ISOcat: Corralling Data Categories in the Wild

Marc Kemps-Snijders[a], Menzo Windhouwer[a], Peter Wittenburg[a], Sue Ellen Wright[b]
[a]Max-Planck-Institute for Psycholinguistics, [b]Kent State University
[a]Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
{Marc.Kemps-Snijders, Menzo.Windhouwer, Peter.Wittenburg}@mpi.nl, swright@kent.edu

Abstract

To achieve true interoperability for valuable linguistic resources different levels of variation need to be addressed. ISO Technical Committee 37, Terminology and other language and content resources, is developing a Data Category Registry. This registry will provide a reusable set of data categories. A new implementation, dubbed ISOcat, of the registry is currently under construction. This paper shortly describes the new data model for data categories that will be introduced in this implementation. It goes on with a sketch of the standardization process. Completed data categories can be reused by the community. This is done by either making a selection of data categories using the ISOcat web interface, or by other tools which interact with the ISOcat system using one of its various Application Programming Interfaces. Linguistic resources that use data categories from the registry should include persistent references, e.g. in the metadata or schemata of the resource, which point back to their origin. These data category references can then be used to determine if two or more resources share common semantics, thus providing a level of interoperability close to the source data and a promising layer for semantic alignment on higher levels.

1. Introduction

Linguistic resources are valuable for many stakeholders, e.g. researchers, language communities, translators, 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 world wide, these resources have to be interoperable. However, to achieve true interoperability of resources from heterogeneous sources, many variations on different levels have to be addressed. In ISO Technical Committee 37 (TC37), *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 (ISO DIS 12620, 2007). Parts of the (meta)data model of 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 categories. 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* [1] and builds on the experience of the earlier *Syntax* implementation (Ide and Romary, 2004). This paper describes the new data model for data category specifications, followed by ways in which external applications can reuse and reference these categories.

2. Modeling Data Categories

Each data category contains general information in the form of at least one *language section* (see Figure 1). This section provides a definition, possibly examples of its use, additional explanations and one or more *name sections*. The name section provides a place to store alternative names in specific application domains. A data category specification can also have multiple language sections, each of which provides the same information in a different language. The data model of the registry states that each data category should have at least one language section for the English language.

There are two basic types of data categories: *complex data categories* and *simple data categories* (see Figure 2). A complex data category can *contain* values, i.e. it has a *conceptual domain*, whereas a simple data category only *appears as* a value. The values for a complex data category may be further constrained. *Open data categories* are only constrained by their definition, but their conceptual domain is open. *Closed data categories*

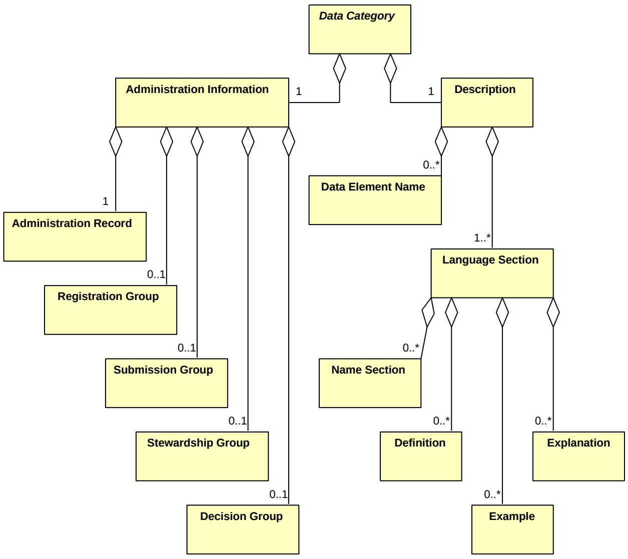


Figure 1: UML class diagram of the general data category structure.

