

Putting Data Categories in their Semantic Context

Marc Kemps-Snijders^a, Menzo Windhouwer^a, Sue Ellen Wright^b

^aMax-Planck-Institute for Psycholinguistics, ^bKent State University

{Marc.Kemps-Snijders, Menzo.Windhouwer}@mpi.nl, sellenwright@gmail.com

Abstract

The TC 37 Data Category Registry (DCR; www.isocat.org) specifies names, authoritative definitions, and other information and constraints for data categories used in a wide range of linguistic resources. Data category selections subsetted and exported from the DCR in the Data Category Interchange Format can be used as the basis for configuring diverse applications. Furthermore, authoritative standardized data category definitions can contribute trustworthy semantic content for the creation of Relation Registries in the extended environment of the DCR in support of external ontologies and other semantic web resources. Resources that reference DCR specifications will require annotation using unique, location-independent, persistent identifiers, and procedures must be established for maintaining ongoing coordination of external resources referencing the dynamic DCR. Developers are currently exploring approaches to data category modeling in RDF(S) and OWL-DL and plotting navigation strategies for traversing a network of data category and relation registries, as well as linguistic resources.

1. Introduction

Linguistic resources are valuable for many stakeholders, e.g. researchers, language communities, translators, and cultural heritage curators. These resources are typically very heterogeneous with respect to structure and semantic encoding, which limits interoperability with respect to search, comparison and merging. To enable these stakeholders to benefit from the wealth of resources worldwide, these resources have to be interoperable to some degree. However, to achieve true interoperability of resources from heterogeneous sources, many variations on different levels have to be addressed. In ISO Technical Committee 37 (TC 37), *Terminology and other language and content resources*, a metadata registry,

called the *Data Category Registry* (DCR), has been developed that will provide a reusable set of (standardized) data categories [1]. Parts of the (meta)data model for a language resource can include these data categories, and thus share common semantics with other resources. Although the sharing of data categories addresses only one level of interoperability problems, this level is close to the core data and promises to provide a solid base for alignment on higher levels. For example, domain ontologies could be built bottom-up or middle-out based on the standardized semantics provided by the data category definitions. Such ontologies would address variations on higher semantic levels.

However, before higher levels can be addressed the foundation has to be laid. To achieve this, TC 37 has started to revise its existing registry and to build a new implementation that will provide a greater level of accessibility to and usability of the data in the registry. The new implementation has been dubbed *ISocat* [2] and builds on the experience of the earlier *Syntax* implementation [3]. This paper describes the data model for data category specifications, followed by ways in which external applications can reuse and reference these categories in the schemata for linguistic resources. It also addresses initial concerns with regard to the next semantic level, i.e. building domain ontologies on top of a selection of data categories (*data category selection* or DCS).

2. Data Category Registry Data Model

The DCR contains a collection of data category specifications. A *data category* (DC) is defined as “the result of the specification of a given data field” and can be viewed as an elementary descriptor in a linguistic structure or annotation scheme. The specification of a data category in the DCR reflects closely the data model of the DCR [4] and consists of three parts:

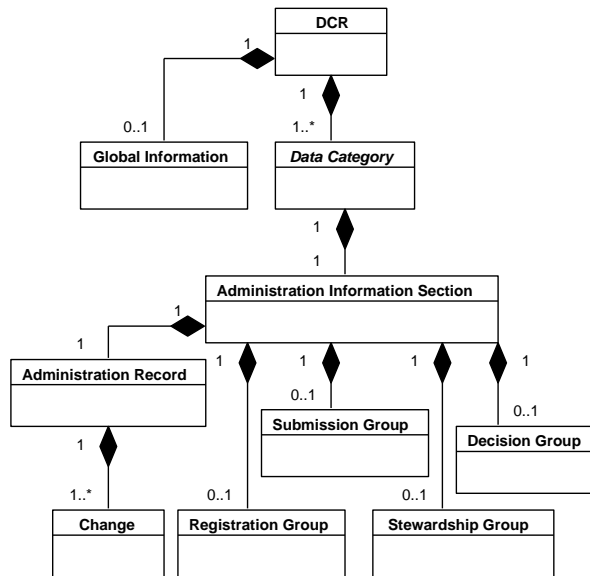


Figure 1 Administrative part

1. An administrative part (see Figure 1): dedicated to the administration and identification of the data category
2. A descriptive part (see Figure 2): dedicated to the documentation of the data category, with a required name and definition in English, plus information in zero or more other working languages, and possibly alternative names used in a given database, format, or application
3. A linguistic part (see Figure 3): dedicated to the documentation of the conceptual domain of the data category in the context of a specific object language

