resources interoperable. Here, we are not concerned with the details or limitations of specific database models.

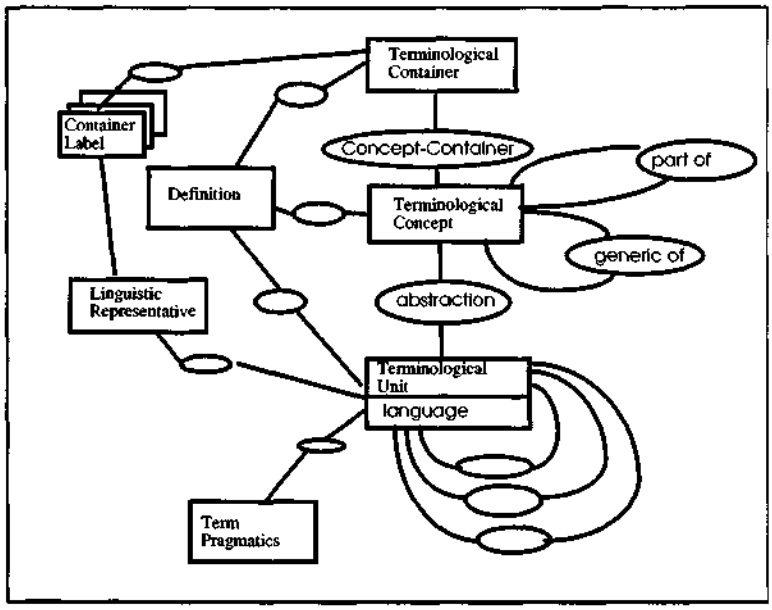

Figure 1: TRANSTERM project's elaborated data model

[4] The TRANSTERM project investigated the connection between terminology data bases and machine-readable lexica. It investigated GENELEX, CEGLEX, PAROLE and other such models. The TRANSTERM model was created in conformity with GENELEX.

A Metamodel for Terminology?

The literature on terminology management system frequently uses the term 'data' for both the values of the attributes of a term and to describe a set of terms. In computing, the term *object* is frequently used, loosely meaning the object of study, that could represent entries in a database, the database itself or the users of the database. Typically, in the TMS literature, the *object* is a term. The data values associated with the term are its (linguistic) attributes. A collection of objects, again in computing terminology, is described as a model or schema; in a TMS, a schema can be compared with the data model used to design and implement a terminology data base system. The description of a number of schemas (or schemata to be more precise) is called a metamodel (how these metamodels are described is called a meta-metamodel). There is no current equivalent of metamodels in the TMS literature, however the forthcoming international standard ISO 16642 provides a description of such a metamodel as used in terminology management. The terminological metamodel (Figure 2) comprises a number of 'containers' or 'sections' into which data about the terms can be stored. For example, a Language Section is used to separate the terms in one language from other languages. The Term Section contains information about the term being described, such as contextual or source information. The metamodel in turn, has been described, or instantiated, with a model (meta-metamodel) developed using the Unified Modelling Language (UML) that decomposes the metamodel into containers and relations and cardinalities between containers (N.B. A tree-structure is not the only possible metamodel for language resources). The terminological metamodel has been developed based on the Methods and Principles of Terminology Work (ISO 704). To demonstrate the effectiveness of the metamodel, the SALT project team developed a program that can map between terminology collections using an implementation independent XML application called the Generic Mapping Tool (GMT), which is presented in ISO 16642.

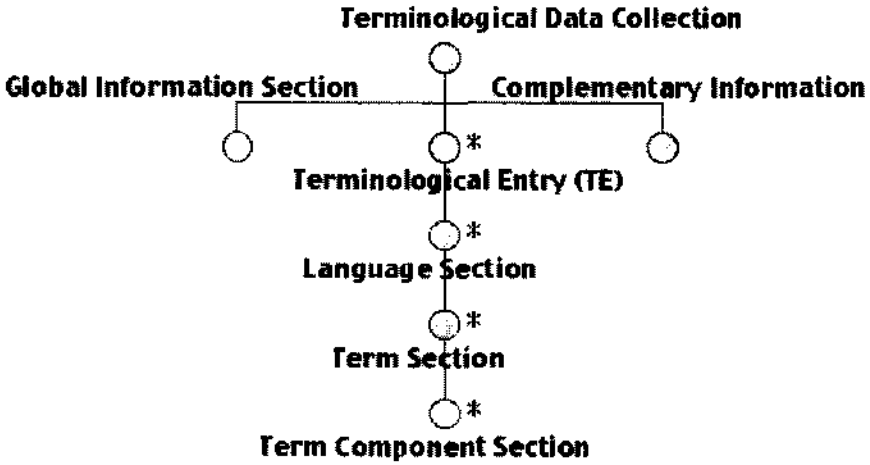

Figure 2: ISO 16642 Terminological Metamodel

Terminological Data Categories

There is a tendency to develop language-based resources in an ad hoc manner, frequently to expedite the functioning of an enterprise. The ad hoc development of these resources leads to the ad hoc use of data, including attributes, values and relations that can cause considerable difficulties in the inter-operability of terminology databases, and can result in the failure of terminology management systems. One item of this data - attribute, value and relation - can be referred to as a Data Category (DC). For example, a *synonym* is a relation between two terms or concepts, which may be realised in a number of ways, providing the realisation implements the semantics.