[1] http://www.isocat.org/

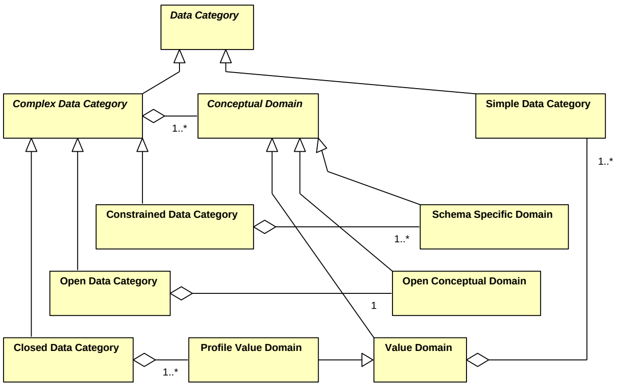


Figure 2: UML class diagram for data category types.

are not only constrained by their definition; their conceptual domain is also limited to an enumerated list of values consisting of simple data categories. Finally, for *constrained (open) data categories*, the conceptual domain is constrained by schema-type specific rules. An example would be a data category which uses the constraint language of W3C XML Schema to specify that it can only contain dates later then January 1st 2008.

The language section (see Figure 3) provides information about a data category for one or more *working languages*. The *linguistic section* records this information for one or more *object languages*. Additional language-specific constraints can be added to a data category. For example, the data category /*gender*/ is a closed data category with three simple data categories in its value domain: /*masculine*/, /*feminine*/ and /*neuter*/. When the object language is French, only a subset of this value domain is applicable: /*masculine*/ and /*feminine*/. Depending on the type of data category and its conceptual domain, the linguistic section allows for the further specific restriction of the conceptual domain.

3. Standardizing Data Categories

Experts can create own data categories in Syntax or ISOcat and share them with other users. It thus becomes possible for small communities to share data categories without the need to go through the formal standardization process. De facto standard data categories may arise in this way. When a data category or, more likely, a set of data categories, is submitted for standardization, it is assigned to a *Thematic Domain Group*. Thematic Domain Groups are initiated by the *Data Category Registry Board*, which ensures that the scope and coherence of the registry are maintained. Each group manages the data categories belonging to a specific class of applications, e.g. metadata or morphosyntax. The chair of the assigned group selects judges who will validate the candidate data category. This process leads in the end to either acceptance or rejection of the data category. Accepted data categories become standard after final approval by the board, and become part of the coherent and consistent set of data categories in the DCR. Rejected data categories are returned to the submitter accompanied by the reason why they have been rejected. The submitter may revise his/her proposal and resubmit it later.

It has been proposed that snapshots of the standardized data categories be submitted biennially to ISO as a Standard as Database (ISO, 2007). Although the data categories will also remain freely available in the DCR, ISO will then retain the right to disseminate this standardized subset according to ISO practices. These data categories can thus also attain the level of ISO standard.

4. Selecting Data Categories

Although the DCR functions as a meeting point to discuss and align the semantics of data categories, its ultimate goal is to promote reuse of the data categories within linguistic resources. Everyone can freely access the registry and collect data categories in so-called *data category selections*. These selections can be kept private

