ISOcat: remodelling metadata for language resources

Marc Kemps-Snijders, Menzo Windhouwer, Peter Wittenburg and Sue Ellen Wright*

Institute for Applied Linguistics,
Kent State University,
109 Satterfield Hall,
Kent, OH 44242, USA
Email: Marc.Kemps-Snijders@mpi.nl Email: Menzo.Windhouwer@mpi.nl
Email: Peter.Wittenburg@mpi.nl Email: sellenwright@gmail.com
*Corresponding author

Abstract: The Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands, is creating a state-of-the-art web environment for the ISO TC 37 (terminology and other language and content resources) metadata registry. This Data Category Registry (DCR) is called `ISOcat` and encompasses data categories for a broad range of language resources. Under the governance of the DCR Board, `ISOcat` provides an open work space for creating data category specifications, defining Data Category Selections (DCSs) (domain-specific groups of data categories), and standardising selected data categories and DCSs. Designers visualise future interactivity among the DCR, reference registries and ontological knowledge spaces.

Keywords: Data Category Registry; metadata registry; ISO TC 37; terminology; `ISOcat`; SYNTAX; ontology; standard as database.

Reference to this paper should be made as follows: Kemps-Snijders, M., Windhouwer, M., Wittenburg, P. and Wright, S.E. (2009) 'ISOcat: remodelling metadata for language resources', *Int. J. Metadata, Semantics and Ontologies*, Vol. 4, No. 4, pp.261–276.

Biographical notes: Marc Kemps-Snijders is a Technical Coordinator and Senior Software Engineer at the Max Planck Institute (MPI) for Psycholinguistics in Nijmegen, The Netherlands. At the MPI he is responsible for the technical coordination of a number of linguistic user applications including the `ISOcat` DCR software. In ISO TC 37 he is involved in the DCR and lexical modelling, and is interested in standards that focus on improving interoperability between linguistic resources.

Menzo Windhouwer is a Researcher and Scientific Software Developer at the University of Amsterdam (UvA) and the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands. At the UvA he is a co-designer and the main developer of the *Typological Database System*: an interface which allows linguists to query a wide range of typological databases using one uniform interface. At the MPI he is co-designing and implementing the `ISOcat` DCR software.

Peter Wittenburg is a Technical Director and a member of the central IT advisory board at the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands. At MPI he and his team have set up one of the largest open standards-based archives for multimedia language resources. In ISO TC 37 his main interests are lexical modelling, setting up an ontology structure based on the ISO DCR and defining other standards for improving interoperability.

Sue Ellen Wright is a Professor in the Kent State University Institute for Applied Linguistics, where she teaches computer applications for translators and terminology studies. She is the co-compiler (with Gerhard Budin) of the *Handbook for Terminology Management*. She chairs the US mirror committee for ISO TC 37, *Terminology and other language and content resources* and is convener for ISO 12620. Her interests in TC 37 include, among others, the design of terminological databases and interchange formats for terminological and lexical data.

1 Introduction

1.1 The TC 37 Data Category Registry

ISO Technical Committee 37, *Terminology and other language and content resources*, is in the process of developing a revised version of its Metadata Registry (MDR) following the principles of the ISO/IEC 11179 family of standards, insofar as this is feasible within the framework of divergent traditions within the terminology management and language resources communities. The goal of the project is to create a universally available resource for language-related metadata that can be used in a variety of applications and environments, ranging from concept-oriented terminology management to machine-readable lexicons to morpho-syntactic markup in natural language processing resources, among others. Typical data categories that are treated in this collection include, for instance, */term/*, */part of speech/*, */definition/*, to name just a few examples. One major goal of the project is to provide uniform naming and semantic principles for such data categories so as to facilitate the interoperability or leveraging of language resources across applications and approaches.

This MDR, commonly referred to as the TC 37 *Data Category Registry* or *DCR*, was initially implemented in the so-called *SYNTAX* system (Ide and Romary, 2004). In this form, it has demonstrated its value as a proof of concept, but users have clamoured for an upgrade. The user interface is considerably improved both in function and linguistic presentation, help files are more user-friendly and bandwidth will be significantly improved. Most importantly for the standards-related implementation of the resource, full functionality for balloting and review will be put in place. The new revised DCR has been christened `ISOcat` and is available at http://www.isocat.org.

1.2 Historical precedence for the term 'data category'

The practice of using the term *data category* to cite individual data elements derives from a difference in terminological usage between the community of discourse dedicated to the computerisation of terminology management solutions and the general metadata community. The variation in terms and conceptual reference is grounded in at least 35 years or more of usage. Felber (1984) describes *data category* as a 'type of terminographic data', and notes that *data categories* can be associated with one or more *data elements*, implying that within original TC 37 usage, *data elements* are instantiations of data categories. Felber's use of *data category* and *data element*, however, tends to be ambiguous because the terms seem at times to be used synonymously in his writings. This may reflect an individual stylistic turn, but more likely the failure of the terms to have fully stabilised in the terminology community at that time.

Felber reports, for instance, on 'the First International Conference on Terminological Data Banks' in 1979, citing efforts 'to undertake or to arrange for a comparative study to be undertaken of the categories (data elements) employed in different term records for ordering and identifying terminological data' (p.119). His efforts to compare and codify terminological data categories focus on the enumeration of elements, as well as identifying and standardising cryptic symbols and short (two and three letter) codes for common data categories, a practice constrained by both the physical limitations of paper-based catalogue cards (terminology *fiches*) and of early computing environments, where inadequate memory and limited field length frequently imposed artificial strictures on data management practice.

The MATER standard (ISO 6156 Magnetic tape exchange format for terminological/lexicographical records) appeared in 1987 after some years in preparation. Here *data element* and *data category* are clearly defined with characteristics that still apply more or less adequately in current TC 37 practice:

data element: smallest separately identifiable portion of a record used for the description and/or representation of terminological or lexicographical data (ISO 6156, 4.4) (Compare to ISO 1087-2:2000, 6.11: *unit of data that in a certain context is considered indivisible*)

data category: uniquely defined type of a terminological/ lexicographical data element which can be used to structure and describe the content of a terminological/lexicographical record (ISO 6156, 4.5; Compare: ISO 1087-2:2000, *data element type, result of the specification of a given data field*)

1.3 Early compilation of data categories

Although Felber provided many examples of data category representations in a variety of hard copy and early database environments, the MATER standard represented the first effort to systematically list and standardise data categories in TC 37 practice, linking an exchange format, a range of modelling constraints, naming conventions and other aspects of data elements in the same standard. With the advent of the personal computer, which parallels the gradual demise of the magnetic tape medium for which the MATER standard was ostensibly designed, developers of early PC-based terminology management programs continued the effort to create three letter codes for data categories (Melby, 1991; e.g. DEF for /*definition*/; CTX for /*context*/, SRC for /*source*/). Although the memory limitations of pre-hard drive DOS applications motivated this effort to transfer paper-based conventions to the computing environment, repeated exponential gains in both working memory and storage capacity, coupled with the emergence of effective graphic user interfaces, particularly Windows™ applications, discouraged further development along these lines. (Other early GUI interfaces notwithstanding, this program development took place almost exclusively in the IBM PC environment.)

Based on Melby's usage and Sager (1990, 142 ff), it is apparent that the notion of the *data category* is firmly established by the early 1990s, although Sager remarks on the lack of consensus on a metalanguage for terminological information. Concerted efforts to define a metalanguage or more precisely, to specify terminological metadata, progressed in the framework of the Text Encoding Initiative (TEI; Budin et al., 1994). Here the term *data element type* is equated with *data category* and *database field type*, in response to the prevalent usage of the *element* designation in the TEI

environment. In the course of the TEI terminology interchange project, Budin and Wright (1994) reports on a comprehensive survey of the then available terminology management systems and documents hundreds of data categories, including both field names and specified values. Most importantly for further elaboration of the data categories, the adoption of SGML markup conventions set the stage for the collection and expression of data categories as standardised metadata labels. This work eventually established the core set later introduced in ISO 12620:1999 as data categories for terminology management.

1.4 Interaction with ISO/IEC Joint Technical Committee 1/Sub-Committee 32

In the meantime, terminological usage with respect to the collection of metadata elements progressed independently in the general metadata community. The Metadata Open Forum in Santa Fe in January of the year 2000 provided the first significant interaction between the two communities, at which point they began to explore areas of common interest and mutual enrichment. While the terminologists and other language specialists re-examined the now more mature metadata standards and explored ways to bring their practice more into line with these specifications, the metadata community sought to integrate terminological best practices with respect to the specification of data elements, particularly the definition of data element concepts. In the ensuing years, the two communities have addressed the issue of 'variety control' with respect to their own terminology and practices in an effort to create a level of interoperability and interchangeability between the two areas of practice, yet without attempting total harmonisation, which many feel would be counterproductive given the established traditions on both sides of the discussion.

2 An accessible repository based on a known, common system

2.1 Historical development of a metadata repository for linguistic resources

At the Metadata Open Forum in New York City (2007), Arofan Gregory introduced his discussion of metadata technology by providing his own description of a metadata registry as a 'repository that provides a single point of visibility into and access to resources relevant to a domain, modelled and maintained according to a known, common system'. These basic principles are reflected in efforts within TC 37 to create an MDR for language resources.

The 'known and common system' upon which the TC 37 MDR is based is primarily a set of ISO standards that govern the metamodels and data category specifications that have been elaborated within the TC 37 community. The initial standards governing the data categories and the modelling of terminological entries included:

- ISO 12620:1999, which actually lists and classifies terminological data categories (DC); this version of ISO 12620 constitutes the then full *data category selection (DCS)* identified for use in modelling terminological data, with the understanding that individual designers would identify subsets of this list for use in specific applications.
- ISO 16642:2003, the Terminological Markup Framework (TMF), which provides a metamodel for Terminology Markup Languages.

