Linking to Linguistic Data Categories in ISOcat

Menzo Windhouwer and Sue Ellen Wright

Abstract ISO Technical Committee 37, Terminology and other language and content resources, established an ISO 12620:2009 based Data Category Registry (DCR), called ISOcat (see http://www.isocat.org), to foster semantic interoperability of linguistic resources. However, this goal can only be met if the data categories are reused by a wide variety of linguistic resource types. A resource indicates its usage of data categories by linking to them. The small DC Reference XML vocabulary is used to embed links to data categories in XML documents. The link is established by an URI, which servers as the Persistent IDentifier (PID) of a data category. This paper discusses the efforts to mimic the same approach for RDF-based resources. It also introduces the RDF quad store based Relation Registry RELcat, which enables ontological relationships between data categories not supported by ISOcat and thus adds an extra level of linguistic knowledge.

1 Introduction

ISO Technical Committee 37 Terminology and other language and content resources established a Data Category Registry (DCR), called ISOcat, to foster semantic interoperability of linguistic resources. ISOcat is based on ISO 12620:2009, which describes the data model and the management procedure for a DCR (ISO 12620, 2009). These procedures follow a grass roots approach, which means that any linguist can add the data categories (s)he needs to the registry. Standardized subsets of these data categories are created by a standardization procedure involving groups of

Menzo Windhouwer
Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands, e-mail: Menzo.Windhouwer@mpi.nl

Sue Ellen Wright
Kent State University, 109 Satterfield Hall, Kent, OH 44242, USA e-mail: sellenwright@gmail.com

international experts who are members of various Thematic Domain Groups (TDGs) and the DCR Board. There are currently over a dozen domains supported by a TDG, e.g., metadata, morphosyntax and terminology. But the aim of improving the semantic interoperability can only be met by the data categories if they are reused by a multitude of linguistic resource types (Kemps-Snijders et al, 2008). A resource indicates its usage of data categories by linking to them (Windhouwer et al, 2010). This paper focuses on how this can be done, and gives special attention to linked open data, i.e., RDF-based, resources.

2 Linking to Data Categories from XML-Based Resources

The focus of ISOcat has been mainly on general XML-based resources. ISO 12620:2009 specifies a small Data Category (DC) Reference XML vocabulary (see http://www.isocat.org/12620/) to annotate XML documents with links to data categories. This vocabulary, using Relax NG compact syntax as the schema language, defines two basic annotation descriptors, as shown in Example 1.[1]

```
default namespace dcr = "http://www.isocat.org/ns/dcr"

dcr_attribute_datcat = attribute datcat { xsd:anyURI }
dcr_attribute_value_datcat = attribute valueDatcat { xsd:anyURI }
```

Example 1: DC Reference attributes specified in Relax NG

The `dcr:datcat` descriptor can be used to annotate any XML element with a link to the equivalent data category in ISOcat. The other descriptor, `dcr:valueDatcat`, is in general used to annotate the textual value of an element or attribute, i.e., to annotate this value with a link to a simple data category.

2.1 Persistent Identifiers

As Example 1 shows, the link to the data category in ISOcat should be established by a URI. In ISOcat each data category has a unique and persistent identifier, also known as the PID (Persistent Identifier). The URI scheme that ISOcat uses for the PID is called a 'cool URI' (Berners-Lee, 1998), which is basically a standard HTTP URL with extra guarantees that these URLs will remain resolvable over a long period of time. The use of cool URIs is only one of the possible approaches to creating

[1] The DC Reference XML vocabulary defines the descriptors both as XML attributes and XML elements. The specific structure of the annotated XML-based resource determines whether either the attribute or the element should be used.

PIDs. ISO TC 37 has recently published a new standard, PISA (Persistent Identification and Sustainable Access, ISO 14619:2011), which describes the requirements to be met by these PID systems.

2.2 Data Category Types

In the structure of a resource various elements play different roles, e.g., some elements can have values while other elements group other elements. To accommodate these various roles there are data categories of different types:

1. Complex data categories have a typed value domain; the DCR data model supports various ways to describe these value domains:
   a. Open data categories can take any value allowed by the associated type;
   b. Closed data categories enumerate their allowed values as simple data categories (see below);
   c. Constrained data categories restrict their allowed values by one or more rules, e.g., any day in the 20th century;
2. Simple data categories are describe values associated with a closed data category;
3. Container data categories don't have a value domain but can be used to group other container or complex data categories together.

ISOcat does not store any relationships beyond the basic value domain relationships between simple and closed data categories. This means that specific structures built using container and complex data categories are not available in ISOcat as this would hamper their reuse. These structures are preferably specified in a resource schema document, e.g., a W3C XML Schema or Relax NG document, annotated with data category references.