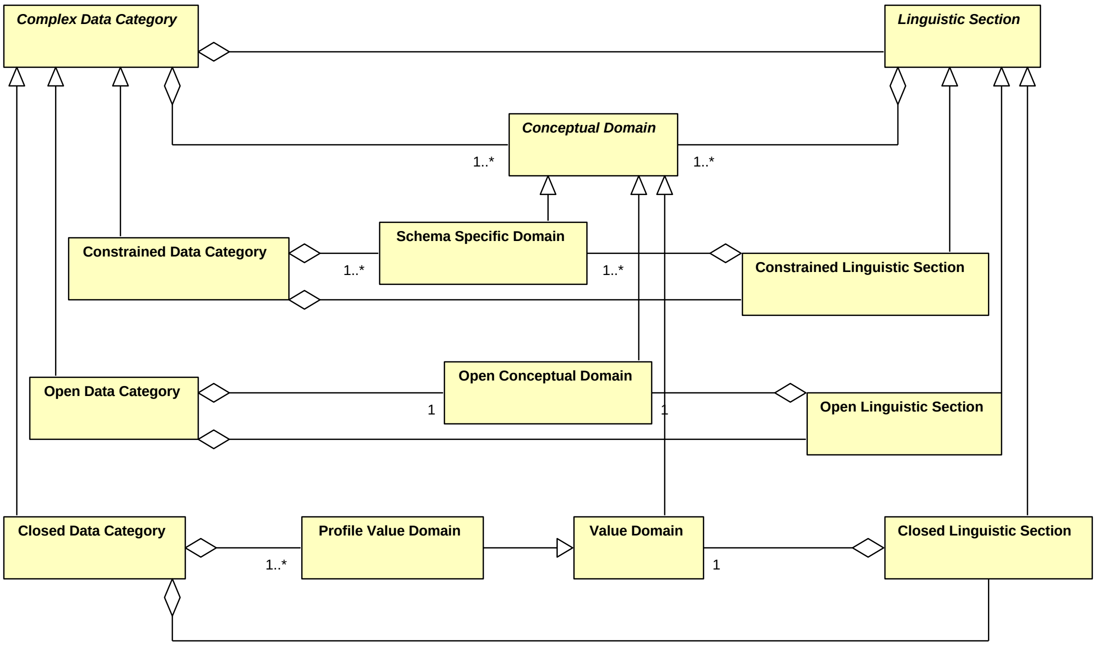


Figure 3: UML class diagram for the linguistic section.

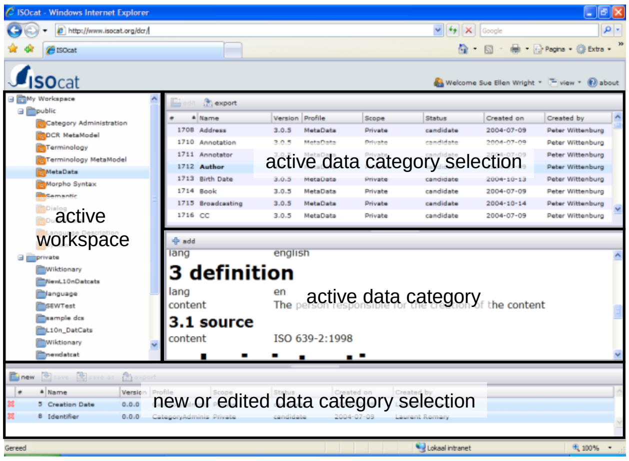
or shared with others. Managing them is an easy process whereby users simply drag and drop data categories into a data category selection using the ISOcat interface (see Figure 4). These data category selections can be exported in a standard XML format. ISOcat will support XSLT-based plug-in modules to facilitate transformation into alternative formats.

5. Referencing Data Categories

To achieve interoperability between various resources, their metadata should be able to indicate which data categories were used, i.e. they should be able to include references to the specific categories in the registry, so that the overlapping set of data categories can be identified. These references should be persistent. (Broeder, Declerck et al., 2007) proposes to use so-called, *handles* for these references. A handle consists of two parts: a *prefix* and a *suffix*, separated by a forward slash ("/"). The prefix is assigned by the *Handle System* [2] top authority and uniquely identifies an institution or organization. The suffix is assigned by this institution and organization, and its *Local Handle System* is designed to resolve these elements and can direct the client to the resource identified by a specific handle. Given this scenario, handles for data categories now take the following form[3]: `42/DC-1232`. The prefix "`42`" indicates the registry authority, e.g. the Max Planck Institute, which hosts the registry, and the suffix, "`DC-1232`", indicates a specific data category, e.g. version 1.3 of the */adverb/* data category owned by the morphosyntax Thematic Domain Group.

