A Global Data Category Registry for Interoperable Language Resources

Sue Ellen Wright
Institute for Applied Linguistics, Kent State University
sewright@neo.rr.com

Abstract

ISO TC 37 is creating a Data Category Registry (DCR) as an online open-source RDF-based resource for use by implementers of electronic language resources, including terminologies, presentational and non-presentational lexical resources, NLP lexica, etc. The DCR will allow dynamic generation of data category selections (DCSs), e.g., subsets of the collection reflecting various thematic domains and different data category classes and functions. The DCR will facilitate interchange and interoperability in heterogeneous environments. Participation of a wide range of experts from the broader computing community is important, as is provision for user-friendly guidance for implementers of databases and other resources.

Data Categories for Language Resources

ISO 12620:1999

ISO Technical Committee 37, Terminology and Other Language Applications, published a standard in 1999 specifying data categories used in terminological resources, ISO 12620:1999, *Computer assisted terminology management — Data Categories*. At that time, it seemed appropriate to create a printed document. TC 37/SC 3 was only working with terminological data categories, i.e., those that are needed for terminology management and used to decorate the meta-model defined later in ISO 16642:2003, *Computer assisted terminology management — Terminology Markup Framework (TMF)*. Furthermore, there was still a consensus at the time for the standard to be oriented toward human readers and database designers rather than toward machine processing. In the meantime, however, the SALT project (see Loria) has developed a platform-independent Java-based tool for manipulating the data categories that can be used to subset the existing collection and to output a variety of resources, including RDF (Resource Description Framework) or various XSL-generated HTML representations of the registry and its subsets.

Although the standard is not yet due for review, in 2003, TC 37/SC 3 initiated a revision of the existing document with the intention of creating a family of data category standards designed to meet the needs of terminologists and other language experts developing a variety of electronic linguistic resources. The intention was to include data categories for a variety of applications, including terminologies, non-presentational electronic lexical resources, presentational dictionaries (both hardcopy and electronic), and machine translation lexica, as well as morphological, syntactic, semantic markup formats, etc. These areas of interest have been designated *thematic domains* and include TC 37/SC 3 (computer applications in terminology), SC 2 (presentational lexicography) and of SC 4 (languages resources, non-presentational natural language processing, e.g., a number of components of the TEI (Text Encoding Initiative) and the ISLE projects). The major objectives of this project are to ensure interoperability among these domains and to avoid any redundancy in data categories that are used across domains. It is also important to ensure linkage with evolving standards for rule-based ontologies. A further goal is to achieve interoperability among resources in integrated heterogeneous computing systems, with an eye to facilitating the implementation of wide-scale information handling environments such as the Semantic Web (Daconta et al., 2003).

The Global Data Category Registry

The objective of this effort is to create an open-source Data Category Registry (DCR) compliant with ISO 11179-3, *Information technology — Metadata registries (MDR) — Part 3: Registry metamodel and basic attributes.* Users will be able to access data category specifications via the Web. This resource must be machine-processable, preferably in some flavor of XML or RDF. Compliance with ISO 11179 will ensure compatibility with other data element registry initiatives. ISO 12620:1999 does not comply in all respects with ISO 11179, primarily because the latter standard was in flux while ISO 12620 was being developed, which made it difficult to maintain coherence between the standards. Today ISO 11179 has matured to the extent that TC 37 can confidently implement its requirements.[1]

In addition to the level of maturity we now see in ISO 11179, TC 37 has developed and is developing high level data models designed to negotiate interoperability on a structural level. ISO 16642:2003 (TMF, cited above) has provided the community with the ability to relate the data categories presented in the TC 37 DCR to a metamodel for terminological entries, thus greatly enhancing the feasibility of wide-scale exchange and interoperability for concept-oriented information. Recent developments in SC 4 are producing a high level meta-model for word-oriented, non-presentational electronic lexical resources such as NLP lexicons. This Lexical Markup Framework (LMF) in its latest iteration is intended to provide a very powerful, highly flexible abstract model that will accom-

[1] One remaining difference is terminological: *data categories* in TC 37 are called *data elements* in JTC 1/SC 32.

modate numerous extensions in order to ensure interoperability among various thematic domains.

A Proposed Family of Data Category Standards

According to resolutions passed during the 2003 meetings of TC 37/SC 3, ISO 12620 would in future comprise a family of standards. The umbrella standard governing this suite of standards would be the current CD ISO 12620-1, which defines the meta-structure for the global DCR as described in the previous section. This standard supports a framework for an unspecified number of data category selections (DCSs), each listing the data categories and their definitions used in a particular thematic domain. According to this model, each DCS would be represented by a normative standard numbered 12620-n. Figure 1 illustrates this model, where the peripheral oval describes the entire DCR and the large overlapping ovals represent individual thematic domain DCSs. The small octagonal structures represent individual applications, which typically feature a very small subset of any given domain DCS.