To keep effective track of Data Categories, a registry is to be proposed in Part 1 of the revised ISO 12620 which will take care of the collection and maintenance of these Data Categories, and provide a means for managing/maintaining, updating and revising the collected DCs more rapidly than is possible through the publication and subsequent revision of standards. This registry will be an implementation of a metadata registry, as specified in 11179-3, realised through the description of so-called Data Elements (see Figure 3 below for attributes used to describe a data element).

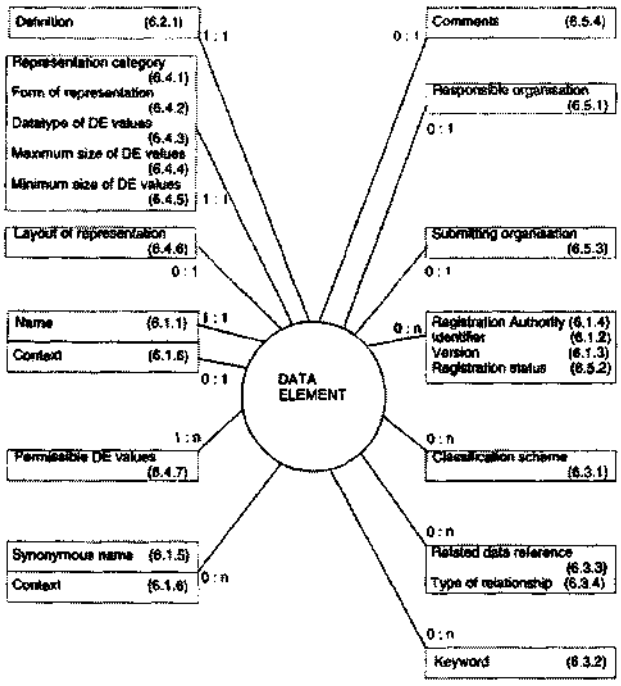

Figure 3: Attributes describing a Data Element, from ISO 11179-3

For the purposes of Data Categories, this description of a Data Element is used to outline the attributes that are filled in for each Data Category (DC). Each DC has

attributes, for which the description are the *Data Elements,* that identify and define it and may provide useful examples and tips on usage. For example, consider the Data Category: *term* (see Table 1). The *term* DC has a number of Data Elements, which are named on the left-hand side of the table. **DC Name** is the name of a Data Element that must be filled in for a Data Category. **Example** is another Data Element for which the Data Category may or may not fill the value. Each of these Data Elements has a Name and a Definition, for which filler values are mandatory. Filling other Data Elements is either conditional (if other values are filled) or optional. Part 2 of ISO 12620 will contain the description of the terminological data categories as at least, **DC Name** and **DC Definition.** For related information, the DCR will be available for registration, monitoring, further information and usage details of the Data Categories. Subsequent parts of this standard are envisaged for, for example, lexicographical DCs such as items used within the Open Lexicon Interchange Format (OLIF) and even DCs for the language identifiers.

term	
DC ID	ISO12620A-01
DC Name	term
DC Definition	A verbal designation of a general concept in a specific subject field.
DC Source Comment	For definition of related term, see ISO 1087-1, 3.4.3Source
Concept-related Comment	Terms can consist of single words or be composed of multiword strings. The distinguishing characteristic of a term is that it is assigned to a single concept, as opposed to a phraseological unit, which combines more than one concept in a lexicalized fashion to express complex situations. Quality assurance system is a term, whereas satisfy quality requirements is a phraseological unit, specifically a collocation.
Example	"radıx" ın annex D, figure D.1.
Data Type	noteText (open)
Level(s)	Term Section, Term Component Section

Table 1: Description of the term Data Category

ISO 12620:1999 contains about 170 data categories (growing and shrinking in revision) that have been and are being described in this fashion, organised into the following groups

- term
- term-related information
- equivalence
- subject field
- concept-related description
- concept relation
- conceptual structures
- note
- documentary language
- administrative information

ISO 12620 could be considered as a terminology of terminology (metaterminology), which would seem to make ISO11179-3 a terminology of terminology of terminology (meta-metaterminology)

The description of these DCs, the values that are filled in for the attributes described by Data Elements, is made using the Resource Description Framework (RDF, Lassila and Swick, 1999), and can be carried out using the SALT suite developed in the SALT project. Part of the resulting file of DCs (the reference file) is shown below in the main area of the SALT Suite (Figure 4). The interface for defining such a Data Category is shown in Figure 5 using the example of the ***definition*** **DC.**