These handles can now be embedded in the metadata of a linguistic resource. Take 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 DIS 30042, 2007))[4]:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE TBXXCS SYSTEM "tbxxcsdtd.dtd">
<TBXXCS name='master' version="0.4" lang="en">
  …
  <datCatSet>
    …
    <termNoteSpec name="animacy"
        datcatId="hdl:42/DC-1902">
      <contents datatype="picklist"
        targetclass="none" forTermComp="yes">
        <termNoteSpec name="animate"
          datcatId="hdl:42/DC-1911"/>
        <termNoteSpec name="inanimate"
          datcatId="hdl:42/DC-1952"/>
        <termNoteSpec name="otherAnimacy"
          datcatId="hdl:42/DC-1953"/>
      </contents>
    </termNoteSpec>
    …
  </datCatSet>
```



#	Name	Version	Profile	Scope	Status	Created on	Created by
1708	Address	3.0.5	MetaData	Private	candidate	2004-07-09	Peter Wittenburg
1710	Annotation	3.0.5	MetaData	Private	candidate	2004-07-09	Peter Wittenburg
1711	Annotator	3.0.5	MetaData	Private	candidate	2004-07-09	Peter Wittenburg
1712	Author	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Peter Wittenburg
1713	Birth Date	3.0.5	MetaData	Private	candidate	2004-10-13	Peter Wittenburg
1714	Book	3.0.5	MetaData	Private	candidate	2004-07-09	Peter Wittenburg
1715	Broadcasting	3.0.5	MetaData	Private	candidate	2004-10-14	Peter Wittenburg
1716	CC	3.0.5	MetaData	Private	candidate	2004-07-09	Peter Wittenburg

lang english

3 definition

lang en

content The person responsible for the creation of the content

3.1 source

content ISO 639-2:1998

#	Name	Version	Profile	Scope	Status	Created on	Created by
5	Creation Date	0.0.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
8	Identifier	0.0.0	CategoryAdminis	Private	candidate	2004-07-09	Laurent Romary

Figure 4: Managing data category selections in ISOcat.

```
  …
</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 cannot incorporate the handles 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. The "`hdl:`" prefix identifies the value as being a handle.

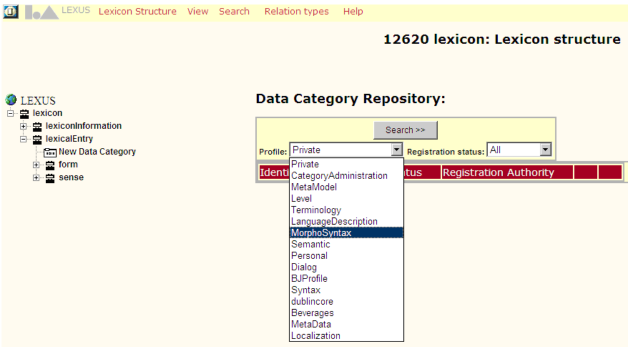
While the TBX standard already provides a placeholder for the handle, other metadata and schema technologies will not. For at least XML-based technologies the DCR provides a small XML vocabulary to embed handles. Here is a small section of the *Integrated Data and Documentation Format* (IDDF) metadata for a typological database (see (Windhouwer and Dimitriadis, 2008))[4]:

```
<?xml version="1.0" encoding="UTF-8"?>
<warehouse xmlns="…"
  xmlns:dcr="http://www.isocat.org/ns/dcr">
  <meta>
    …
    <notion id="56" name="number" scope="tds"
        type="top" dcr:datcat="hdl:42/DC-1902">
      <label>
        Specification of the count of
        participants in an event.
      </label>
      <values datatype="enum">
        <value dcr:datcat="hdl:42/DC-252">
          <literal>sg</literal>
          <label>singular</label>
        </value>
        …
      </values>
      …
    </notion>
  </meta>
  …
</warehouse>
```

[2] http://www.handle.net/

[3] At the moment of writing the prefix for the handle has not yet been assigned, so "`42`" is just an example. The exact format of the suffix is also still subject to discussion.

[4] The ellipses (…) indicate places where parts of the XML document have been omitted.