The TEI-Term interchange format has advanced to become ISO 30042:2008. TermBase eXchange (TBX), via a series of acronymic stages: TIF, MARTIF, SALT, XLT and the LISA TBX standard.[1] Geneter (2007) features an alternate interchange format that utilises the same basic set of data categories. ISO 16642 was developed in order to address issues involving interchangeability and interoperability between the TBX family of data representations and Geneter, which led to the abstract notion of the Terminology Markup Language (TML). Whereas TMF provides for the structure of a TML by specifying a metamodel, it supplies semantics by 'decorating' that model with a vocabulary consisting of a DCS subsetted from the DCR. Data categories included in the vocabulary are constrained by their specific relationships to levels in the metamodel. The actual syntax of such a TML is manifested primarily in its *style*. Examples of such TMLs can be seen on the Geneter/Gentrix and the LISA/TBX websites.

Style in this context involves primarily the choice of xml markup style used in a given environment, whereby, for instance, 'Geneter style' declares major data categories to be Generic Identifiers (GIs), and TBX follows a TEI-inspired approach of declaring a short set of meta-GIs and governs the actual validation of the data categories at the schema level. This procedure is based on an early discussion in the TEI environment that separated most of the data categories into several primary meta-level classes: term, term-related information, descriptive (concept-related) information and administrative information. Together with notes and various types of relations, all the terminological data categories in the 12620:1999 collection can generally be categorised according to these classes. This classification according to meta-metadata classes is based at least in part on Sager's earlier efforts to identify term-specific information, concept-related definitional information and pragmatic contextual information, accompanied by administrative data (Sager, 1990, p.129 ff).

Around the time that 12620 and 16642 were published, however, the scope of TC 37 expanded considerably to include an increasing variety of linguistic resources. Table 1 reflects the expanding number of so-called *Thematic Domain Groups (TDGs)*, i.e. groups of experts responsible for defining data categories used for the representation of various kinds of language resources in this new framework. Comparison of the data category needs for these new TDGs reveals considerable overlap among their DCSs. For instance, many, if not most, use standard grammatical and lexical information (/*part of speech*/, /*grammatical gender*/ and the like), as well as language documentation categories such as /*definition*/, /*source*/, etc. Obviously, the many stakeholders involved in these TDGs (researchers, language communities, translators, cultural heritage curators, terminologists and lexicographers, to name just a few) can benefit from a system that will allow for interchangeability and interoperability among these various resources.

Table 1 ISO Technical Committee 37 Thematic Domain Groups

TDG 1 Metadata
TDG 2 Morphosyntax
TDG 3 Semantic Content Representation
Activity 1 Discourse Relations
Activity 2 Dialogue Acts
Activity 3 Referential Structures and Links
Activity 4 Logico-semantic Relations
Activity 5 Temporal Entities and Relations
Activity 6 Semantic Roles and Argument Structures
TDG 4 Syntax
TDG 6 Language Resource Ontology
TDG 7 Lexicography
TDG 8 Language Codes
TDG 9 Terminology
TDG 11 Multilingual Information Framework (MLIF)
TDG 12 Lexical Resources
TDG 13 Lexical Semantics

Furthermore, realities of the computing environment dictated the ready availability of DCS information in digital form, as well as the need for a global web-based environment for proposing, elaborating, storing and retrieving data categories in an open environment. Thus the decision was made to create the current DCR in the form of a web service (Ide and Romary, 2004; Wright, 2004). This effort has been further supported by additional standards:

- ISO/IEC 11179 family of standards
- ISO 24613:2008, the Lexical Markup Framework (LMF), which provides a complex metamodel with extensions for a variety of language resources, including machine readable dictionaries
- a variety of other NLP-markup standards elaborated in TC 37/SC 4
- ISO 12620:2009, which provides a framework for the development of the `ISOcat` DCR as an open, online repository for data categories classified according to TDGs and subordinate activities (see Table 1), along with procedures for administering a subset of the DCR as a 'standard as database'.

2.2 *Syntax*

The DCR was first developed by LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications) and INRIA (l'Institut National de Recherche en Informatique et en Automatique) and was originally hosted by L'INstitut de l'Information Scientifique et Technique du CNRS (INIST-CNRS) in Nancy, France. Called *SYNTAX*, the online interactive utility has enabled TC 37 experts to elaborate over 1737 data categories, including the core ISO 12620:1999 set for terminology. With this interface, however, the DCR was not as 'visible and accessible' as the community desired, for the service suffered from bandwidth issues and did not provide a user-friendly functional interface. Hence an effort has been made to provide a greater level of accessibility by providing for mirror sites worldwide and enhanced usability of the data in the resource via a clearly defined Data Category Interchange Format (DCIF). In addition to increased user friendliness, the new system will implement the ISO balloting system and add a number of powerful output modalities designed to facilitate actual application of the data categories in the variety of environments cited in Table 1.

2.3 `ISOcat`

The Max Planck Institute for Psycho-Linguistics (MPI) in Nijmegen, The Netherlands, is responsible for redesigning and redeploying the DCR in the new `ISOcat` configuration. MPI is the formal Registration Authority for the DCR. In this context, however, it is important to note that the DCR was originally conceived of as a venue where a community of linguists could develop data categories for use in language resources, including both official Thematic Domain Group specialists appointed by TC 37 working groups and self-declared experts working in a broad range of related research communities.

Further in keeping with explicit MPI policy, the DCR is intended to be a freely and persistently accessible, interactive environment for generating, discussing, and interchanging linguistic metadata. All software used in the context of the DCR will be open-source. MPI is also committed to persistent archiving of data by preserving stable snapshots of the data collection at six month intervals. Older versions will be accessible via the IMDI metadata catalogue. Copies of the `ISOcat` DCR will be maintained at two other computer centres in order to ensure persistent data security and availability. As noted above, future plans include the operation of actual mirror sites in order to provide enhanced broadband functionality for the interactive online interface, and discussion is under way with regard to the creation of batch upload utilities. MPI has been fully integrating in-house applications with the DCR, underscoring its commitment to the resource.

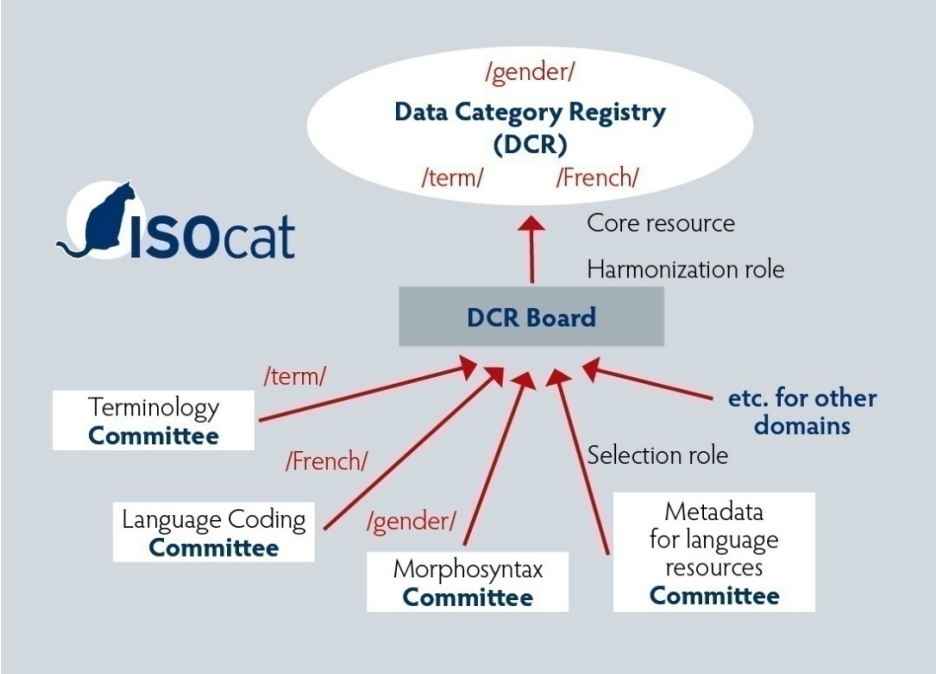
2.4 *Private spaces, public access and official standards*

As indicated above, TC 37 projects are represented by Thematic Domain Groups (TDGs), each with its respective chairperson. In addition to these standards-related groups, any groups or individuals who have registered as `ISOcat` experts can create individual data categories or assemble DCSs in their own private spaces within the DCR. The work conducted in these areas is indeed private and not subject to the control of TC 37. It is also planned that users will eventually be able to download `ISOcat` software to create individual data category registries, which may then be maintained independently or, if desired, batch uploaded to the DCR for sharing, joint editing, and dissemination to other users. It is foreseeable that de facto standardisation may occur within these private work areas.

Despite the openness described here, and the broad accessibility with respect to workspaces and availability of data categories, the system is also being designed to accommodate

the formal standardisation process. TC 37 TDGs, or conceivably, other standards groups, need not only to specify individual data category names and definitions, but also to determine preferred DCSs as standardised subsets of the DCR. The process to be followed for standardisation is described in Section 8 of ISO 12620:2009 in compliance with the ISO Directives, Normative Annex ST: *Procedure for the development and maintenance of standards in database format*. ISO 12620 spells out procedures for data category submission and review, whereby individual experts and TDGs can submit new data category proposals to a DCR Board for consideration as standardised data categories approved for official use by anyone wishing to maintain compliance with specifications of individual TC 37 TDGs. This approach is designed to ensure optimum reusability, interchangeability and interactivity among different data resources. Figure 1 illustrates the configuration of the DCR itself under the administration of the DCR Board. According to the requirements of the ISO policy on standards as databases, this board consists of two functioning units, i.e. an administrative board comprising the TDG chairs, and a balloting board, comprising representatives appointed by each of the P(articipating) members of TC 37.