2.3 An Annotated LMF Document

The ISO standard for the Lexical Markup Framework, ISO 24613:2008, encourages the use of ISOcat data categories although it omits to mention the DC Reference XML vocabulary. Example 2, taken from the standard (ISO 24613, 2008), has been annotated with ISOcat data category PIDs.[2]

It is apparent that the annotation of an instance quickly becomes verbose. A general solution is to annotate the schema of the resource, e.g., the W3C XML Schema or Relax NG schema. This way many instances can be annotated in using a single resource.

[2] Due to space limitations the common ISOcat cool URI prefix `http://www.isocat.org/datcat` has been replaced by elipses.

```
<LexicalResource xmlns:dcr="http://www.isocat.org/ns/dcr">
  <GlobalInformation>
    <feat att="languageCoding" dcr:datcat=".../DC-2008"
      val="ISO 639-3"/>
  </GlobalInformation>
  <Lexicon>
    <feat att="language" dcr:datcat=".../DC-1969" val="eng"/>
    <LexicalEntry>
      <feat att="partOfSpeech" dcr:datcat=".../DC-1345"
        val="commonNoun" dcr:valueDatcat=".../DC-1256"/>
      <Lemma>
        <feat att="writtenForm" dcr:datcat=".../DC-1836"
          val="clergyman"/>
      </Lemma>
      ...
      <WordForm>
        <feat att="writtenForm" dcr:datcat=".../DC-1836"
          val="clergymen"/>
        <feat att="grammaticalNumber" dcr:datcat=".../DC-1298"
          val="plural" dcr:valueDatcat=".../DC-1354"/>
      </WordForm>
    </LexicalEntry>
  </Lexicon>
</LexicalResource>
```

Example 2: Annotated LMF example

The example doesn't show the use of container data categories as this is a recent addition to the DCR data model not even covered by ISO 12620:2009. For the LMF core model and its extensions these container data categories have not been specified yet. However, it does show an open data category, i.e., */writtenForm/* (http://www.isocat.org/datcat/DC-1836), an simple data category, i.e., */commonNoun/* (http://www.isocat.org/datcat/DC-1256) which is an instance of the value domain from a closed data category, i.e., */partOfSpeech/* (http://www.isocat.org/datcat/1345).

3 Linking to Data Categories from RDF-Based Resources

Using the descriptors described above any XML document can refer to data categories to make the semantics explicit for elements, attributes and values. However, determining the level of semantic interoperability still involves additional processing, e.g., determining the overlap in semantics by finding the shared data categories. In the Linked Data world of RDF resources this interoperability is built into the data model used, i.e., data categories can play a direct role in that model as they are resources identifiable by a (cool) URI.

3.1 RDF Resource and Data Category Types

At first glance there seems to be a natural correspondence between data category types and RDF types: container data categories correspond to RDF classes, complex data categories correspond to RDF properties, and simple data categories correspond to literal values. However, this simple mapping comes with some drawbacks:

- Data categories would be far more prominent in the RDF model than in the XML model, as the direct use of the data categories would impose their non-semantics bearing URIs on the model being constructed. In the case of an XML model the fact that data categories are references leaves the choice of terminology more to the model builder. This is also more in line with the terminology management background of the DCR, which is also reflected by the data model which allows the specification of various (technical) terms used for a data category in a variety of circumstances, e.g., language or application specific.
- Literal values cannot be annotated in an RDF model, which means that the appropriate simple data category cannot be referred to when the literal value is ambiguous in the profile value domains of the closed data category. One possible solution could be to use simple data categories as individuals, but in this case the Cool URL of simple data category would again feature prominently in the RDF model.
- RDF models can actually be used to fine tune the ontological relationships between data categories, and specifying fixed type would hinder this as these relationships might require different mappings to RDF types.

For now ISOcat leaves the actual mapping to RDF types to the model builder and outputs the data categories as related to RDF resources. The model builder can then decide which types are appropriate. Also the data category cool URIs are, just like in the XML world, used to annotate these RDF resources so the model builder can fine tune these resources, e.g., using his/her own terminology. Experiments with approaches to mapping will continue, and either simple or more advanced forms of mapping might even be used for a (semi-)automatic conversion to RDF for annotated (either inline or by their schema) XML documents.

Example 3 shows an RDF specification for a simple model for a dictionary annotated with data category references.

3.2 RDF Predicates and Data Category Links

The example also shows the use of the `dcr:datcat` predicate to associate the RDF resource with the data category (Example 4).