12620 lexicon: Lexicon structure

Data Category Repository:

Profile: Private Registration status: All

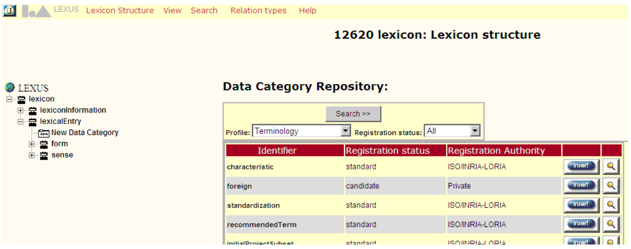
Figure 5: Selecting a thematic profile

The `dcr:datcat` attributes are added and used to embed persistent handles to the related complex and simple data categories in the metadata and schema section of the IDDF document.

6. Web Service Support

The previous section dealt with embedding references to data categories in the metadata of a linguistic resource. Collecting references to data categories may be achieved through manual lookup using Syntax or ISOcat. For applications providing user specific data models, however, a *web services API* is available. This API allows direct lookup and extraction of data category information (Kemps-Snijders, Ducret et al., 2006). The web services implemented by Syntax uses a RPC style. The ISOcat system will also support RESTful (Richardson and Ruby, 2007) and WSDL SOAP (Christensen, Curbera et al., 2001; Gudgin, Hadley et al., 2007) variants. Notice that all these interfaces provide read-only access. Creating and updating data categories is only done through the web interface.

Lexus (Kemps-Snijders and Wittenburg, 2006), for example, implements the *Lexical Markup Framework* (Francopoulo, George et al., 2006) This core model can be fleshed out by a user in various ways, one of which involves the interaction with the data category registry using Syntax's web services API and the selection of relevant data categories. Figures 5 to 8 show a series of screenshots of the interaction between Lexus and the DCR. In the first figure the user connected the DCR and selects one of the thematic profiles. The system then presents a list of data categories (see Figure 6). And, possibly after inspecting the details of a single data category (see Figure 7), the user adds the data category to its lexicon schema (see Figure 8). Other interactions are also possible, e.g. doing a keyword search for a data category.



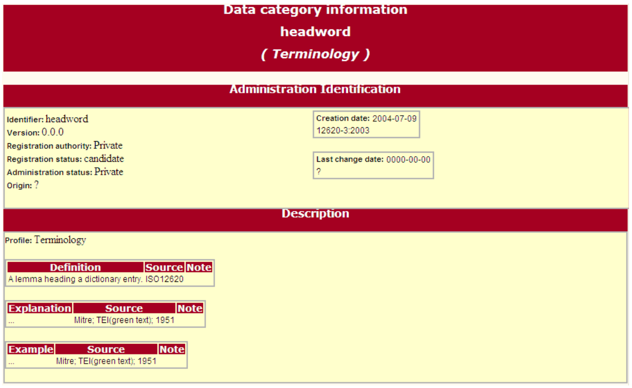
12620 lexicon: Lexicon structure

Data Category Repository:

Profile: Terminology Registration status: All

Identifier	Registration status	Registration Authority		
characteristic	standard	ISO/INRIA-LORIA	Insert	🔍
foreign	candidate	Private	Insert	🔍
standardization	standard	ISO/INRIA-LORIA	Insert	🔍
recommendedTerm	standard	ISO/INRIA-LORIA	Insert	🔍
initialProjectSubset	standard	ISO/INRIA-LORIA	Insert	🔍

Figure 6: Selecting a data category


Data category information
headword
(Terminology)

Administration Identification

Identifier: headword
Version: 0.0.0
Registration authority: Private
Registration status: candidate
Administration status: Private
Origin: ?

Creation date: 2004-07-09
12620-3:2003

Last change date: 0000-00-00
?