Figure 1 The DCR and DCR Board, together with examples of Thematic Domain Groups (shown here as committees) (see online version for colours)



Process-related specifications described in ISO 12620 must be implemented pragmatically in the DCR software itself. The steps required for standardisation, i.e. Registration, Submission, Decision and Stewardship, are reflected with clear status records documenting the history of individual data categories and DCSs. Roles, responsibilities and work flow are clearly defined in working procedures and pragmatically implemented in the DCR (see also Figure 3, Section 3.1) When an individual or group originates a new data category, the choice can be made to maintain that data category in a private space or to submit it to a TDG for consideration as a member of that group's DCS. The approval process unfolds in a series of specified steps, involving the possible appointment of additional expert contributors to review the data category, the appointment of judges from the TDG or TDGs potentially interested in adopting the data category, and the consideration and possible revision of the data category before consensus is reached on its inclusion in the DCR. In the event of rejection, it is important for the submitter to be informed of the conditions under which the data category (or possibly the DCS) has been rejected:

1 The data category concept may already exist in the TDG in question in the form of a data category specification associated with a different name from the one that has been proposed. In these cases, its submitters should map their data category to the existing data category for exchange purposes. If the variant data category name is current in applications or working environments, the alternate name can be included as a variant in an instantiation of the Data Element Name class of the existing data category specification (see Section 3.2).

2 The data category concept may already exist in the DCR, but is currently assigned to some other TDG(s). In this case, the submitters should just add their TDG to the profile attribute for the existing data category specification. If name issues exist as noted under Item 1 above, that procedure should be followed in this case as well.

3 The data category is inappropriate for any of the existing TDGs, but it is advisable to create a new TDG/DCS for this and other related data categories.

In this case, experts will be advised to submit a request to the DCR Board for the creation of a new TDG within the framework of the DCR.

4 The data category is inappropriate for the DCR because it appears to be associated with a class of metadata outside the framework of linguistic data. Experts should examine other metadata environments to determine whether their data category actually applies to or derives from other formats (e.g. SKOS, OWL, SOAP, etc.). In some cases it may be desirable to create a crosswalk to another metadata environment. Other formats, such as MathML or LaTeX, etc., can be used in conjunction with linguistic resources, but do not have to be incorporated into the DCR.

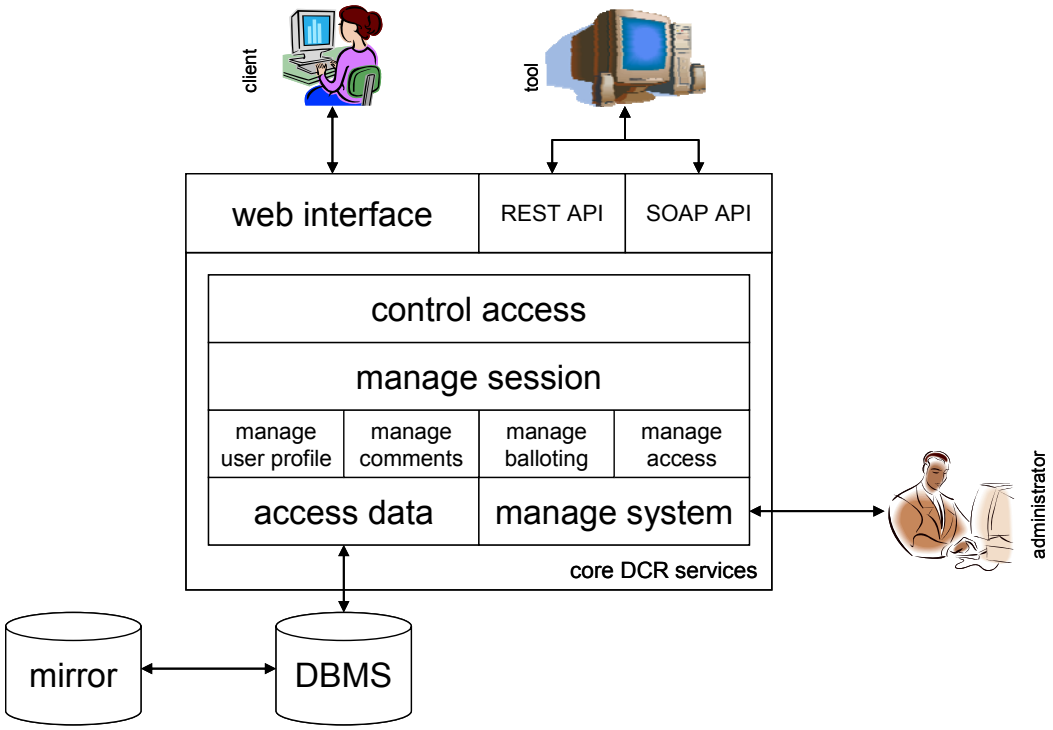
5 The data category appears to be inappropriate to the DCR in its current form and does not fit any of the above conditions. If this is the case, only a significant revision of the submission will make it possible for the data category to be accepted.

It has been proposed that those data categories that come to represent the standardised subset of data categories in the DCR will be archived separately on a biennial basis and submitted to ISO as a 'standard as database'. Although the data categories will also remain freely available in the DCR, ISO will then retain the right to disseminate this standardised 'snapshot' of the database according to common ISO practices.

2.5 System architecture

Figure 2 illustrates the system architecture of `ISOcat`. It shows a layered set of backend modules which provide the core DCR services. Attached to these are frontend modules to provide various ways to communicate with the outside world. Modules dedicated to system management and data storage and retrieval occupy the lowest level. The two top core modules provide access control and session management. In the middle level the main functionality of the DCR is implemented.

Figure 2 Schematic representation of the DCR system architecture and functions (see online version for colours)



The management of user profiles incorporates a wide range of functionalities, from creating and editing a DC to DCS related procedures. The variety of comments types that appear in the data model of the DCR are all handled in a common way by the *Manage Comments* module. The balloting process forms the culmination of the standardisation phase, and this whole process, including the initiation stage (submission, assigning judges, etc.) and the results roll-out stage (acceptation or rejection), is managed by a single module. Access information is attached to various entities in the data model, e.g. DCs and DCSs, and the *Access Management* module 'knows how' to interpret the interactions between these access rules and can thus determine whether a particular user can access a particular DC and with what rights (read-only, comment-only, write access).

One of the frontend modules provides the state-of-the-art web interface for `ISOcat`. Clients can access this interface using a modern browser. For interaction with other applications, the system provides two APIs, each implemented in a different module: a set of RESTful Web services and a set of SOAP/WSDL Web services. Both interfaces will support the functionality to browse the DCR as described by Kemps-Snijders et al. (2006).

To provide high availability of the DCR, mirrors of the registry will be created at other computer centres. The database management system provides the functionality to synchronise these mirrors regularly.

Implementation of this system architecture is currently underway. In its initial milestone release, `ISOcat` provides, based upon the existing SYNTAX database, guest access to the registry both for clients and tool. In the current beta version, a guest or registered user can select a new DCS by accessing existing DCs and export the DCS to DCIF or RDF formats. Registered users can also create and edit new DCs and save them to their own private working space. The standardisation process will then be added in the third milestone release, at which point the core DCR services will be fully supported. In the last phase of the implementation the database mirrors will be created.

All code developed within the `ISOcat` project will be freely available and open source, allowing any individual or group to host its own DCR and even to adapt the code to specific needs. The project itself also uses (commercial) open source software. The state-of-the-art web interface is being developed using TIBCO General Interface (TIBCO Software Inc, 2007). The application server is 1060 NetKernel (1060 Research Ltd, 2007) and hosts the described modules, and the DBMS is PostgreSQL (PostgreSQL Global Development Group, 2007).

3 DCR data model

3.1 The three-part model

The data model for the DCR is specified in Unified Modelling Language (UML), extended where necessary with additional constraints expressed in the Object Constraint Language (OCL). The complete data model is shown in three parts in Figures 3, 4 and 5, where the Data Category class serves as the linking element between the parts. The DCR itself consists of two classes, the Global Information (GI) class shown in Figure 3, and one or more Data Category (DC) specification classes. Further enhancements can be added to the data category specification to describe the conceptual domain of the data category, if relevant, or to describe the names associated with the conceptual domain in a variety of languages.

Figure 3 illustrates primarily the Administration Information Section of the data category specification, which documents procedural functions and roles involved in the standardisation of individual data categories and of thematic domain-specific DCSs. This process was described in detail in the previous section.

Figure 3 UML class diagram for Part 1 of the DCR data model, the Administrative part