All types of data categories specify the first two parts of the model. The third part, the linguistic part, is only populated for complex data categories, i.e., non-terminal data categories that have a conceptual domain, which may consist of an enumeration of simple data categories (data categories without a conceptual domain). Different types of conceptual domains also provide a further distinction between different types of complex data categories. An open conceptual domain contains no further restriction on the permissible values for a complex data category and is associated with a so-called *open data category*. A closed conceptual domain consists of an enumeration of simple data categories that are permissible values and is associated with a *closed data category*. In order to allow more elaborate types of constraints, e.g., $date > 1980$, a third type of conceptual domain has been added to allow the constraint for a conceptual domain to be specified in a schema-specific language. No specific schema language is enforced on the user here; it is left to the creator or validators of a data category to ensure that

the specified constraint is valid for a schema language and in line with other constraints (possibly in other languages) specified for the data category. A *constrained data category* is associated with at least one schema-specific domain.

Persons wishing to contribute to the work of the DCR may register as expert users. Independent experts are assigned their own workspaces in which they may freely create or modify their data categories or select existing data categories for inclusion in their own DCSs. A DCS is generally used for collecting data categories for a specific application domain. For a number of thematic domains, standardized data category selections are available in the DCR, each of which is managed by a Thematic Domain Group (TDG; see www.isocat.org for the list of currently active TDGs). Inclusion of a new or modified data category in one of these thematic domains is subject to acceptance by the TDG. After a data category has been accepted for consideration, the governing DCR Board will decide on the standardization of the data category. The DCR Board may also decide to install new TDGs in order to introduce data categories for thematic domains not yet captured in the DCR. Standardized data categories will comprise an ISO standard as database [5], but the DCR itself will remain an open resource for the linguistics community.

3. Referencing Data Categories

A basic requirement for interoperability among resources using data categories is that references to the data categories used are included in the data or metadata of the resource. These references should be

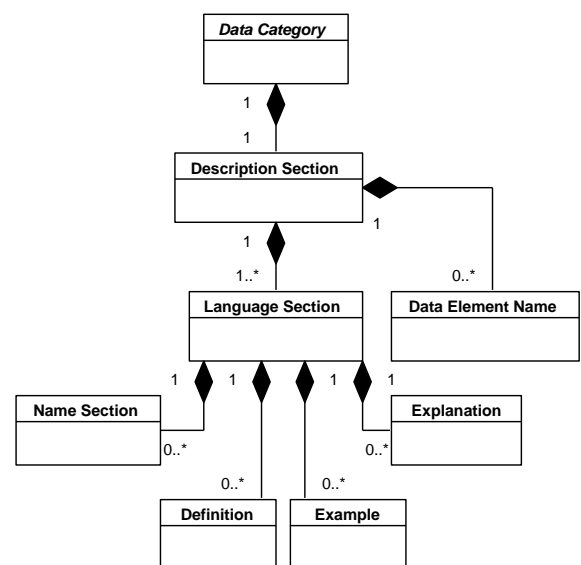


Figure 2 Descriptive part

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; [6]), handles [7], Archival Resource Keys (ARK; [8]), and Uniform Resource Names (URN; [9]).

The DCR uses cool URIs [10], 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://isocat.org/datcat/ISO-DC-1345>. The prefix of these URLs, <http://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. Using HTTP content negotiation other representations, e.g., HTML or RDF, can be requested.

4. Annotating Linguistic Resources

Existing or new linguistic resources can be annotated using this data category referencing mechanism. Increasingly, the format of these resources is XML-based and consists of a data document and an associated schema. Some schema languages support components that can be used to express equivalence to data categories from the DCR.