The overlapping ovals in Figure 1 reflect the fact that there is considerable redundancy between the data categories used in different kinds of language resources, especially with respect to terminology, lexicography, and machine translation lexica. Elements such as "definition", "context", "source", etc., are very likely to appear in several, if not all environments, as are specific language-related elements such as "part of speech", "grammatical gender", and the like. It is important that these descriptive attributes be fully portable and leverageable between resources and resource types in interoperable computing environments. Of course, it should be noted in this context that not all data elements behave in the same way in different applications, e.g., some data elements used in machine translation lexica play a different function within their database environments from their role in electronic terminological resources (ETRs) or electronic lexical resources (ELRs). This phenomenon underscores the reality that interoperability will be relative at best and that there is no absolute guarantee of a lossless roundtrip for all exchange transactions.

Implementing the DCS Documents

As implied above, the 2003 discussions in TC37 visualized the identification of a manageable subset of the DCR for each thematic domain, which led to the elaboration of the proposed new work items ISO 12620-2 to replace the original 12620:1999 and ISO 12620-3 (for non-presentational electronic lexical resources). This work was undertaken with the intention of including other similar collections in the future. The initial abstract model for these subsets of the DCR envisioned fairly clear divisions within the global set, but actual elaboration of the standards revealed that the subsets are not that well defined. The new work item proposals, together with the ensuing comments on these collections, reveal huge overlaps among the various thematic domains. Figure 1 (which is not statistically proportionate) is a modification of the original, apparently more manageable model in that it shows the high level of redundancy among the different DCSs. Here the small, non-overlapping segments of the DCS ovals (identified in Figure 1 by the letter X) represent those data categories that are specific to each thematic domain. Given these considerations, representing the large, unwieldy lists of shared data categories in printed DCS standards is not necessarily useful or easy for potential users to interpret because of the size of the collection and the apparent repetitions from one DCS to the next.

In this light, several issues arise with respect to subsetting the global collection. The DCR is large, as are the proposed DCSs cited here. ISO 12620: 1999 attempted to provide a clear overview of the collection by classifying the data categories into eleven major groups, but this classification was difficult to arrive at and does not satisfy anyone. Confronted with various schemata for reclassifying the collection, TC 37/SC 3 decided in 2002 that it would be both counter-productive to spend a great deal of time on such a project and potentially confusing to propose multiple views on the collection. Given these considerations, the data categories are now simply alphabetized, and in fact, the future intention is not to present them in their entirety in printed form, although users can easily generate html tables or other printed resources from the master file or its subsets. The old position numbers from 12620: 1999 will be retired, and the data category names themselves become the unique theoretically non-mnemonic identifiers.

The previous paragraph outlined the now obvious difficulties involved in classifying the data categories according to thematic domain. Nevertheless, the notion presents itself that all these different classification schemes, together with various other subsetting criteria, could be used to identify data categories within the DCR in order to dynamically generate numerous subsets as needed. This procedure would allow users with different objectives to create faceted views of the collection, with these subsets intersecting each other in various ways.

Figure 2 presents one view of the kind of subsetting described above. Here the DCR is represented by the familiar flow-chart resource container. The columns represent potential DCSs as originally conceived in the Oslo model. Within these columns, the data categories are described according to their respective functions within language resources.

Those data categories that appear in the bottom sections of the individual columns are essentially identical across thematic domains, which means that they are redundant throughout at least some of the DCSs. This phenomenon accounts for the extreme overlap in data categories between some of the different thematic domains illustrated in Figure 2, and underscores criticisms registered in the comments from the NWIP ballot for 12620-3 in particular. This phenomenon is an argument in favor of a finding a different way of presenting the data categories from the solution proposed in Oslo. The data categories represented in the top sections enclosed by the oval are different for each domain. In some respects, these data categories may be more interesting

than the others, but they are very difficult to isolate in the kinds of printed documents proposed for NWIP 12620-2 and 12620-3.

The 11179 Model

In addition to internal TC 37 discussion and comment on the structure of the DCR and of potential DCS standards, discussions between TC 37 experts and representatives from JTC 1/SC 32, which is responsible for the ISO 11179 family of standards, have also supported a different approach to the configuration of the DCR and any subsets. Although the members of JTC1/SC 32 do not discount the value of a standard such as NWI 12620-2 as a successor to 12620:1999, they recommend maintaining the DCR as a TC 37-sponsored external resource that is not itself a standard, but that is, of course, clearly based on ISO 11179-3 and on an ISO 12620-1 (with the obvious understanding that there might not be a part number if there were no other actual components to the standard). This approach reflects procedures already followed in a number of existing data registries.