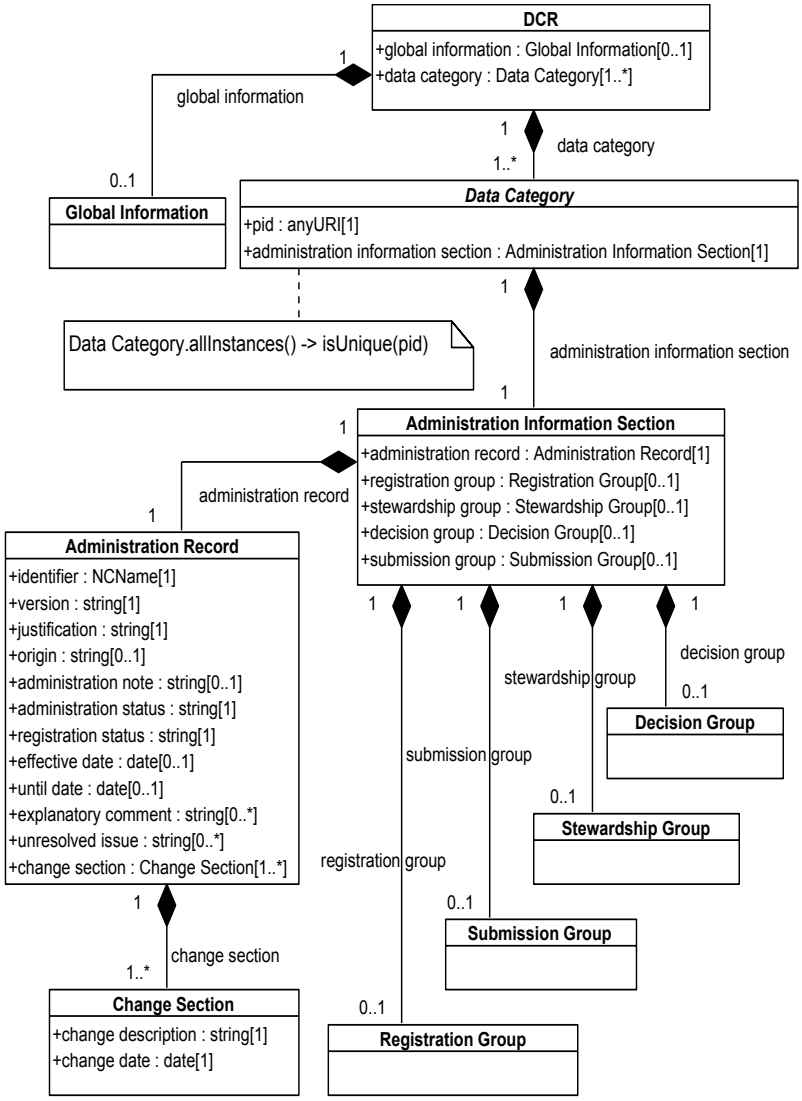


Figure 4 illustrates the Description Section of the data category specification. The description part can be viewed as having both a set of Language Section classes and a set of Data Element Name classes. Each of the Language Section classes documents the data category for a specific working language. For each data category there should always be an English Language Section with at least one definition, at least one Name Section and at least one note. The obligatory note is expected to justify the relevance of the data category to the field of language resources. The Data Element Name classes document the use of a data category in a given database, format or application. The profile attribute in the Description Section declares the thematic domain or domains to which the data category is assigned. Data categories introduced in private work spaces are classed as *Private* by default and do not require the assignment of a thematic domain profile, but as soon as the data category is submitted for inclusion in the standardised component of the DCR, one must be declared.

Figure 5 combines information about the Linguistic Section of the data category specification with detailed treatment of conceptual domain constraints based on differentiation by data category subtype. The fundamental purpose of the Linguistic Section entails the specification of conceptual domain values for a given object language.

Figure 4 UML class diagram for Part 2 of the DCR data model, the Data Category specification class

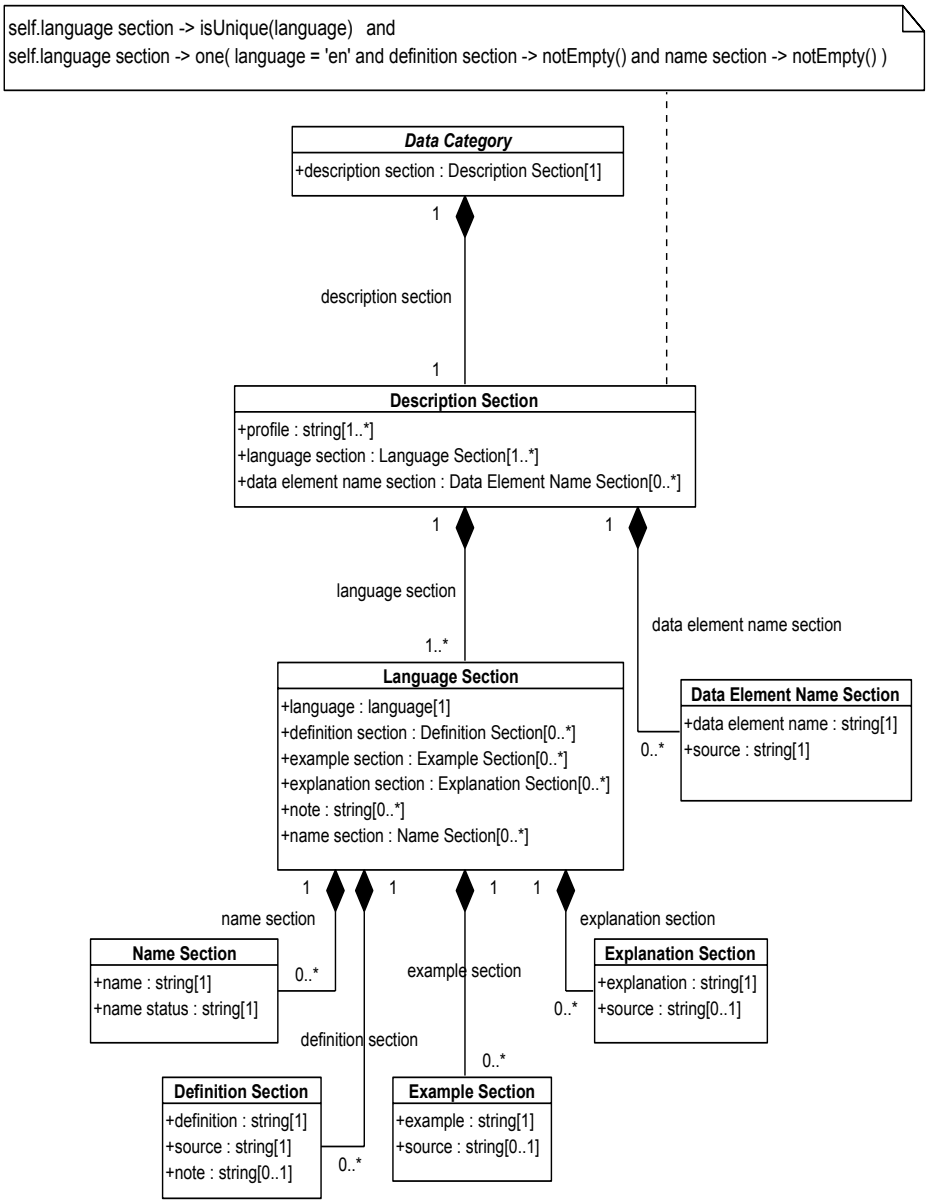
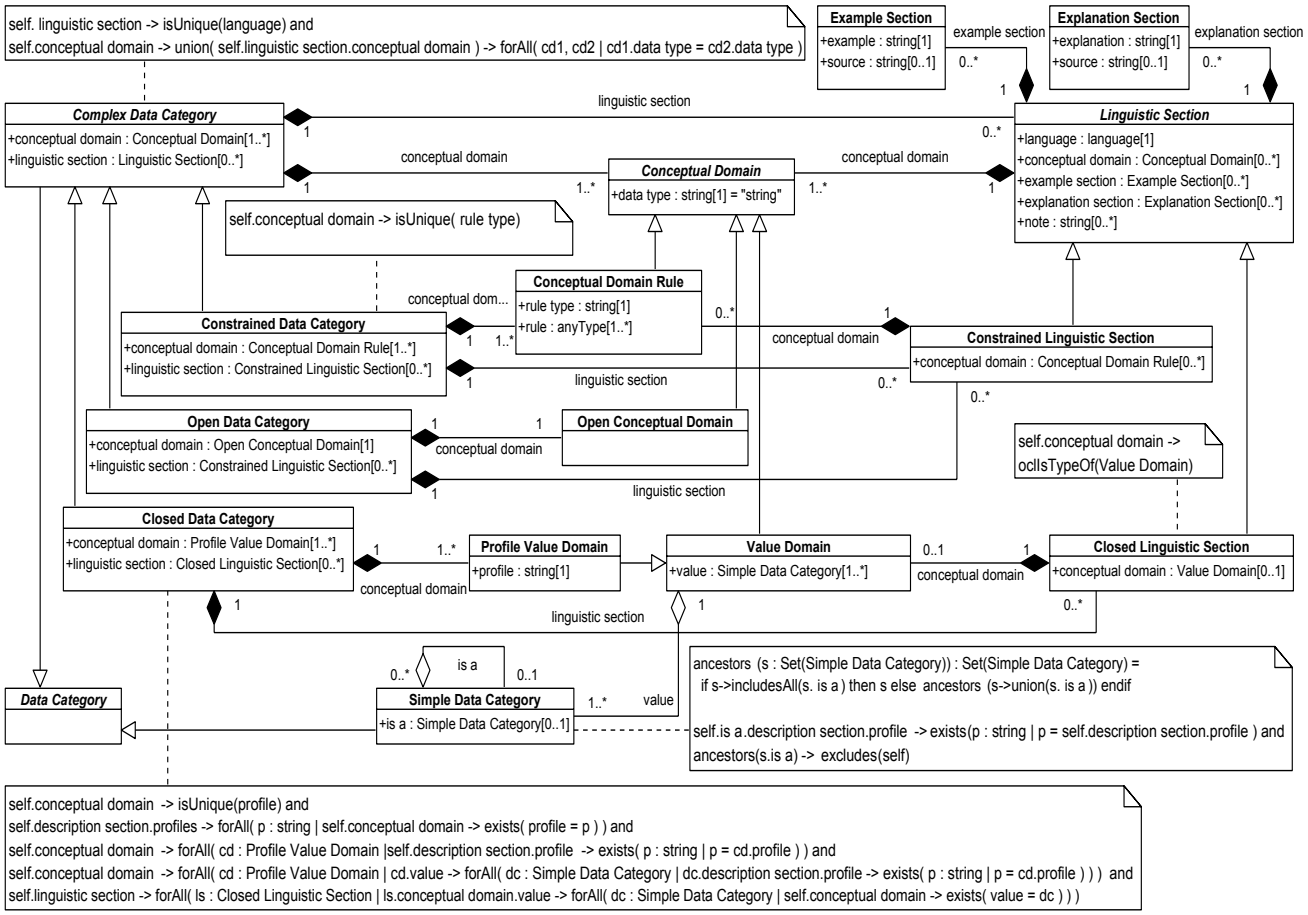


Figure 5 UML class diagram for Part 3 of the DCR data model, the Linguistic Section class



3.2 Data category subtypes

In TC 37 practice, which diverges somewhat from ISO/IEC 11179, data categories are classified as *complex data categories* and *simple data categories*. In terms of data modelling, a complex data category is one that has a *conceptual domain*, which is to say that the complex data category can function as the label for a container (e.g. a field in a data resource) that is constrained in some way. All complex data categories are hence subject to constraint, but with graduated degrees of restriction. *Open Data Categories* are constrained by the characteristics set down in their definitions, but their conceptual domains are not restricted to an enumerated set of values. In other words, they are restricted by constraints other than a value range. As an example, the *Open Data* category */term/* is constrained by its definition from properly containing, e.g. contextual information or grammatical information, but it would be impossible to enumerate all values that might in some subject field be designated as a *term*. Needless to say, automatic validation of the content of *Open Data Categories* is generally impossible, although some items, such as rigorous definitions, can be processed for certain characteristics and utilised for automatic generation of other resources, such as ontologies.