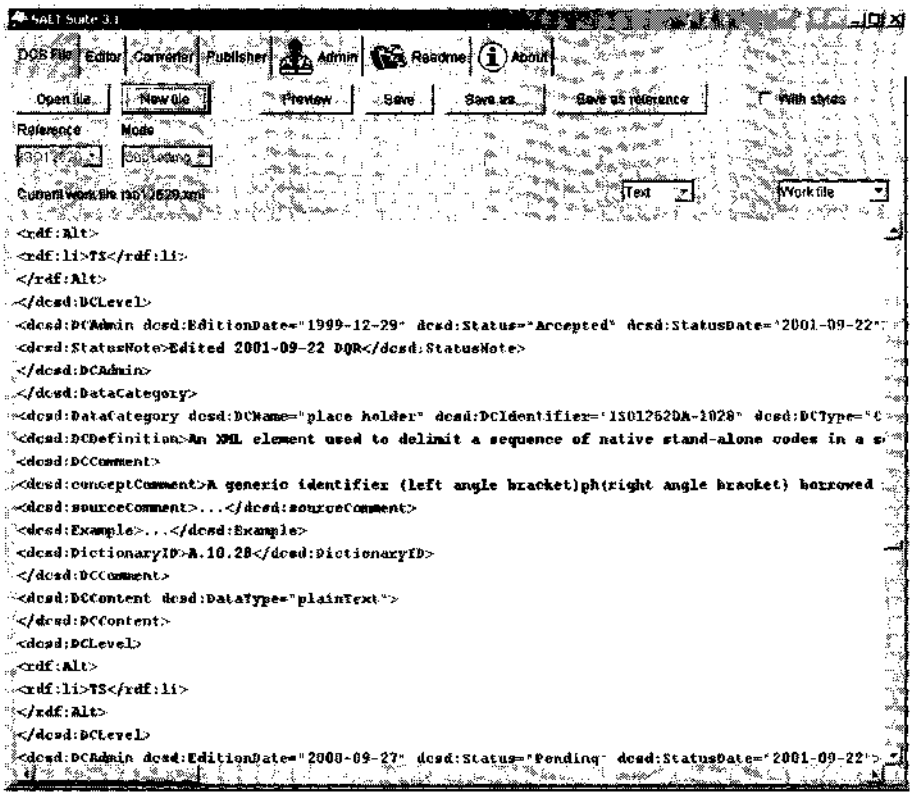

Figure 4: RDF-based Data Category Specifications as viewed through the SALT Suite

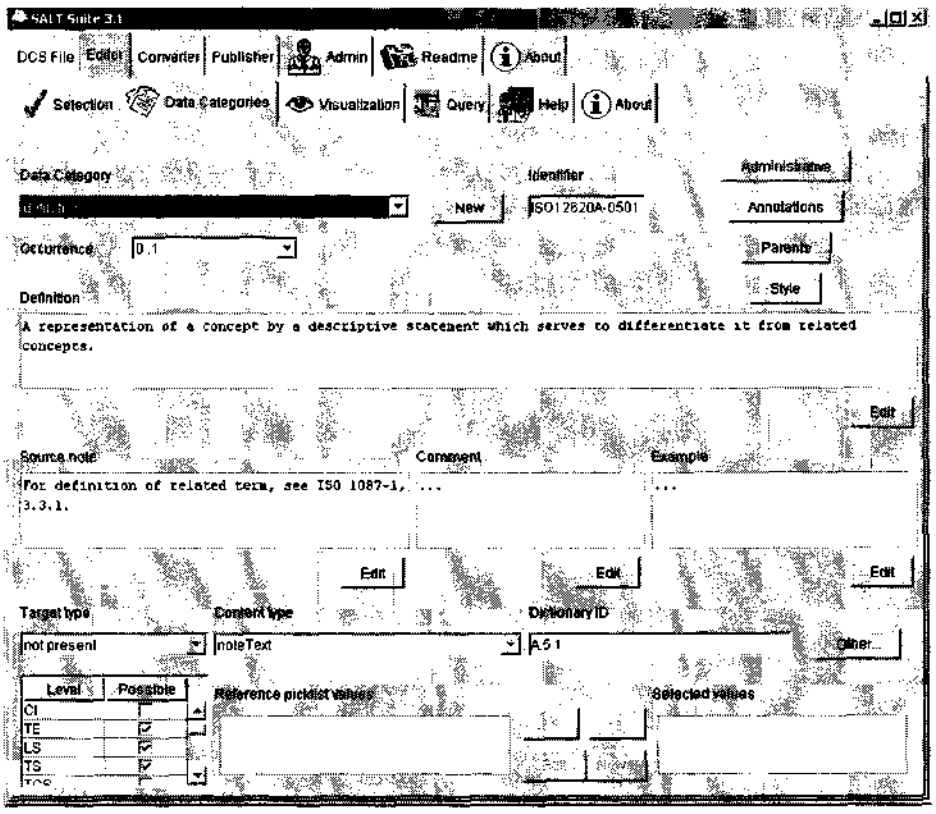

Figure 5: Defining a Data Category - the example of *definition*

Defining DCs within this environment, using the XML-based RDF enables a number of possible displays of the results, including the ability to search this consistently defined set of data, by making use of the extensible Stylesheet Language for Transformations (XSLT, Clark, 1999), extensible Hypertext Markup Language (XHTML, Pemberton, 2000) and a variety of other XML-based technologies. Through this tool, the data set is searchable by a number of criteria (see Figure 6).