Reflecting on this suggestion, recent TC 37/SC 4 discussions have produced the concept that it would probably be more consistent to maintain the entire registry outside the standard and to abandon the notion of publishing subset-oriented individual DCS standards. This would result in a single ISO 12620 standard that reflects more or less the current ISO CD 12620-1. As indicated above, this standard would support an external DCR that would reside in an online, open-source resource administered and maintained by TC 37.

Data Category Profiles

In this model, alternative views of the collection, such as the assignment of individual data categories to such meta-classifications as "administrative data", "linguistic data", or to individual thematic domains, would be documented in the form of attributes included in the individual data category specifications in the DCR. These attributes would make up the *data category profile* of each data category specification in the DCR. Based on this information, subsets of data categories of various kinds could be generated as needed instead of maintaining them in printed standards. Thus the DCSs as originally envisioned in Oslo would be identifiable based on dynamic subsetting of the DCR, depending on user needs.

In the scenario described here, the DCR would be administered by a Registration Authority (RA) under the auspices of ISO TC 37. An RA is defined by a standard and maintains a resource external to the standard. In a framework involving multiple DCSs expressed as printed standards, each thematic domain would be responsible for establishing a Maintenance Authority (MA) to maintain and update each standard as new data categories are proposed, old ones deprecated, or other changes are introduced. If rather than creating separate standards, the various DCSs are expressed as sortable subsets of the DCR, then in all likelihood it would be feasible for designated task forces representing each thematic domain to administer changes to the data category profiles within the DCR without having to create MAs for this purpose. The critical issue here, particularly with respect to the balance of interests within TC 37 and across the thematic domains it represents, is that all SCs having a vested interest in the DCR and in the related abstract data models (e.g., LMF) have the opportunity to participate in discussions and to provide input for the evolution of the resource. While this may seem obvious, implementing the idea is not a trivial issue, given scheduling conflicts between the different groups. In future much of the work may be done online in order to facilitate broad group participation.

User-friendly Guides for Data Category Application and Management

The online availability of the DCR will be user-friendly for developers actually implementing databases according to the model outlined in ISO 16642:2003 and in the evolving LMF standard, but dissolving the hierarchical subsets of data categories presented in ISO 12620:1999 does have its downside. Newcomers to the standard can search for individual data categories online, but obtaining a clear overview of certain types of data categories and identifying sets of elements used for specific application areas (terminology planning, for instance, or controlled authoring) is not an easy prospect. The subsetting ability discussed here may not be transparently helpful to such users if they cannot visualize ahead of time the kinds of subsets they may want to be able to initiate.

It is not surprising that many potential users find the current data category collection unwieldy, and this concern is not going to improve as data categories are added to facilitate an expanded variety of electronic language resources. By the same token, it does not make sense to deny the admission of new, useful data categories just because the collection is already large. The most obvious solution to this dilemma is for TC 37 to produce a technical report designed to provide guidance to users wishing to establish their own language resources based on TC 37 standards**.** ISO 12616:2002 *Translation-oriented terminography* already provides information on data categories that are most useful for creating translation-oriented resources. There have been plans for several years to take up revision of the now withdrawn technical report, ISO/TR 12618:1994**,** *Computer aids in terminology — Creation and use of terminological databases and text corpora*. Concerns surrounding the creation of the DCR point to the utility of such a project.

Furthermore, some critical pieces of information will no longer be available as the older standards are withdrawn or superceded. For instance, the discussion of data modeling variance contained in informative Annex D of ISO 12620:1999 provides essential information for users of the standard unfamiliar with the approach taken in specifying the data categories. As an example, the item *synonym* can function as a field name in one database and in another database model as a permissible instance (member of a data domain) associated with the data category *term type*. Another example involves the explanation of mapping procedures for variant data category names used in individual applications, for ISO 12620 only requires the use of standardized names for interchange and inter-

operability and does not affect stand-alone legacy resources. The logical place for these kinds of information is in a user-friendly guide.

As already indicated, SC 3 has decided not to recommend any sort of standardized hierarchical concept system or classification schema. Nevertheless, SC 4 has discussed the feasibility of providing guidelines for creating an ontology or ontologies based on the DCR, with the notion that these kinds of ordering systems would be linked to OWL (Web Ontology Language) efforts and could incorporate rule-based inferential capabilities. The information needed to generate such embedded resources could also be included in the individual data category profiles described above. The SCs need to explore the options for facilitating the creation of such user-defined resources.

Outlook

The mission of TC 37, particularly of SCs 3 and 4, is expanding and changing to meet the needs of new classes and generations of creators and users of linguistic resources. The accelerated schedule of meetings for SC 4, for instance, is evidence of the need to provide timely guidance to the language processing community in the form of market-responsive standards.