Closed Data Categories, in contrast, possess conceptual domains that can be limited to an enumerated set of picklist values. Called *value domains*, these sets of permissible instances can be specified by either a Thematic Domain Group or by the designers of a private application. Up to this point, the data category types listed here reflect the same conventions specified by the ISO/IEC 11179 family of standards. Beyond these specifications, TC 37 also expressly treats enumerated values as so-called *simple data categories* that are dependent on their associated closed data category. As defined in ISO 12620, a simple data category does not itself *have* a conceptual domain, but rather *is* a member of a value domain. In conceptual terms, the value domain usually comprises the extension of the data element concept represented by its closed data category. Individual DCs are generally associated with one single closed data category concept, with the caveat that when different DC owners create multiple versions of a data category specification (e.g. */part of speech/* for morphosyntax and a second version for */part of speech/* for terminology), this results in a situation whereby a simple DC (e.g. */noun/*) can be shared by multiple closed DCs of the same fundamental class. Discussion is still underway to resolve issues surrounding the notion of more than one parent for a DC or set of value domain DCs.

Closer analysis has led to the definition of a further class of complex data categories that have been singled out as *constrained data categories,* for which the conceptual domain cannot be expressed as an enumeration, but rather is expressed in some schema-specific constraint. The schema type for such a *schema-specific domain* identifies the schema used for specifying the constraint. Examples of this type of data category would be to specify a constraint in W3C XML Schema or, for instance, RelaxNG, stating that a date field should contain only values after a specific date or that values must fall within a certain range.

3.3 Data category names and linguistic variants

Individual data category specifications are associated with several kinds of identifiers. To achieve interoperability between various resources, their metadata should be able to indicate which data categories were used. This means that resource metadata should include references to the specific categories in the DCR, so that the DCs in the subset (DCS) can be identified. These references should be represented as globally unique, location independent and persistent identifiers (PIDs) that enable lookup in the DCR of the data categories they represent. Examples of persistent identifiers that are currently commonly used include Digital Object Identifiers (DOI; DOI, 2001), handles (Sun et al., 2003), Archival Resource Keys (ARK; Kunze and Rodgers, 2007) and Uniform Resource Names (URN; Moats, 1997).

The DCR uses cool URIs (Berners-Lee, 1998), i.e. a stable URI scheme, to provide persistent data category references. To achieve this persistence, ownership of the internet domain, isocat.org, is bound to the Registration Authority of ISO 12620, the standard describing the DCR. This means that although the Registration Authority controlling the DCR may change over time, the DCR will be hosted with the same URI scheme at the same internet location.

Data category URLs take the following form: http://www.isocat.org/datcat/ISO-DC-1345. The prefix of these URLs, http://www.isocat.org/datcat/, is the location of the DCR resolver and the suffix, ISO-DC-1345, the unique identifier of a specific data category. The Data Category Interchange Format (DCIF) is the default representation used for the data category specification returned by the resolver. HTTP content negotiation can be used to request other representations, e.g. HTML or RDF.

These references can be embedded in the metadata of a linguistic resource. For example a small section of the TBX XCS (TermBase eXchange eXtensible Constraint Specification) for master data category selection (based on Annex B.2 in (ISO 30042: 2008)), might look something like the following. Note that the ellipses (…) indicate places where parts of the XML document have been omitted.