For example, ODD (One Document Does it all; [11]) from the Text Encoding Initiative (TEI) provides the `<equiv/>` component:

```
...
<elementSpec ident="pos">
  <equiv name="partOfSpeech"
    uri="http://isocat.org/datcat/ISO-DC-369"/>
  ...
</elementSpec>
...
```

Example 1

However, not every schema or schema language has specific facilities for establishing these kinds of data category equivalences. For XML-based schema languages and resources, the data category reference XML vocabulary can be used. This small vocabulary consists of the attribute `dcr:datcat` which establishes

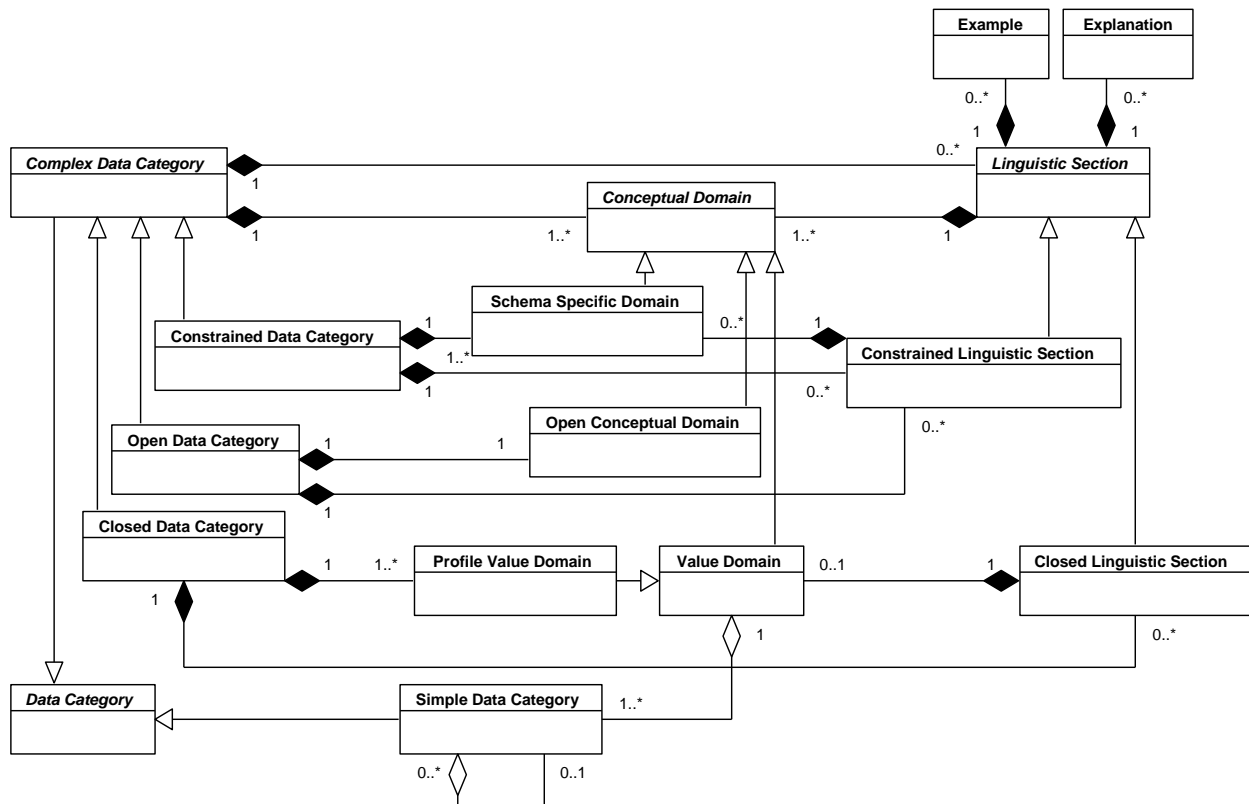


Figure 3 Linguistic part

an equivalence relationship between the current XML element in the case of a resource or the element the schema construction describes and a complex or simple data category. The following RELAX NG schema [12] is an example of annotating an XML schema using the `dcr:datcat` attribute:

```
...
<rng:element name="partOfSpeech"
  dcr:datcat="http://isocat.org/datcat/ISO-DC-369"
>
  <rng:choice>
    <rng:value
dcr:datcat="http://isocat.org/datcat/ISO-DC-370">
      verb
    </rng:value>
    <rng:value
dcr:datcat="http://isocat.org/datcat/ISO-DC-371">
      noun
    </rng:value>
    <rng:value
dcr:datcat="http://isocat.org/datcat/ISO-DC-372">
      proper noun
    </rng:value>
  </rng:choice>
</rng:element>
...
```

Example 2

ISOcat will support the creation and annotation of these XML schemata by being able to export DCSs in popular XML schema languages. A user can use these DCS exports as a starting point for creating a new schema for a linguistic resource.