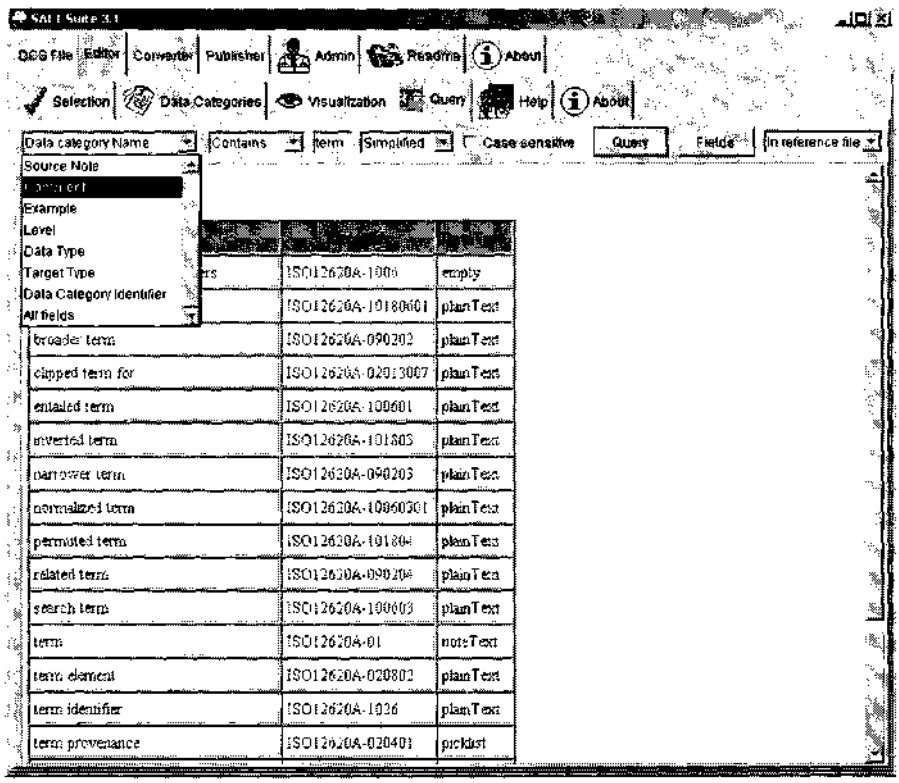

Figure 6: Searching particular values of Data Elements of the DCs for, in this example, those containing *term*.

The SALT Suite provides an environment in which to describe and maintain Data Categories, and is likely to form the basis of a tool for the registry for Part 1 of the revised ISO 12620, and for DC collections built that conform to it in future. We have yet to describe the combination of the standards for describing terminology collections. This is the focus of the following subsection.

Combining Data Categories and the Metamodel

To use ISO 12620 Data Categories within the ISO 16642 framework, which is itself a framework that can be implemented using XML, a number of items need to be declared for each Data Category. We therefore assume conformity with the levels used in ISO 16642. There are two possible ways in which to make use of the combination of these standards: first, to document an existing format, and second, to describe a new format. In both cases, the goal is to have a Terminological Markup Language (TML) that, through its conformity to these standards becomes interoperable with other TMLs. It is important to note that MARTIF and GENETER have both been described in this manner, and so any TML should be interoperable with these formats and hence make use of various applications that make use of data in these formats.

Describing a new TML

If we consider an example of defining the use of the ISO 12620 *term* and *language identifier* data categories, we have to define:

- **Style** - e.g. <XMLElement> or XMLAttribute
- **Vocabulary** - e.g. called "termString" or "lang"
- **Anchor** - <XMLElement> anchored on structural node of 16642; XMLAttribute anchored on an <XMLElement>
- **Possible Values**

So, for the chosen data categories, the following table (Table 2) shows possible definitions for these DCs:

	Term	*Language Identifier*
Style	XMLElement	XMLAttribute
Vocabulary	TermString	lang
Anchor	TS [16642 Term Section]	term
Possible Values	String	List from ISO 639-1

Table 2: Values for 'styling' DCs for use in a terminology collection

Such descriptions could be used produce (and validate) the following XML fragment:

<termString lang="en">.....</termString>

A full set of such definitions identifies the structure and content of a TML, and the comparison of two such sets of definitions enables the degree of interoperability between such TMLs to be defined. The labels chosen for the DCs are not as important as the fact that a common DC is used in two different TMLs. This ability to refer to common sets of information provides for greater degrees of reusability and also documents the system being developed.

Describing an existing format

Consider the following entry (Example 1) from an Automotive Engineering terminology collection, encoded in XML as shown below. Note that the text (strings) in angle brackets (<, >) refer to implementations of Data Categories in this collection - the left angle bracket (<) indicates the start of the descriptor for a given data category and the right bracket demarcates the descriptor. We then have the value associated with the category followed by a left angle bracket and a slash ('/') and the descriptor name enclosed finally by the right angle bracket. The slash indicates the end of the associated data.

```
<termBank>
        <tbid>00aa</tbid>
        <tbDescription>Automotive Engineering</tbDescription>
        <conceptEntry>
                <domainOfConcept>ABS</domainOfConcept>
                <conceptLastModified>21-08-2001</conceptLastModified>
                <termGroup>
                        <languageCode>Deutsch</languageCode>
                        <termDefinition> Bauteile, die die elektronischen Steuer- und
                        Regelvorgänge für die Blockierregelung und die
                        Antriebsschlupfregelung übernehmen.</termDefinition>
                        <termString>ABS/ASR-Steuerung</termString>
                        <usageDescriptors>
                                <usedIn>Germany</usedIn>
                                <usedIn>Switzerland</usedIn>
                        </usageDescriptors>
                        <wordClass>n</wordClass>
                        <wordGender>f</wordGender>
                        <termLastModified>21-08-2001</termLastModified>
                </termGroup>
                <termGroup>
                        <languageCode>English</languageCode>
                        <termString>ABS/ASR control</termString>
                        <usageDescriptors>
                                <usedIn>Britain</usedIn>
                        </usageDescriptors>
                        <wordClass>n</wordClass>
                        <termLastModified>20-08-2001</termLastModified>
                </termGroup>
        </conceptEntry>
</termBank>
```