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <!DOCTYPE TBXXCS SYSTEM "tbxxcsdtd.dtd">
3 <TBXXCS name='master' version="0.4" lang="en">
    …
4  <datCatSet>
    …
5    <termNoteSpec name="animacy" datcatId="hdl:42/
     DC-1902">
6      <contents datatype="picklist"
       targetclass="none" forTermComp="yes">
7        <termNoteSpec name="animate" datcatId=
         "http://www.isocat.org/datcat/DC-1911"/>
8        <termNoteSpec name="inanimate" datcatId=
         "http://www.isocat.org/datcat/DC-1952"/>
9        <termNoteSpec name="otherAnimacy" datcatId=
         "http://www.isocat.org/datcat/DC-1953"/>
10       </contents>
11     </termNoteSpec>
    …
12   </datCatSet>
   …
13 </TBXXCS>
```

The `datcatId` attribute values contain the handles which refer back to the DCR. The TBX XCS DTD already declares the `datcatID` attribute, which is used to store handles to complex data categories. However, the current version of the handles system cannot incorporate the references to simple data categories, as the picklist is implemented as a space-separated sequence of values. The example fixes this by also using the `termNoteSpec` element for simple data categories.

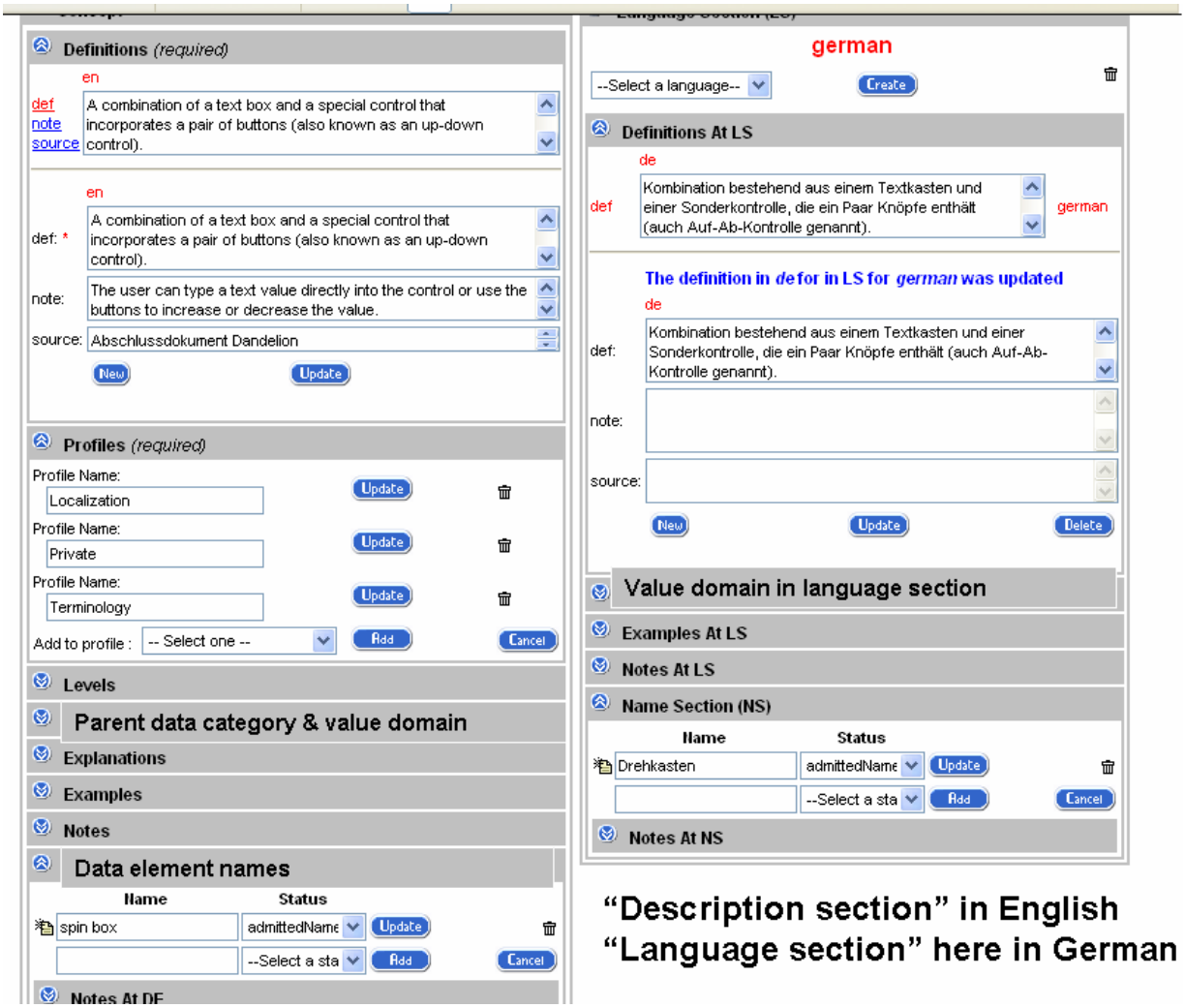
For human users, data categories are also assigned meaningful names, frequently names that represent best practice naming conventions in linguistic resources. These names are officially assigned in English, as noted above, and are viewed as international in character. They are invariant for purposes of data exchange and interoperability, but in actual practice, such data category names are subject to many of the same variations that compromise the semantic content of terms in human language. Although the data category names are generally stable inside individual environments, variation between individual applications or across thematic domain boundaries tends to be significant at times. Although use of standardised names can be encouraged, different developers, working groups, projects and communities of practice frequently have to deal with variant legacy data or well-established traditions that militate against the adoption of standardised names. As a consequence, applications frequently differ in their interface usage, but declare mapping procedures in order to facilitate data compatibility. Furthermore, language resource developers working with multiple languages or focusing on a single local language have a need to specify preferred or standardised forms of the data category name and other information (definitions, notes) in other languages.

In the SYNTAX version of the DCR, the data category name and synonyms or aliases in English have been entered in a portion of the data category specification called the *Description Section*. Parallel information in multiple languages can be added in the *Language Section* as shown in Figure 6. Considerable discussion has arisen in TC 37 concerning the need to distinguish usage for both the *working language* of an application environment and the *object language* (the immediate language under discussion). Figure 6 has been modified from the model used in SYNTAX in order to accommodate a more nuanced and logical treatment of the data structure.

As a consequence of describing the Conceptual Domain for a given object language in the `ISOcat` model, a complex data category contains a *Linguistic Section*, describing the conceptual domains for a specific object language, whereas a simple data category does not, because, by definition, it has no conceptual domain, as shown in Figure 5. This approach follows the definition very closely and yields a much cleaner model than SYNTAX did, which makes the distinction more intuitive.

At the Linguistic Section level, the distinction between an *Open Data Category*, a *Constrained Data Category* and a *Closed Data Category* is modelled explicitly to distinguish between the various types of conceptual domains that exist for these data categories, i.e. *Open Conceptual Domain, Schema Specific Domain* and *Value Domain.* An open conceptual domain is only restricted to the type of data that may appear as a value. A schema specific domain is bound by a constraint specification in a given schema language. A value domain is defined here as a conceptual domain which consists of a value range.

Figure 6 SYNTAX -based Definition, Data element names and Language Sections; see the text and `ISOcat` diagrams for changes in the model from SYNTAX to `ISOcat` (see online version for colours)



3.4 Data Category Selection (DCS)

Much attention has already been paid to Data Category Selections (DCS). The following discussion is intended to clarify the technical functionalities involved in making and maintaining DCSs within the DCR.

3.4.1 Purpose of a DCS

A Data Category Selection (DCS) is a collection of Data Categories (DCs) which for some system-internal or user-specific reason belong together. ISO 12620: 2009 defines a DCS as follows:

> set of **data categories** selected from the **DCR**
>
> NOTE A DCS can represent the data categories used within a **thematic domain** or a selection of data categories used for a specific application or project. In the latter case, the DCS may draw data categories from more than one thematic domain.
>
> NOTE A DCS can be expressed as a simple list of data categories, or it can be output in a form that contains the entire content of their associated data category specifications, thus incorporating the full set of constraints associated with the DCS. It can also be expressed using a schema notation such as XML schema or RelaxNG, which also comprises the list of data categories together with their associated constraints.

The second note describes mainly options available for the serialisation of a DCS. Internally the DCS data model maintains context information required for users to be able to restrict value domains to their specific needs.

DCSs are the major navigation or action instrument users have within the DCR, as each action they perform results in the creation of sets of DCs, i.e. a DCS, either through the selection of existing DCs, their modification, or the addition of new DCs. Furthermore, each DC specification includes a *Profile* field denoting the TDGs with which the DC is associated, although there is no mechanism for tracking the use of DCs in individual applications or private workspaces.

3.4.2 The SYNTAX implementation

SYNTAX featured three types of DCSs:

1 *Private:* identified a DCS (in some case as well as individual DCs) and stored the DCS a user creates privately.

2 *Public:* stored information on the thematic domain group(s) (TDG) to which a DC belongs, indicated by the *profile* information associated with a DC.

3 *Shared*: stored information on any DCS that users (creators) had marked as shared with other (expert) users.

3.4.3 The `ISOcat` implementation

The new implementation of the DCR features two internal types of DCS, whereby they both adhere to the ISO 12620 definition of a DCS, but each has different implementation strategies.

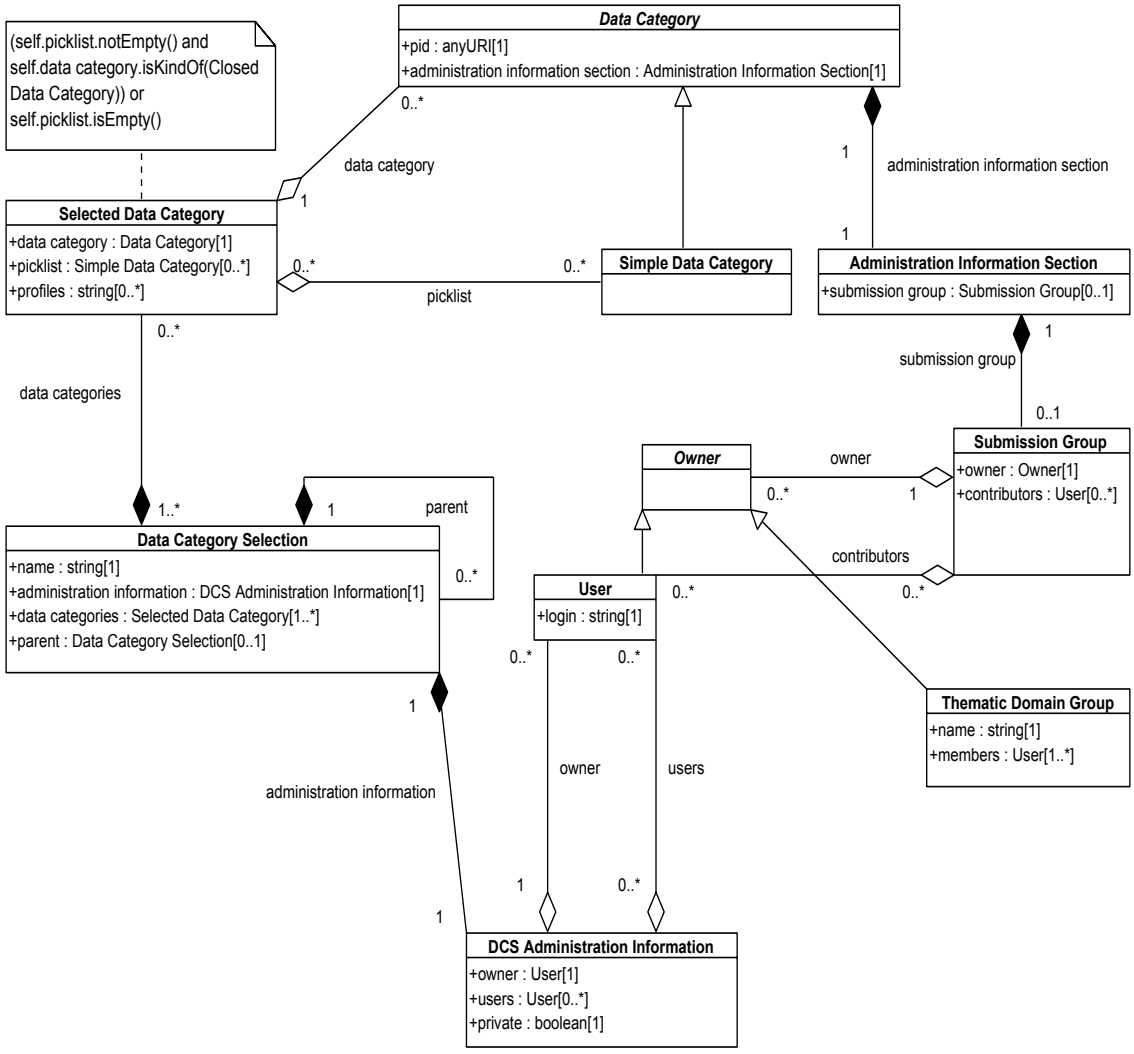
1 *Implicit DCSs*: The first type is analogous to the TDG related DCSs in SYNTAX: they are basically assembled as the result of a query on the properties of a DC, i.e. the profile information. These DCSs are dynamic: they change when relevant content in the DCR changes. For implementation this means that when users request such a DCS, the (implicit) query associated with it is (re)run against the complete set of DCs in the DCR. A consequence of this dynamic nature is that such implicit DCSs are inherently read-only, i.e. a DC can only be added or removed from this kind of DCS by changing the properties of the DC or by changing the query that generated the DCS.

2 *Explicit DCSs*: The second type of DCS is explicitly created by the users of the system. Access information can be attached to this kind of DCS, making it possible to share these DCSs with other `ISOcat` users. Information about this type of DCS needs to be persistently stored in the database. Notice that `ISOcat` only has one explicit DCS type, instead of the two provided by SYNTAX.

3.4.4 The DCS data model

The data model (see Figure 7) as described in this section only applies to explicit DCSs. This leads to the constraint that *owners* can only be *users*, as TDG-owned DCSs are always implicit because they are the result of queries on the DC-related profile information.

Figure 7 UML class diagram for the Data Category Selection



Each explicit DCS has Administration Information associated with it, which provides information about which user owns the DCS, whether the DCS is private (the attribute */private/* is set to */true/*), shared (the attribute */private/* is set to */true/*, and there are additional users) or public (the attribute */private/* is set to */false/*). Notice that write access is handled in the Submission Group associated with a Data Category. The */contributors/* attribute in that class lists the users, next to the owner, who can edit the DC.