5. Linguistic Knowledge Bases

The embedding of the kinds of references cited here provides advantages with regard to interoperability, but this is not to say that all possible ambiguities and interchange issues are solved in the process. The schemata in which the references are embedded provide the semantic context of the data categories. To achieve true interoperability as well, these contexts need to be (formally) described in the form of linguistic knowledge bases.

The DCR should support a wide range of applications and domains. It is hard enough to create a set of data categories, let alone at the same time creating ontological relationships among them which are valid for all these applications and domains. For this reason the DCR stores only two kinds of relationships:

1. the relation between a closed complex data category and the simple data categories in its value domain;
2. one super type relation between simple data categories and one other simple data category in order to facilitate management of large value domains.

As an example of modeling other possible ontological relationships, this paper will propose an approach for storing ontological relationships among data categories in an semantic web-based representation.

5.1 Data categories as RDF resources

RDF (Resource Description Framework; [13]) and RDFS (RDF Schema; [14]) form the basis of OWL (Web Ontology Language; [15]) and other semantic web technologies like SKOS (Simple Knowledge Organization System; [16]). The main part of the descriptive information from the data category specifications in the DCR is easily transformed into a set of RDF(S) statements (using the N3 notation [17]):

```
@prefix : <http://isocat.org/rest/dcs/139.rdf#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-
schema#> .
@prefix dcr: <http://isocat.org/ns/dcr.rdf#> .

:headword
  dcr:datcat <http://isocat.org/datcat/DC-258> ;
  rdfs:label "head word"@en ; rdfs:comment "A
lemma heading a dictionary entry."@en ;
  rdfs:label "lemma"@nl ;
  rdfs:comment "Het eerste woord van een
artikel in een woordenboek."@nl .

:partOfSpeech
  dcr:datcat <http://isocat.org/datcat/DC-396> ;
  rdfs:label "part of speech"@en ;
  rdfs:comment "A category assigned to a word
based on its grammatical and semantic
properties."@en .
```

Example 3

As was the case with the `dcr:datcat` attributes for annotating schemata (see the previous section) the `dcr:datcat` statements identify which data categories are reused. This transformation creates RDF resources. The decision to create an RDF class or property based on a data category is left to the knowledge designer and may vary depending on point of view or the focus of a particular RDF resource. For example, in one domain modeled by an RDFS knowledge base, */headword/* (this typographic convention implies a reference to a data category using its mnemonic identifier) may be a class while */partOfSpeech/* is a property of that class:

```
@prefix : <http://isocat.org/rest/dcs/139.rdf#> .

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-
syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-
schema#> .
```

```

:headword a rdfs:Class .

:partOfSpeech a rdf:Property ;
  rdfs:domain :headword .

```

Example 4

Although this distributed setup (the N3 document of Example 4 refers to the document of Example 3 using the <http://isocat.org/rest/dcs/139.rdf> URI) is a theoretically correct approach, most tools don't allow users to create an RDF vocabulary by merging multiple RDFS documents. This means that these tools only allow the knowledge designer to visualize the local part of the vocabulary, i.e., the tool will not load the label and the description of the data category-based class or property. One may also choose to download the RDFS export and edit it locally. When the DCS only contains standardized DCs, this is a viable approach, as their specifications will be stable. However, DCs from private workspaces (which perhaps have been made public by their owners) may be more volatile, which means that the DC specification stored locally may diverge over time from the specification available in the DCR. In the future ISOcat can provide a service which allows the user to upload an RDFS document and synchronize the descriptive information of classes and properties which claim equivalences with data categories.