Description

Profile: Terminology

Definition	Source	Note
A lemma heading a dictionary entry.	ISO12620	

Explanation	Source	Note
...	Mitre; TEI(green text); 1951	

Example	Source	Note
...	Mitre; TEI(green text); 1951	

Figure 7: Inspecting data category details

7. Conclusion

The new implementation of the DCR provides a coherent framework for dealing with various aspects of data category standardization. Accessibility and reusability of data categories for user communities will increase by creating a market place for both standardized and non-standardized data categories. This process is supported by a number of mechanisms which allow easy selection of data categories, supported by stable referencing schemes. Tools support is provided through service APIs which feature search, browse and look-up support, thus allowing external tools to provide automated support for users in their selection process or directly to extract data category information using the DCR.

8. References

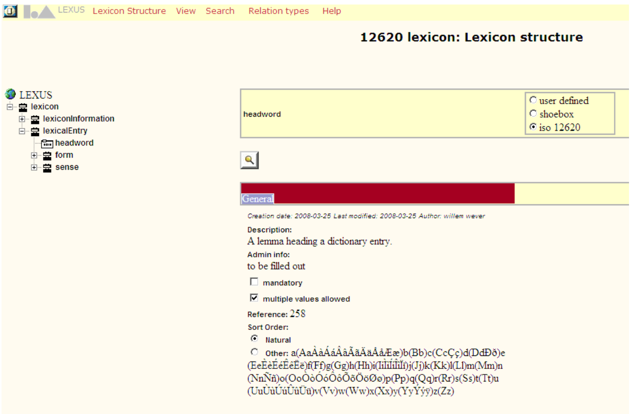
Broeder, D., T. Declerck, et al. (2007). *Citation of Electronic Resources*. Pragmatic Applications for TC 37 Standards, Provo UT.

Christensen, E., F. Curbera, et al. (2001). "Web Services Description Language." Version 1.1. http://www.w3.org/TR/wsdl.

Francopoulo, G., M. George, et al. (2006). *Lexical Markup Framework (LMF)*. International conference on Language Resources and Evaluation, Genoa, Italy.

Gudgin, M., M. Hadley, et al. (2007). "SOAP." Version 1.2. http://www.w3.org/TR/soap12-part1/.

Ide, N. and L. Romary (2004). *A Registry of Standard Data Categories for Linguistic Annotation*.



12620 lexicon: Lexicon structure

headword

○ user defined
○ shoebox
● iso 12620

Creation date: 2008-03-25 Last modified: 2008-03-25 Author: willem wever

Description:
A lemma heading a dictionary entry.

Admin info:
to be filled out

□ mandatory

☑ multiple values allowed

Reference: 258

Sort Order:

● Natural

○ Other: a(AaÀàÁáÂâÃãÄäÅåÆæ)b(Bb)c(CcÇç)d(DdÐð)e
(EeÈèÉéÊêËë)f(Ff)g(Gg)h(Hh)i(IiÌìÍíÎîÏï)j(Jj)k(Kk)l(Ll)m(Mm)n
(NnÑñ)o(OoÒòÓóÔôÕõÖöØø)p(Pp)q(Qq)r(Rr)s(Ss)t(Tt)u
(UuÙùÚúÛûÜü)v(Vv)w(Ww)x(Xx)y(YyÝýÿ)z(Zz)

Figure 8: Embed the data category

International conference on Language Resources and Evaluation, Lisbon, Portugal.

ISO (2007). Procedure for the development and maintenance of standards in database format, International Organization for Standardization (ISO).

ISO DIS 12620 (2007). 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).

ISO DIS 30042 (2007). TermBase eXchange (TBX) Format Specification, International Organization for Standardization (ISO).

Kemps-Snijders, M., J. Ducret, et al. (2006). *An API for Accessing the ISO Data Category Registry*. International conference on Language Resources and Evaluation, Genoa, Italy.

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

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

Windhouwer, M. and A. Dimitriadis (2008). Sustainable operability: Keeping complex resources alive. *LREC Workshop on Sustainability of Language Resources and Tools for Natural Language Processing*. Marrakech, Morrocco.