Example *1:* XML-encoded example from a German Automotive Engineering Terminology collection

This entry contains a German term and its associated attributes (Table 3), including subject domain *(automotive engineering),* the specific terminology project it was created for, the usual attributes found in most term bases, i.e. definition, grammar, the specialised subject field *(ABS - Anti-lock braking systems),* and the modification history of the term (Table 3). Note that the DCs contained in this example have associated values. Analysis of this entry in terms of ISO 12620 produces Table 3, below. From this table, it is evident that some of these values are DCs in their own right (the so-called *simple* DCs that can not be expanded by further DCs) and are indicated in column 2. Column 2 also shows the relation to DCs that could be constructed from other standards such as ISO 639 and ISO 3166 (Country Codes). Table 3 also contains the reference of the specific ISO 12620 DC in column 4.

Data Category	DC of Value	Value	ISO 12620 Reference
Subject Field		ABS	**12620A.4**
Term		ABS/ASR-Steuerung	**12620A.1**
Part of Speech	**ISO 12620: A.2.2.5.2 - common noun**	Noun	**12620A.2.2.1**
Language Identifier	**ISO 639-1: de**	de-639.1	**12620A10.7.1"**
Grammatical Gender	**ISO 12620: A.2.2.2.2 - Feminine**	feminine	**12620A.2.2.2**
Geographical Usage	**ISO 3166-1: DE**	DE-3166.1	**12620A.2.3.2**
Geographical Usage	**ISO 3166-1; CH**	CH-3166.1	**12620A.2.3.2**
Definition		Bauteile, die die elektronischen Steuer- und Regelvorgänge für die Blockierregelung und die Antriebsschlupfregelung übernehmen	**12620A.5.1"**
Subset Identifier-			**12620A.10.3**
Modification	**ISO 12620A.10.2.1.3: Modification Date**	21-08-2001	**12620A.10.1.3**
Project Subset		Automotive Engineering	**12620A.10.3.3**

Table 3: Analysis of a German-based terminological entry

By carrying out such an analysis of the information encoded into the example, we can identify the ISO 12620-based DCs that have been employed to describe this example. We now have an example that can be described in terms of a common reference. We can also identify the XML markup that corresponds to the ISO 16642 metamodel - in this case, **termbank** corresponds to **Terminological Data Collection (TDC), conceptEntry** corresponds to **Terminological Entry** (TE) and termGroup corresponds to **Language Section** (LS). Completion of the analysis requires the manipulation of these data to produce the **Global Information Section** (GIS) and the introduction of a **Term Section** (TS), which can be made without loss of information. One final step requires the manipulation of the so-called **Modification** since this attribute and its value together represent two pieces of information, a transaction and a date of the transaction. This dependency is not obvious in the current specification. The resulting splitting produces the **brack** structure shown below in the GMT-based ISO 12620/ISO 16642 conformant example (Example 2).

```
<struct type="TDC" xml:lang="en">
        <struct type="GIS">
                <feat type="subsetIdentifier-12620A.10.3">00aa</feat>
                <feat type="projectSubset-12620A.10.3.3">Automotive Engineering</feat>
        </struct>
        <struct type="TE">
                <feat type="subjectField-12620A.4">ABS</feat>
                <brack>
                        <feat type="terminologyManagementTransactions-
                        12620A10.1">modification-12620A.10.1.3</feat>
                        <feat type="modificationDate-12620A.10.2.1.3">21-08-2001</feat>
                </brack>
                <struct type="LS" xml:lang="de">
                        <feat type="languageIdentifier-12620A10.7.1"> de-639.1 </feat>
                        <feat type="definition-12620A.5.1">Bauteile, die die
                        elektronischen Steuer- und Regelvorgänge für die
                        Blockierregelung und die Antriebsschlupfregelung
                        übernehmen.</feat>
                        <struct type="TS">
                                <feat type="term-12620A.1">ABS/ASR-Steuerung</feat>
                                <feat type="geographicalUsage-12620A.2.3.2"
                                xml:lang="en"> DE-3166.1 </feat>
                                <feat type="geographicalUsage-12620A.2.3.2"
                                xml:lang="en"> CH-3166.1 </feat>
                                <feat type="partOfSpeech-12620A.2.2.1">n</feat>
                                <feat type="grammaticalGender-12620A.2.2.2">feminine-
                                12620A.2.2.2.2</feat>
                                <brack>
                                        <feat type="terminologyManagementTransactions-
                                        12620A10.1">modification-12620A.10.1.3</feat>
                                        <feat type="modificationDate-12620A.10.2.1.3">21-
                                        08-2001</feat>
                                </brack>
                        </struct>
                </struct>
                <struct type="LS">
                        <feat type="languageIdentifier-12620A.10.7.1"> en-639.1 </feat>
                        <struct type="TS">
                                <feat type="term-12620A.1">ABS/ASR control</feat>
                                <feat type="geographicalUsage-12620A.2.3.2"
                                xml:lang="en"> GB-3166.1 </feat>
                                <feat type="partOfSpeech-12620A.2.2.1">n</feat>
                                <brack>
                                        <feat type="terminologyManagementTransactions-
                                        12620A10.1">modification-12620A.10.1.3</feat>
                                        <feat type="modificationDate-12620A.10.2.1.3">21-
                                        08-2001</feat>
                                </brack>
                        </struct>
                </struct>
        </struct>
</struct>
```

Example 2: GMT-based ISO 12620/LSO 16642 conformant example produced through analysis of term collection example

The result show above in the format-neutral GMT can be converted into an existing well-defined format such as MARTIF or GENETER, as mentioned previously. By identification of the 16642 metamodel and identification of the ISO 12620 Data Categories being used, as well as the splitting out of certain items of data that are encoded into the tags themselves, we have produced an example that can be used in two other environments, emphasising the reusability potential of this approach.