5.2 Expressing TBX structures using OWL-DL

The initial RDF(S) DCS export as provided by the DCR can form the basis of various patterns which can be used to describe the semantic context of the data categories in the DCS. As the DCS export can only capture the descriptive information associated with the data categories, it is important that the patterns used should correctly capture the ontological relationships stored in the DCR. As an example, we will discuss an effort to build an OWL ontology reflecting the hierarchical data category structures elaborated in the TBX (TermBase eXchange; [18]) terminology interchange standard. TBX is designed to unambiguously exchange terminological data between potentially heterogeneous termbases, in part by imposing a data structure that is consistent with the Terminology Markup Framework (TMF; [19]). TBX was developed as a stand-alone XML-based language roughly within the framework of the TEI initiative. For some time, however, it has been the intention of TBX developers to also create an RDF representation of TBX structures in an effort to facilitate a crosswalk to the semantic web environment. After examining the potential for SKOS in this regard, Wright and Summers [20] concluded that differences in the semantics of SKOS

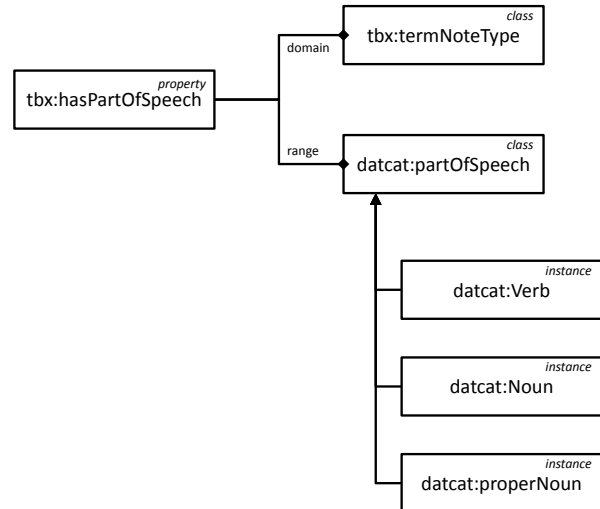


Figure 4 Structure of Example 5

and the semantics of DCR-related data categories made it more appropriate to model TBX using OWL-DL. This particular exercise is also interesting from the standpoint that in this context the data categories are viewed as data category objects structured according to the TBX data model. This distinction leads to solutions that diverge from the examples we have demonstrated above.

The mappings from complex data category types are as follows:

- a closed data category becomes a class, the simple data categories in its value domain become instances of that class, and a property is created with the class itself declared as its range;
- an open data category becomes a class with a corresponding data type property;
- a constrained data category also becomes a class with a corresponding property; if the constraints are expressed in an OWL compatible rule language, the constraints are attached to the property.

This approach allows for a fairly powerful utilization of the various logical expressions provided by OWL-DL. The following example (see also Figure 4) of the mapping is based on the closed data category */partOfSpeech/*. It is interesting to note that rather than citing */headword/* as the domain, as was the case with the previous mapping, the superordinate TBX class “TermNoteType” (equivalent to TBX: `<termNote type="...">`) acts as the domain value:

```

@prefix : <http://isocat.org/rest/dcs/139.rdf#> .
@prefix tbx: <http://www.lisa.org/tbx#> .

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

```

```

@prefix owl: <http://www.w3.org/2002/07/owl#> .

tbx:TermNoteType a rdfs:Class .
:partOfSpeech a rdfs:Class .

tbx:hasPartOfSpeech a rdf:Property ;
  rdfs:domain tbx:TermNoteType ;
  rdfs:range :partOfSpeech .

:Verb a :partOfSpeech .
:Noun a :partOfSpeech .
:properNoun a :partOfSpeech .

```

Example 5

As noted above, the class “TermNoteType” is not a true data category. `<termNote>` is a generic identifier in TBX, one of its so-called meta data categories, but it is not a linguistic data category in the DCR. Furthermore, the construct “TermNoteType” comprises a conflation of the meta data category together with the *type* attribute, which forms a logical super class that is convenient for modeling a sizable collection of data categories from the TBX data category selection in the DCR. Although this element reflects actual TBX markup, in assigning OWL-DL logic to the TBX model, it has proven convenient (indeed necessary at times) to create other virtual pseudo classes that are not only not true data categories – they are not even equivalent to actual TBX markup conventions. For instance, at one point in the TBX data model there is a logical structure that serves as a kind of container or switch that enables the insertion of either one of two structures. This virtual logical element does not actually exist in the form of a markup feature and is only used to explicitate a constraint within the model. In order to facilitate this constraint in the RDF representation, it was necessary to create a named class to represent this “container”, the pseudo-class “TermInfoSlot”.