DCSs can be nested in each other, e.g. like the well-known directory structures in a file system. The nesting should result in a hierarchy, which is to say that cycles are not allowed. A DCS consists of the union of all DCs in its descendant DCSs, i.e. when serialising the DCS, the nesting is lost. For example, the TBX (Termbase eXchange) format declares a specified DCS listing the DCs that are permissible values compliant with the TBX schema-like functionality; individual user groups, however, can subset (or in some cases even superset) this selection in order to establish their own application or environment-specific DCS, such as the Localisation Industry Standards Association's TBX-Basic subset (LISA 2007).

Depending on the (implicit) DCS from which a DC is selected and added to the DCS under consideration, different subsets of the conceptual domain may play a role. To keep track of this phenomenon, the DC is wrapped in the Selected Data Category class, which includes a reference to the DC, keeps track of which profiles are involved and, in the case of a closed data category, of which simple data categories have been selected for this DCS. The individual TDG or application specific profiles may already limit the value domain to a subset of the standard value domain, but users can also delete some simple data categories from the DCS, thus limiting their own value domain even further. For instance, the closed data category */termType/* provides for a long list of possible simple data categories: */entryTerm/ /synonym/ /internationalScientificTerm/ /fullForm/ /transcribedForm/ /symbol/ /formula/ /equation/ /logicalExpression/ /commonName/ /abbreviation/ /variant/ /shortFormOfTerm/ /transliteratedForm/ /sku/ /partNumber/ /phraseologicalUnit/ /synonymousPhrase/ /standardText/ /string/ /internationalism/*, but one application might choose to limit this list to: */synonym/ /fullForm/ /symbol/ /abbreviation/ /variant/*.

3.5 Customisation of output formats

`ISOcat` supports a standard output format to exchange information. The Data Category Interchange Format (DCIF) specified in (ISO 12620:2009) is used to exchange (parts of) the Data Category Registry within TC37 and to external applications. The SYNTAX DCR was originally intended to use the 16642 Generic Mapping Tool (GMT), which is a high-level mapping tool for use with TMLs, as a standard means for exchanging data category information. In the meantime, developers have come to the realisation that this standard exchange format does not meet the requirements for expressing DCSs for specific application domains, especially in cases where the status of *Open* and *Closed Data Categories* has been altered or where value domains have been subsetted as described in the previous section. Furthermore, data category specifications do not necessarily mirror terminological entries in all respects.

The TBX sample fragment presented earlier shows a `datCatSet` which expresses a DCS in a TBX-specific manner. For communities that have such specific needs regarding the DCS output, `ISOcat` provides the option to add a style sheet to a DCS which transforms the DCIF to their own formats. However, any constraints that cannot be expressed in a style sheet will have to be implemented in the specific application domain and are not handled by the `ISOcat` software. Furthermore, owners of a DCS must be willing to create these style sheets and be willing to share them with other users of the DCS. Also, since `ISOcat` will simply perform the transformation as embodied in the style sheet, the validity of such a resulting DCS cannot be guaranteed within the DCR environment. It is, for example, possible to transform a *Closed Data Category* into an *Open* one or vice versa. In case a DCS is managed by a TDG, it is expected that the specification of the style sheets should reflect the original data category specification.

3.6 Ontology infrastructures vis-à-vis the DCR

3.6.1 Isolating separate functions

It is widely understood that one of the major challenges in our time is to achieve semantic interoperability at a number of levels based on the development of various ontologies and knowledge representation structures. In developing the DCR, we are presented with the temptation to incorporate relational information into the system – indeed, the SYNTAX implementation did provide for the specification of a 'broader concept generic' for each data category, and this functionality is carried over into `ISOcat` with a provision for specifying a single generic '*isA*' relation between any simple data category and another simple data category that is perceived as its broader concept (e.g. */common noun/* is a */noun/*. (It should be noted that such simple binary relations, as well as those between closed data categories and their respective permissible instances, are not construed to be full concept systems in the TC 37 environment.)

There is, however, no provision for referencing multiple systems or hierarchies. Given the potential for multiple hierarchies involving the same members, such simple binary relations is not particularly useful for building more complex systems. The original ISO 12620:1999 was configured to reflect one possible systematic order, focusing, of course,

strictly on terminological data, but there is no total consensus on this particular order, and various experts have suggested multiple relational schemes just for the terminological DCS alone, depending on the specific needs of and viewpoints reflected in individual environments.

Given this experience, the decision has been made to separate the concept definition activity associated with specifying data categories inside the DCR from efforts to establish relations among the categories. (Major contributors to this discussion include TC 37 colleagues Hasida Koiti, Nicoletta Calzolari, Laurent Romary, Andreas Witt, Gerhard Budin and Daan Broeder, in addition to the authors of this article.) Defining domain specific concepts that are used for tagging purposes so that they will be widely accepted by communities of practice as a reference for semantic knowledge is already a very difficult task, but as has been shown within TC 37, the task seems to be manageable. Although it would not be impossible to build in modalities for creating complex hierarchies, this would complicate the data model and exceed the original mandate of the DCR. Hence the decision has been made to move the creation of relational systems outside the context of the DCR.

3.6.2 Using standardised data categories to create federated repositories

The interactive functionality implied in the philosophy and technical organisation of the DCR is that by referencing the same entry or entries in the DCR, creators of multiple lexical schemas or other language resources will be able to search across the multiple lexica without additional effort by informing search engines to exploit the reference contained in the schemas. This scenario is further complicated, however, when there is a need for data resources to interact where one or both have defined additional data categories not included in the common core. It has been proposed that these inconsistencies can be ameliorated by inserting a relation such as 'schema_element_X' is_subclass_of 'datcat_Y'. This relation would typically be stored in a light-weight ontology external to the DCR, here called a 'Relation Registry' (RR). Again, if the search engine is informed that it should make use of the lexical schemas as well as Relation Registries, it could carry out search operations on multiple lexica instantiated by the schemas, as illustrated in Figure 8. Expanding on this simple model, it is foreseeable that more complex environments could evolve on a step-by-step basis whereby federated systems could eventually be designed to reference multiple DCRs and RRs. This procedure could be applied for all schemas that define structured language resources such as metadata, annotations, knowledge spaces, etc.

3.6.3 Relation extensions and knowledge spaces

These simple RRs need to be distinguished from true ontologies which form Knowledge Spaces that lend themselves for inferencing, i.e. they need to include definitions eventually extracted from DCRs and they will include relations, properties, etc. to form logically complete systems. While RRs could be stored in simply structured XML files, Knowledge Spaces need to be represented in knowledge representation schemas such as RDF, SKOS, etc. Of course, it needs to be ensured that relation types to be used in RRs are compliant with types found in RDF-S, OWL, etc. It is up to smart tools to extract definitions and relations to form Knowledge Spaces (see Figure 9). Making this choice to separate definitions and relations allows users to generate several sets of relations that even may include conflicting knowledge. The first layer ontologies could also be called 'Relation Registries', since they store simple relation triples such that they can easily be used by search engines and manipulated easily by simple editors.

Figure 8 Schematic diagram showing the interaction of DCR information with Relation Registries and application-specific schemas in support of federated search behaviour (see online version for colours)

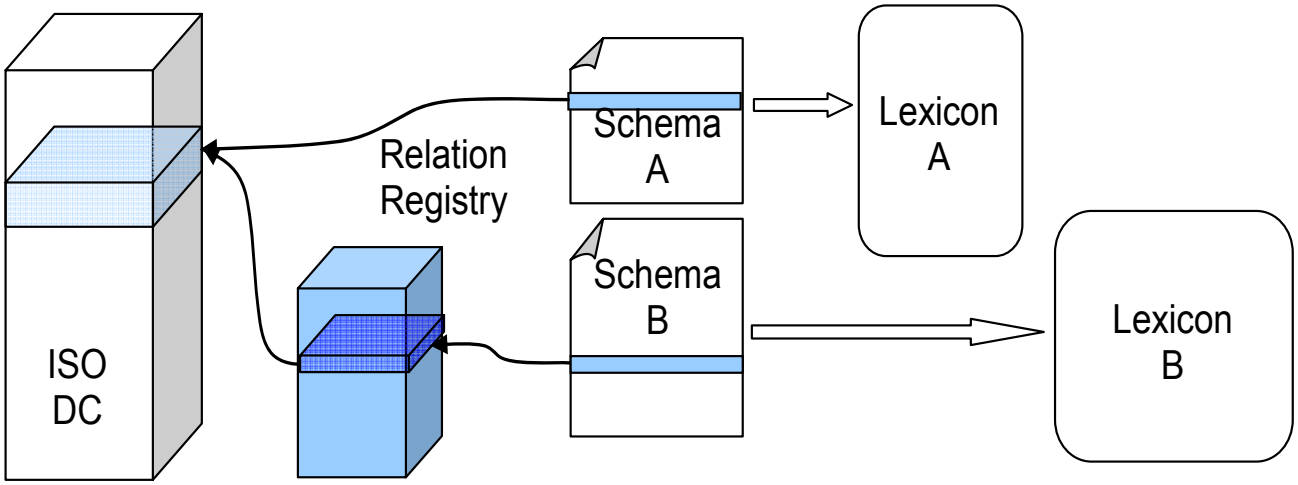