Terminology collection, reuse and interchange

Terminology collections, stored as data files, databases and prototypical knowledge bases, are an essential component of a range of enterprises and are sponsored by governmental and non-governmental organisations. Well known, well-populated multilingual terminology resources include:

- **Eurodicautom,** the European Commissions multilingual termbank containing around 5 million terms in upto 11 languages
- **TIS**, the terminological database of the European Council containing around 600,000 terms in upto 11 languages
- **Euterpe,** the terminology of the European Parliament containing over 1 million terms

These three major multilingual terminology resources use different formats to store individual terms. By themselves, these differences are minor but put together collectively they can impede the transfer of terminology data from one system to the other. This is fact is not lost on the Translation Centre for the Bodies of the European Union in their setting up of the Inter-Agency Terminology Exchange (IATE).

Information exchange within and across a domain requires an understanding of the catalogue of everything that makes up that domain. This involves an understanding of how knowledge is organised. A lack of standardisation is not restricted to Europe - the terminology collections in medicine are a good example here, the US-based *Unified Medical Languages System* (UMLS) (represented as a large 'semantic network') containing medical terms, is organised differently to the terminology database of the World Health Organisation. A number of private enterprises give away or sell terminology collections in finance and commerce. Terminology collections play an important role in enabling multi-lingual and often linguistically divided communities to communicate amongst themselves, in less favoured or minority languages, and with each other. Examples here include South African efforts in creating terminology collections in 11 of its languages; the Canadian government's Anglo-French terminology collections; and the rise in such collections in Eastern and Central Europe.

Conforming to standards is an initial step towards reusability. Provided that we are able to reference a common system or common set of systems, and that the language marking is consistent with these systems, collections that employ the system and the identifiers would appear to have some degree of possible convergence or reusability. The choice of the codes used from the reference set would depend on the coverage of the standard. Even with the comprehensive standards mentioned above, it is possible to find information related to one language missing in one collection and present in another which will make the task quite complicated.

We have shown how the application of two standards in tandem, both of which contain fairly recent developments from the world of computing, can assist in the process of both designing and reusing terminology collections. The purpose of such standards is the interoperability of systems so that every new system does not have to be developed in an ad-hoc fashion. Using an abstract format (GMT) to produce a TML enables this interoperability - conformity with ISO 16642 and ISO 12620

ensures much greater degrees of interoperability than is currently possible. By conversions to existing well-defined formats (MARTIF, GENETER) which have been defined by ISO 12620 and ISO 16642 and are themselves TMLs, data manipulation can be carried out with existing tools *(leveraged).* By taking this approach, we improve the potential for computer-mediated validation of content and can ease the problems associated with import/export/reuse of terminology. Furthermore, terminology collections can be harmonised by their fit to a common structure and to common data sets. The problem remains of how to map between sets of values which do not align so easily, or, as seen previously, where information is coded in the markup.

Knowledge Exchange - Notes on Ontology

So far, we have addressed mainly the problem of form of terminology. Terminology collections contain values which are not so easy to align. The simplest of these is the mapping between value scales, where one system uses the values 1 to 3 and another uses 1 to 4: the values at the ends of the scale should be easy to handle, but how do values in the middle of the scale map? Probably the most complex of values to map are the subject field identifiers used such as the Lenoch Universal Classification (LUC) and the Universal Decimal Classification (UDC). These values are coded in highly granular hierarchies, where the developers of these hierarchies are possibly amongst the few who understand them. For this reason, alignment of values within such systems will take time, and is most certainly beyond the scope of this paper.

Ontologies for Data Categories

To harmonise terminology collections, and indeed to provide harmony amongst any collections containing metadata, it is important to ensure that issues of granularity and content can be managed at least at the simpler levels, to enable more complex issues to be dealt with. An overview of the Metadata Encoding and Transmission Standard (METS[5]) shows the following metadata description (Example 3).

```
<dmdSec ID="dmd002">
  <mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core
Metadata">
      <dc:title>Alice's Adventures in Wonderland</dc:title>
      <dc:creator>Lewis Carroll</dc:creator>
      <dc:date>between 1872 and 1890</dc:date>
      <dc:publisher>McCloughlin Brothers</dc:publisher>
      <dc:type>text</dc:type>
</mdWrap>
</dmdSec>
```

Example 3: XML example from the Metadata Encoding and Transmission Standard

[5] see http://www.loc.gov/standards/mets/METSOverview.html. last visited 1 October 2002

In this example, the **dc:date** tags are used to code data that would have to be parsed by machine in order for it to be used by the machine, or to make it available in translated form. This could be considered as a form of *tag abuse.* It tends to indicate that the tags required are not granular enough to express the range of dates, and this type of coding prevents the interchange of data between systems. This simple example shows a need for monitoring of the use of such markup, and perhaps the need for reporting mechanisms that identify non-conformant usage.