It would be very tempting for purposes of completeness to add a data category to the DCR for each of these special concepts needed to model the knowledge base, but doing so would lead to ‘pollution’ of the DCR, especially in cases where these elements cannot be construed as actual data categories. To a certain extent, this is partially already true: for example the DCR contains `/termSection/` and `/globalInformation/`. These are, however, logical containers in which data categories can be placed, and they or their aliases do indeed occur in actual markup environments. Nevertheless, they do not have their own conceptual domains, or at least not analogous to the way we have treated conceptual domains until now in the DCR data model. Another option, however, would be to establish a parallel resource environment adjacent to the current DCR where thematic domains and interest groups like the TBX community could make their constructs of

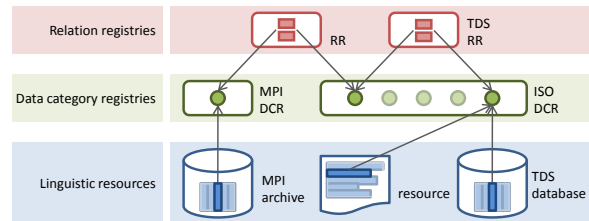


Figure 5 Example of the registry network.

this nature available for use in conjunction with the actual data categories resident in the DCR.

6. A network of registries

The DCR software will be open source, and it will thus be possible that multiple DCRs will become operational. As mentioned at the end of the previous section resource schemata and knowledge bases, like the TBX OWL-DL ontology, could also be made available in central registries, also known as *Relation Registries* (RRs). One could even envision that some of the RRs will be populated by machine learning algorithms instead of manual labor. As there is the possibility of multiple DCR instances, these RRs may thus also refer to data categories from various DCRs. In this way a network of registries will be built. If these registries are uniformly accessible, i.e., by the use of common standard formats and Application Programming Interfaces (APIs), software agents will be able to traverse these networks and assist users in finding relevant resources and interpreting them.

To visualize how this network would work, let us assume that a linguist with access to this registry network (see Figure 5), has found an interesting linguistic phenomenon in a resource in the language archive of the Max Plank Institute (MPI). Software agents with knowledge of the network can now analyze (part of) the resource for relevant data categories and track their location in the network. Using the various registry hubs, the software agent may thus identify that the *Typological Database System* (TDS; [21]), among other resources, contains typological information about the appearance of this interesting linguistic phenomenon in the same or even other languages of the world. The linguistic resources discovered in this way may be ranked using the relationship features. Traversal of certain types of relationships, e.g., ‘equivalence’ relationships may be considered to be stronger than loose relationships like ‘related to’. Also, ranking is influenced by the amount of context taken into account, e.g., when more of the context is inspected, results will be more accurate, but may take longer to be collected. Finally, properties of the visited registries (e.g., manually populated RRs may be more

trustworthy then algorithmically populated RRs) will lead to a ranking of the identified resources. In addition to identifying resources that are relevant to the linguist's research query, the software agent can also assist in the interpretation of these resources by allowing the linguist to navigate the semantic network encoded in the registries.

7. Conclusions and future work

The DCR is establishing itself as a source for authoritative standardized data category definitions. References to data category specification provide an essential ingredient for achieving interoperability between heterogeneous resources and an important step towards establishing a network of registries and software agents that interpret linguistic features encoded in these resources. Various ways of modeling linguistic resources are currently being reviewed and the implications on the DC specification delivery by the DCR are being analyzed. The examples for modeling linguistic knowledge bases indicate that the delivery mechanism for DC specifications for RDF(S) resources is model specific. The mechanism for delivery of DC specifications must be able to cope with this type of flexibility. Further modeling examples will be examined to provide more information on the type of DC specifications that is required from different modeling techniques.

The organizational structure maintaining DCR activities will be augmented in the next few months through the formal establishment of the DCR Board. National standardization bodies will be represented to oversee the standardization activities and provide support to the various Thematic Domain Groups. Thematic Domain Groups will be extended with additional experts from the field to assist in evaluating submitted data categories through their standardization process.