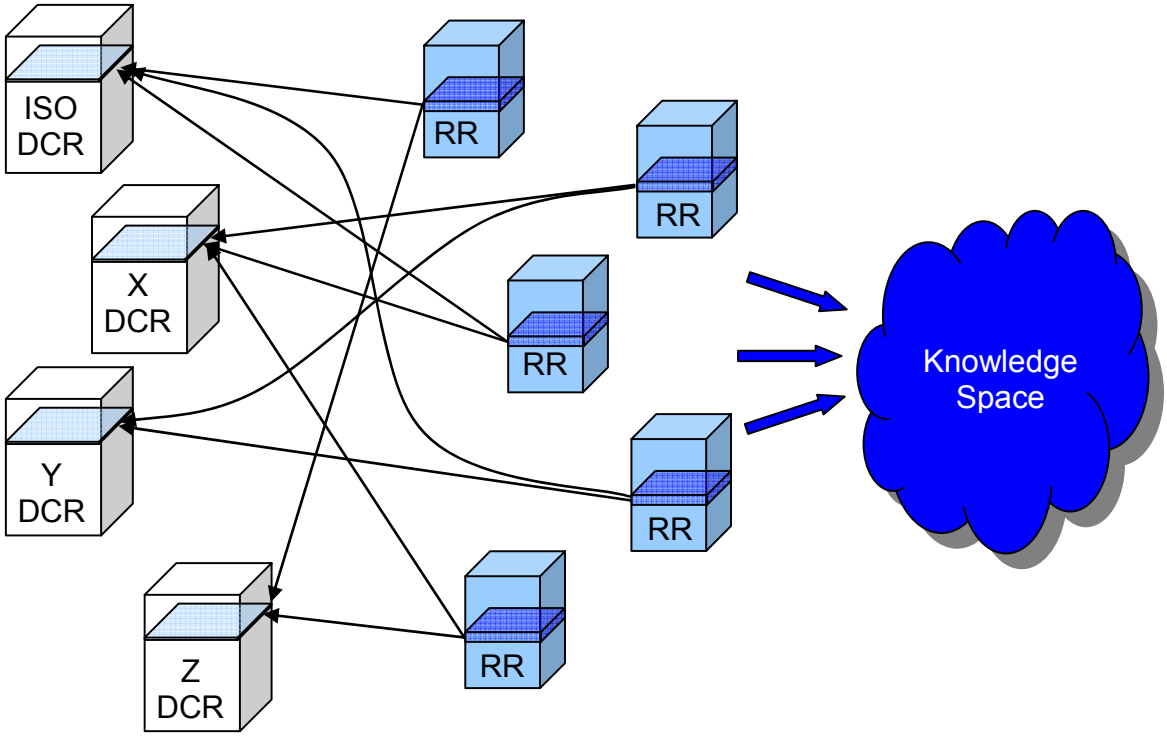


Figure 9 Schematic diagram showing the interaction of multiple DCRs with light-weight relation registries designed to support an ontological knowledge space (see online version for colours)



3.7 Web service support

References to data categories can be embedded manually in the metadata of linguistic resources using SYNTAX or `ISOcat`. For applications that provide user-specific data models, however, a web service API is available. This API allows direct lookup and extraction of data category information (Kemps-Snijders et al., 2006). The web services implemented by SYNTAX use a REST interface (Richardson and Ruby, 2007). The `ISOcat` system will also support a SOAP/WSDL (Chinnici et al., 2007) variant.

Lexus (Kemps-Snijders and Wittenburg, 2006), for example, implements the *Lexical Markup Framework* (Francopoulo et al., 2006). This core model can be fleshed out by a user in various ways, one of which involves interaction with the DCR using the SYNTAX web services API and the selection of relevant data categories.

4 Outlook

We have attempted to address the full range of issues involved in the collection of data categories for language resources, including both the technical aspects of the web-based resource itself and the administrative procedures necessary to maintain a dual purpose repository that is both an open-source environment for the elaboration of metadata and the focal source for an ISO 'standard as database'. We anticipate that the creation of this asset will not only facilitate the dissemination of standardised data categories and data category selections, but will also contribute to higher levels of interoperability and interactivity generated by interaction with differently weighted external ontological resources. These goals are dependent on the following criteria:

- the underlying model must remain comparatively simple
- the concept definitions in the DCR must be excellent and widely accepted
- the DCR infrastructure must be flexible and simple to extend so that research groups and individuals can create their own concept domains
- consensus must be reached concerning ontology mechanisms for relating concepts that are used in different DCRs and resource schemas
- more tools need to support the `ISOcat` API
- an API must be developed for accessing the RR.

The intention over the course of the next few years is to address these multiple issues under the auspices of ISO and within the framework of the EU-sponsored Common Language Resources and Technology Infrastructure project (CLARIN), whose mission is to create an infrastructure which makes language resources (annotated recordings, texts, lexica, ontologies) and technology (speech recognisers, lemmatisers, parsers, summarisers, information extractors) available and readily usable to scholars of all disciplines, in particular the humanities and social sciences.

References

1060 Research Ltd (2007) *1060 NetKernel*. Available online at: http://www.1060.org/

Berners-Lee, T. (1998) *Cool URIs don't change*. Available online at: http://www.w3.org/Provider/Style/URI

Budin, G., Melby, A.K., Shreve, G. and Wright, S.E. (1994) 'Chapter 13: Terminological Databases' TEI P3', in Sperberg-McQueen, C.M. and Burnard, L. (Eds.): *Guidelines for the Encoding and Interchange of Machine Readable Texts*, Text Encoding Initiative, Chicago & Oxford, pp.371–390.

Budin, G. and Wright, S.E. (1994) 'Data elements in terminological entries: an empirical study', *Terminology*, Vol. I, No. 1, pp.41–59.

Chinnici, R., Moreau, J.J., Ryman, C.A. and Weerawarana, S. (2007) *Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language*, W3C. Available online at: http://www.w3.org/TR/wsdl20/

CLARIN (2007) *Common Language Resources and Technology Infrastructure*. Available online at: http://www.cs.kuleuven.be/~liir/projects.php?project=168

Corporation for National Research Initiatives (2007) *The Handle System*. Available online at: http://www.handle.net/

DOI (2001) 'The Digital Object Identifier (DOI) System', International DOI Foundation 2001; Sun, S., Lannom, L. and Boesch, B. *Handle System Overview*, Internet Engineering Task Force, 2003. Available online at: http://www.ietf.org/rfc/rfc3650.txt

Felber, H. (1984) *Terminology Manual*, UNESCO & Infoterm, Paris.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. and Soria, C. (2006) 'Lexical Markup Framework (LMF)', *Language Resources and Evaluation*, Genoa, Italy.

Geneter (2007) *TermBridge Semantic Repository*. Available online at: http://www.geneter.org/

Ide, N. and Romary, L. (2004) 'A registry of standard data categories for linguistic annotation', *Proceedings of the IVth LREC International Conference on Language Resources and Evaluation*, Lisboa, Portugal, pp.135–138.

Handle (2008) *The Handle System®*. Available online at: http://www.handle.net/

ISO (2007) *ISO Directives Annex ST (normative): Procedure for the development and maintenance of standards in database format*, ISO, Geneva.

ISO 1087-2 (2000) *Terminology Work – Vocabulary – Part 2:Computer Applications*, ISO, Geneva.

ISO 6156:1987. *Magnetic tape exchange format for terminological/lexicographical records (MATER)*, ISO, Geneva.

ISO/IEC 11179-1 (2004) *Information Technology Vocabulary – Metadata Registries (MDR) – Part 1: Framework (Draft)*, ISO, Geneva.

ISO/IEC 11179-2 (2002) *Information Technology – Specification and Standardization of Data Elements – Part 2: Classification for Data Elements*, ISO, Geneva.

ISO/IEC 11179-3 (2003) *Information Technology Vocabulary – Metadata Registries (MDR—Part 3: Registry Metamodel and Basic Attributes*, ISO, Geneva.

ISO/IEC 11179-4 (2004) *Information Technology Vocabulary – Metadata Registries (MDR—Part 4: Formulation of Data Definitions*, ISO, Geneva.

ISO/IEC 11179-5 (2005) *Information Technology Vocabulary – Metadata Registries (MDR—Part 5: Naming and Identification Principles*, ISO, Geneva.

ISO/IEC 11179-6 (2004) *Information Technology Vocabulary – Metadata Registries (MDR—Part 6: Registration*, ISO, Geneva.

ISO 12200 (1999) *Computer Applications in Terminology – Machine-readable Terminology Interchange Format (MARTIF) – Negotiated Interchange,* ISO, Geneva.

ISO 12620 (1999) *Computer Applications in Terminology – Data CATEGORIES,* ISO, Geneva.

ISO DIS 12620 (2008) *Computer Applications in Terminology – Data Categories – Specification of Data Categories and Management of a Data Category Registry for Language Resources*, ISO, Geneva.

ISO 16642 (2003) *Computer Applications in Terminology – TMF (Terminological Markup Framework)*, ISO, Geneva.

ISO DIS 24613 (2006) *Language Resource Management – Lexical Markup Framework (LMF),* ISO, Geneva.

ISO DIS 30042 (2007) *Computer Applications in Terminology – TermBase eXchange (TBX) Format Specification*, ISO, Geneva.

ISOcat (2008) Available online at: http://www.isocat.org/

Kemps-Snijders, M., Ducret, J., Romary, L. and Wittenburg, P. (2006) 'An API for Accessing the ISO Data Category Registry', Genoa, Italy.

Kemps-Snijders, M. and Wittenburg, P. (2006) 'LEXUS – a web-based tool for manipulating lexical resources', *Language Resources and Evaluation*, Genoa, Italy.

Kunze, J. and Rodgers, R.P.C. (2007, July) *ARK Persistent Identifier Scheme* (Internet Draft, updated). Available online at: http://www.ietf.org/internet-drafts/draft-kunze-ark-14.txt

LaTeX. *LaTeX – A Document Preparation System*. Available online at: http://www.latex-project.org/

LISA/OSCAR (2007) *SIG News & Updates: TBX-Basic Data Category Specification*. Available online at: http://www.lisa.org/sigs/terminology/

Melby, A. (1991) 'MicroMATER: A proposed standard format for exchanging lexical/terminological data files', *Meta: La terminologie dans le monde: orientations et recherches*, Vol. 36, No. 1, pp.136–159.

Moats, T. (1997) *URN Syntax*, in IETF RFC 2141: Internet Engineering Task Force, 1997. Available online at: http://www.ietf.org/rfc/rfc2141.txt

OMG (Object Modelling Group) (2008) *Unified Modelling Language (UML®)*. Available online at: http://www.uml.org/#UML2.0

OMG (Object Modelling Group) (2008) *UML 2.0 OCL (Object Constraint Language) Specification*. Available online at: http://www.omg.org/docs/ptc/03-10-14.pdf

PostgreSQL Global Development Group (2007) *PostgreSQL*. Available online at: http://www.postgresql.org/

Richardson, L. and Ruby, S. (2007) *RESTful Web Services*, O'Reilly, Cambridge, MA.

Sager, J.C. (1990) *A Practical Course in Terminology Processing*, John Benjamins Publishing Company, Amsterdam and Philadelphia.

SYNTAX. Available online at: http://syntax.inist.fr/

TIBCO Software Inc (2007) *Tibco General Interface*. Available online at: http://www.tibco.com/devnet/gi/

Wright, S.E. (2004) 'A global data category registry for interoperable language resources', *Proceedings of the IVth LREC International Conference on Language Resources and Evaluation*', Lisboa, Portugal, pp.123–126.

Note

1 TIF = Terminology Interchange Format; MARTIF = Machine Readable Terminology Interchange Format; SALT = Standards-based Access to multilingual Lexicons and Terminologies; XLT = XML interchange format refined during the SALT project; LISA = Localisation Industry Standards Association.