Where such data are conformant, it may be that to enable reuse, values need to be mapped across collections of DCs. In the simplest case, there may be a one-to-one mapping between values in data sets, for example the mapping between the French word for 'English', the code for English in ISO 639-1, and the code for English in a possible ISO639-4:

Anglais - ISO639-l:en - ISO639-4:50-eng

We could easily consider mapping between collections that have XML data denoted in such ways (again using XSLT). Of course, this is also possible if we have acquired 2 sets of transforms since we can iterate the translation process

Anglais - ISO639-l:en
ISO639-l:en - ISO639-4:50-eng

By iterating in this fashion, and by treating the transformation process in such a modular fashion, we can consider collections that can be mapped between increasing numbers of systems which share the same reference, but which may wish to migrate to larger scale reference sets, for example, to move from 2-letter codes of ISO 639 to 3-letter codes of ISO 639.

We can further consider the W3C version of **date** as being a restriction of the syntax that is possible in ISO 8601 (ISO 8601, Date formats). Similarly, **dc:creator,** the creator of a particular resource, from the METS example above, could be described as a conjunction of, for example, the vCard[6] electronic business card objects **vcard:family** and **vcard:given. xml:lang** can be considered as a conjunction of ISO 639 and 3166. This hints at a rule-base for the framework for interoperability between data sets. Indeed, from such combinations, a grammar is possible, for example:

xml:lang = '"ISO 639'" + '"-' + 'ISO 3166"'
dc:creator = "'vcard:given' + '' + 'vcard:family'"

The potential for a rule-based approach to the interpretation of hidden semantics within terminology resources would further enable interoperability - again, we are moving to more granular systems. The specialisation and generalisation via such rules requires further investigation.

[6] see: http://www.w3.org/TR/vcard-rdf - last visited 1 October 2002

Using a terminology in an ontology?

An ontology is variously defined as a "specification of a conceptualization". Terminology specifies concepts within concept systems. We have hinted above at the means with which to harmonise collections of terminology as one relation between terminology and ontology. From the above definition, it is possible to consider the population of an ontology *with* the concepts from a terminology collection. As such, terminology collections could provide the vocabulary for conceptualisations and be used in the Semantic Web by conversion from, say, GMT to a format such as the Web Ontology Language (OWL)[7]. As standard terminologies have been created by committees of experts, the requirement for access to experts in the production of knowledge-based systems (part of the knowledge acquisition bottleneck) becomes lessened. It is worth considering the leveraging of such terminology collections within the ontology field itself.

Future work of ISO TC37

The revision of ISO 12620 is currently in progress, in line with many of the details described in this paper. ISO 12620 and ISO 16642 in combination are also being used as models for work in ISO TC37 SC4, where the current focus is on the creation of a Linguistic Annotation Framework. This work is in the early stages, however with lessons learned from these standards, the progress should be rapid. There is, however, significant work already being carried out on new parts of ISO 639 for language codes, described below.

ISO 639

Language is a universal and consistent parameter in documentation. The anticipated increase in spoken as well as written documentation and software, and the consequent need for a coherent global system of language identifiers require a transparent, accurate and unambiguous scientific tagging and referential coding of all the world's languages and speech-communities.

ISO 639 and ISO 3166 (Country Codes) identify some of the world's written languages and all corresponding countries. These coding systems also form an essential component of the World Wide Web Consortium's extensible Markup Language (XML), where the *xml:lang* attribute takes values constructed from at least these standards as identified in the Internet Engineering Task Force *Request for Comments 3066* (IETF RFC3066). The increasing usage of XML in all forms of business communication hints at the need for an extended and unified system for representing language tags. As the XML community is rapidly expanding into business communication, and with the working scope of ISO Technical Committee 37 (TC37) including language resources such as speech where XML-based industry

[7] http://www.w3.org/TR/2002AVD-owl-ref-20020729/

standards are the norm, an XML-based implementation of such an unambiguous referential system is essential.

The need for a global system of language identifiers and coding goes far beyond the existing use of convenient abbreviations of commonly used language names. A reformed and expanded system needs to incorporate an informative function, providing an unambiguous key to the identification, nomenclature, relationships, varieties, locations and relative dimensions of all known languages and speech-communities in today's world. In multilingual translation and interpretation it is particularly important to measure and record the proximity of related languages.

TS/1 has invited the Linguasphere Observatory in Wales to propose a standardised alphanumeric system for tagging and coding all the world's languages. This system, based on the Linguasphere Register (Dalby, 2000), builds upon existing ISO 639 (Language Codes), provides a method to deal with their inconsistencies, and covers the majority of languages as yet uncoded by ISO The Linguasphere Register 1999/2000 has identified, classified and coded 13,840 inner languages (plus 8,881 constituent dialects) within 4,994 outer languages and 694 linguistic sets. Each *set* of languages is classified and coded within one of 100 referential *zones* within one of 10 referential *sectors* (one of 5 *phylosectors* or 5 *geosectors).*

It is intended that the proposed system of Linguasphere identifiers (identifying tags plus referential and relational codes) will support the alpha-2 and alpha-3 sets represented by ISO639-1 and ISO639-2. This system will provide the necessary systematic mapping required for migration to this more extensive set of identifiers at the same time as maintaining and amplifying the core of written standard languages. The dissemination and usability of such a system is therefore of great importance to business and government users alike who share a common need for transmission of information. Influencing and assisting international consortia by the introduction and use of this system can be seen as an important scientific contribution from the UK perspective. This system is being considered as ISO 639 Part 4 for dialects, with ISO 639 Part 3 being based on the Summer Institute of Linguistics' (SIL) Ethnologue (Grimes, 2000).