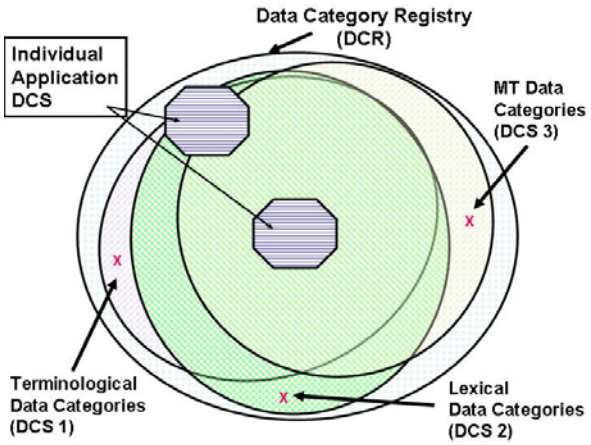

Figure 1: DCSs as subsets of the DCR

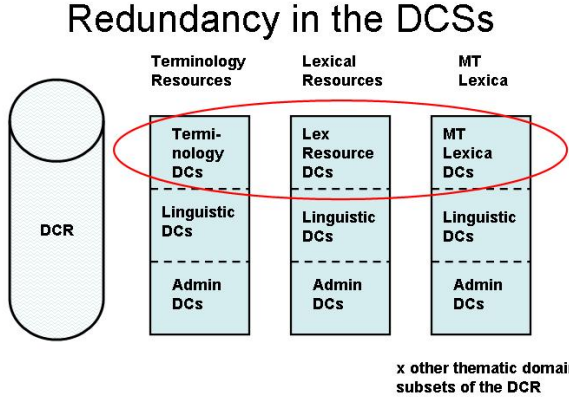

Figure 2: DCS data categories viewed for their function across resource types (thematic domains)

Acknowledgements

This paper reflects the collective work of ISO TC 37/SC 3 and SC 4, and particularly close collaboration with Prof. Laurent Romary, convener for ISO CD 12620-1 and Chair of SC 4.

References

Budin, G.; Lonsdale, D.; Melby, A.; & Wright, S. E. (1999). Integrating Translation Technologies Using SALT. In: Proceedings of ASLIB. London: ASLI, unnumbered pages.

Daconta, M. C.; Obrst, L. J.; & Smith, K.T. (2000). The Semantic Web: A guide to the Future of XML, Web Services, and Knowledge Management. Indianapolis, Indiana: Wiley Publishing, Inc.

ISLE. International Standards for Language Engineering. http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm. Accessed 2004-03-07.

ISO 11179: Specification and Standardization of Data Elements. See: http://www.iso.ch/iso/en/CombinedQueryResult. CombinedQueryResult?queryString=ISO+11179 http://www.diffuse.org/meta.html#ISO11179 Accessed 2003-09-29.

ISO/TR 12618:1994. Computer aids in terminology — Creation and use of terminological databases and text corpora (withdrawn)

ISO 12620:1999. Computer Applications in Terminology—Data Categories. Geneva, International Organization for Standardization, 1999.

ISO 12616:2002 .Translation-oriented terminography.

ISO 16642:2003. Computer assisted terminology management — Terminology Markup Framework. Geneva, International Organization for Standardization.

ISO TC 37/SC 4 N088. (2003). Terminology and other language resources—Lexical Resource Markup Framework (LMF). New Work Item Draft.

ISO TC 37/SC3 N 488. (2003). ISO/CD 12620-1. Terminology and other language resources—Data categories—Part 1: Specification of data categories and management of a data category registry for language resources.

ISO TC 37/SC3 N486. (2003). Terminology and other language resources—Data categories—Part 2: Data category selection (DCS) for electronic terminological resources (ETR).

ISO TC 37/SC4 N085. (2003). Terminology and other language resources—Data categories—Part 3: Data category selection (DCS) for electronic lexical resources (ELR).

Loria. SALT Technical Website. http://www.loria.fr/projets/SALT/saltsite.html Accessed 2003-10-27.

Melby, A. & Wright S. E. (1998). Data Analysis for Data Modeling and Mapping Data Categories Based on ISO 12200 and ISO 12620. In: Proceedings of the 4th TermNet Symposium Terminology in Advanced Microcomputer Applications. Tools for Multilingual Communication (TAMA '98), (pp. 273-302). Vienna: TermNet.

OWL. (2003). OWL Web Ontology Language Guide. http://www.w3.org/TR/2003/PR-owl-guide-20031215/ Accessed 2004-03-07.

TEI. Text Incoding Initiative. http://www.tei-c.org/. Accessed 2004-03-07.

Wright, Sue Ellen. (2001). Data Categories for Terminology Management. In Handbook of Terminology Management, Wright, S.E. & Budin, G. eds. (pp. 552-571). Amsterdam and Philadelphia: John Benjamins Publishing Company.