Additional functionality will be added to continue to provide support to the communities modeling activities and the DCR team will remain actively involved in exploring the various forms that use of the DCR can take.

Abbreviations

ARK	Archival Resource Key
DC	Data Category
DCIF	Data Category Interchange Format
DCR	Data Category Registry
DCS	Data Category Selection
DL	Description Logic
DOI	Digital Object Identifier
HTML	HyperText Markup Language

HTTP	HyperText Transfer Protocol
MPI	Max Planck Institute
ODD	One Document Does it all
OWL	Web Ontology Language
N3	Notation 3 (RDF)
PID	Persistent Identifier
RDF	Resource Description Framework
RDFS	RDF Schema
RR	Relation Registry
SKOS	Simple Knowledge Organization System
TDG	Thematic Domain Group
TDS	Typological Database System
TBX	TermBase eXchange
TEI	Text Encoding Initiative
TMF	Terminological Markup Framework
URI	Uniform Resource Identifier
URN	Uniform Resource Name

References

- [1] ISO DIS 12620, "Terminology and other language resources – Data categories – Specification of data categories and management of a Data Category Registry for language resources", International Organization for Standardization (ISO) 2008-08-01 2008.
- [2] "ISOcat", 2008, <http://www.isocat.org/>.
- [3] N. Ide and L. Romary, "A Registry of Standard Data Categories for Linguistic Annotation", in *International conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004.
- [4] M. Kemps-Snijders, M. A. Windhouwer, P. Wittenburg, and S. E. Wright, "A Revised Data Model for the ISO Data Category Registry", in *Proceedings of the 8th International Conference on Terminology and Knowledge Engineering*, B. N. Madsen and H. E. Thomsen, Eds. Copenhagen, Denmark, 2008.
- [5] ISO, "Procedure for the development and maintenance of standards in database format", International Organization for Standardization (ISO) 2007.
- [6] DOI, "The Digital Object Identifier (DOI) System.", International DOI Foundation 2001.
- [7] S. Sun, L. Lannom, and B. Boesch, "Handle System Overview": Internet Engineering Task Force, 2003, <http://www.ietf.org/rfc/rfc3650.txt>.
- [8] J. A. Kunze and R. P. C. Rodgers, "ARK Persistent Identifier Scheme", 2007.
- [9] R. Moats, "URN Syntax", in *IETF RFC 2141: Internet Engineering Task Force*, 1997, <http://www.ietf.org/rfc/rfc2141.txt>.

- [10] T. Berners-Lee, "Cool URIs don't change", 1998, <http://www.w3.org/Provider/Style/URI>.
- [11] TEI, "Getting started with ODDs", <http://www.tei-c.org/Guidelines/Customization/odds.xml>.
- [12] J. Clark, "RELAX NG home page", 2003, <http://relaxng.org/>.
- [13] "Resource Description Framework (RDF)", 2004, <http://www.w3.org/RDF/>.
- [14] D. Brickley and R. V. Guha, "RDF Vocabulary Description Language 1.0: RDF Schema", 2004.
- [15] D. L. McGuinness and F. v. Harmelen, "OWL Web Ontology Language Overview ", 2004.
- [16] "Simple Knowledge Organization System (SKOS) – Home Page", 2008, <http://www.w3.org/2004/02/skos/>.
- [17] T. Berners-Lee, "Notation 3", 2006, <http://www.w3.org/DesignIssues/Notation3>.
- [18] ISO DIS 30042, "TermBase eXchange (TBX) Format Specification", International Organization for Standardization (ISO) 2008.
- [19] ISO 16642, "Computer applications in terminology – TMF (Terminological Markup Framework).", International Organization for Standardization (ISO) 2003.
- [20] S. E. Wright and D. Summers, "Crosswalking from Terminology to Terminology: Leveraging Semantic Information across Communities of Practice.", in *LREC 2008 Workshop: Uses and Usage of Language-Related Standards*, A. Witt, F. Sasaki, E. Teich, N. Calzolari, and P. Wittenburg, Eds. Marrakech, Morocco, 2008, pp. 21-30.
- [21] TDS, "The Typological Database System Project", 2008, <http://language-link.let.uu.nl/tds/>.