References

Berners-Lee,T. (1999) "Weaving the Web: The Past, Present and Future of the World Wide Web by its Inventor." London: Orion Business Books

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E. (eds.), 2000. "Extensible Markup Language (XML) Version 2.0". W3C Recommendation. http://www.w3.org/TR/REC-xml

Clark, J. (ed.) (1999). "XSL Transformations (XSLT)". Version 1.0. W3C Recommendation. http://www.w3.org/TR/xslt

Dalby, D. (2000) "Linguasphere Register of the World's Languages and Speech Communities". Hebron (Wales). Two volumes: ISBN 0 9532919 1 X and 0 9532919 28

Fellbaum, C. (ed.) (1998) "WordNet: An Electronic Lexical Database". The MIT Press, Cambridge, US

Goldfarb, C.F. (1990). "The SGML Handbook" Y. Rubinsky (Ed.), Oxford University Press, Oxford, UK.

Grimes, B.F. (Ed.) (2000) "Ethnologue: Volume 1 Languages of the World" 14th Edition 866 pp., ISBN 1-55671-103-4

ISO8601:2000 "Data elements and interchange formats — Information interchange — Representation of dates and times". International Organization for Standardization

ISO/IEC 11179:1994 Information technology — Specification and standardization of data elements. International Organization for Standardization

- Part 1: Framework for the specification and standardization of data elements
- Part 2: Classification for Data Elements
- Part 3: Basic Attributes of Data Elements
- Part 4: Rules and Guidelines for the Formulation of Data Definitions
- Part 5: Naming and Identification Principles for Data Elements
- Part 6: Registration of Data Elements

Lassila, O., Swick, R.R., 1999. "Resource Description Framework (RDF) Model and Syntax Specification". W3C Recommendation. http://www.w3.org/TR/REC-rdf-syntax

Miller, G. A., Beckwith, R., Fellbaum, C. Gross, D. Miller, K. (1993) "Introduction to Wordnet: An On-line lexical database" Princeton University, USA

Pemberton, S. (2000) "XHTML 1.0: The Extensible HyperText Markup Language". World Wide Web Consortium, Recommendation REC-xhtml 1-20000126

Acknowledgements

This work was partially funded by the European Union through the Standards-based Access to Lexicographical and Terminological multilingual resources (SALT: IST-1999-10951) and Generic Information-based Decision Assistant (GIDA: IST-2000-31123) Projects and the EPSRC through the Scene of Crime Information System (SOCIS: GR/M89041/01) project.

Annex A: Standards published by ISO TC37

TC Level	ISO 639:1988	Code for the representation of names of languages	Provides a list of two-letter symbols for the representation of names of languages
TC 37/SC 1	ISO 704:2000	Terminology work -- Principles and methods	Provides methodological guidance for preparing and compiling terminologies
	ISO 860:1996	Terminology work -- Harmonization of concepts and terms	Provides methodological guidance for international harmonization of concepts, concept systems, definitions, terms and term systems
	ISO 1087-1:2000	Terminology work -- Vocabulary -- Part 1: Theory and application	Provides a set of fundamental terms and concepts for applied terminology work.
TC 37/SC 2	ISO 639-2:1998	Codes for the representation of names of languages -- Part 2: Alpha-3 code	Provides a list of three-letter symbols for the representation of names of languages
	ISO 1951:1997	Lexicographical symbols and typographical conventions for use in terminography	Specifies symbols and layout conventions for use in specialized dictionaries and terminology databases
	ISO 10241:1992	International terminology standards -- Preparation and layout	Establishes rules for use in preparations for and layouts of international terminology standards.
	ISO 12199:2000	Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet	Provides information on the ordering of terminological and lexicographical data
	ISO 12616:2002	Translation-oriented terminography	Provides application methods for translators, localisers. etc. for compiling glossaries
	ISO 15188:2001	Project management guidelines for terminology standardization	Provides guidelines for standard terminologies
TC 37/SC 3	ISO 1087-2:2000	Terminology work -- Vocabulary -- Part 2: Computer applications	Specifies terms and definitions for language and information processing for applications in terminology work and terminography
	ISO 6156:1987	Magnetic tape exchange format for terminological/ lexicographical records (MATER)	Provides rules for the exchange of terminological and lexicographical data on magnetic tape among large-scale terminology data banks
	ISO 12200:1999	Computer applications in terminology -- Machine-readable terminology interchange format (MARTIF) -- Negotiated interchange	Provides guidance for programmers and analysts in designing export and import softwares designed for data interchange among terminological databases
	ISO/TR12618:1994 (Technical Report)	Computational aids in terminology -- Creation and use of terminological databases and text corpora	
	ISO 12620:1999	Computer applications in terminology -- Data categories	Defines data categories for recording terminological data in both computerized and noncomputerized environments and for the interchange and retrieval of terminological information