The first version of the ISOcat RDF export used `owl:sameAs`. But the drawback of that approach is that by using OWL semantics the annotation has impact on the OWL model being built, i.e., it quickly pushes the OWL model to OWL Full.

```
@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 ;
...
: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: Annotated RDF resource

```
@prefix dcr: <http://isocat.org/ns/dcr.rdf#> .

dcr:datcat a owl:AnnotationProperty ;
  rdfs:label "data category"@en ;
  rdfs:comment "This resource is equivalent to this data
    category."@en ;
  skos:note "The data category should be identified by
    its Persistent IDentifier (PID)."@en ;
...
```

Example 4: `dcr:datcat` annotation property

This is an unwanted side effect and is prevented by specifying a dedicated annotation property. Once more the RDF model builder can fine tune this. Depending on the actual RDF type of the annotated RDF resource the `dcr:datcat` predicate can be replaced by the following OWL (2) predicates: `owl:equivalentClass` for classes, `owl:equivalentProperty` for properties and `owl:sameAs` for individuals. The use of these specific predicates limits the impact of ISOcat data categories on OWL semantics.

4 Ontological Relationships

ISOcat basically contains a flat list of data categories, i.e., it doesn't store (ontological) relationships between container and/or complex data categories. In addition to value domain relationships between simple and closed data categories, only a subsumption hierarchy between simple data categories is stored, but only one such a subsumption hierarchy is allowed, i.e., a simple data category can only be a child of one other data category. The storage of these ontological relationships in ISOcat is due to legacy issues and its usage is actually discouraged.

The reason that ontological relationships aren't stored in ISOcat is that they are highly domain or even application dependent and thus would hamper standardiza-

tion of data category specifications. However, they are important to make the semantics of linguistic resources explicit. To support this a companion registry to ISOcat named RELcat is under construction (Schuurman and Windhouwer, 2011). In RELcat anyone or any group can store (ontological) relationships between data categories and/or concepts from other registries.

```
@prefix relcat    : <http://www.isocat.org/relcat/set/> .
@prefix rel       : <http://www.isocat.org/relcat/relations#> .
@prefix dc        : <http://purl.org/dc/elements/1.1/> .
@prefix isocat    : <http://www.isocat.org/datcat/> .

relcat:cmdi {
    isocat:DC-2573 rel:sameAs dc:identifier .
    isocat:DC-2482 rel:sameAs dc:language .
    ...
    isocat:DC-2556 rel:subClassOf dc:contributor .
    isocat:DC-2502 rel:subClassOf dc:coverage .
}
```

Example 5: Relations between data categories and Dublin Core elements

Example 5 shows the set of relationships between data categories in ISOcat and Dublin Core elements. This set is in use by the metadata search engine for the CLARIN MetaData Infratructure (Broeder et al, 2008, CMDI). Mappings to support crosswalks to any other (linguistic) metadata element set, e.g., OLAC (Simons and Bird, 2003), or ontologies or taxonomies, e.g., GOLD (Farrar and Langendoen, 2010), can be added in the same vein.

4.1 Ontological Relationship Types

RELcat supports the following ontological relationship types:

1. related
   a. same as (a symmetric and transitive relationship)
   b. almost same as (a symmetric relationship)
   c. broader than (a transitive relationship and the inverse of the 'narrower than' relationship)
      i. superclass of (a transitive relationship and the inverse of the 'subclass of' relationship)
      ii. has part (a transitive relationship and the inverse of the 'part of' relationship)
         A. has direct part (the inverse of the 'direct part of' relationship)
   d. narrower than (a transitive relationship and the inverse of the 'broader than' relationship)

   i. sub class of (a transitive relationship and the inverse of the 'super class of' relationship)
   ii. part of (a transitive relationship and the inverse of the 'has part' relationship)
      A. direct part of (the inverse of the 'has direct part' relationship)

Although inspired by OWL and SKOS these relationship types may seem to be an impoverished set. But they are already an extension to the original purpose of RELcat, which mainly dealt with (almost) same-as relationships. However, this shallow taxonomy is just a first start. Other relationship types from other richer vocabularies, e.g., complete OWL or SKOS, can be inserted at the proper place in this subsumption hierarchy:

1. related
   a. same as (a symmetric and transitive relationship)
      i. `owl:equivalentClass`
      ii. `owl:equivalentProperty`
      iii. `owl:sameAs`
      iv. `skos:exactMatch`
   b. almost same as (a symmetric relationship)
      i. `skos:closeMatch`
   c. ...

Now sets of relations using these vocabularies can be loaded into RELcat, and be combined and exploited in their usual fashion, e.g., by an inferencing engine. For example, this is done for the GOLD ontology of linguistic concepts (Farrar and Langendoen, 2010). However, the upper part of the taxonomy can be used by generic algorithms to traverse the large graph created by the combined relationships.

```
PREFIX rel:<http://www.isocat.org/relcat/relations#>
PREFIX isocat:<http://www.isocat.org/datcat/>

SELECT ?rel WHERE { isocat:DC-2482 rel:related ?rel . }
```

Example 6: SPARQL query for relations with */languageID/* (`http://www.isocat.org/datcat/2482`DC-2482)

The query in Example 6 returns both the Dublin Core `dc:language` metadata element and the Language (http://purl.org/linguistics/gold/Language) GOLD concept, although their relationships with the */languageID/* (http://www.isocat.org/datcat/2482) data category have been expressed using different RDF vocabularies. One of the purposes of RELcat is to provide this information to semantic search engines to enable the retrieval of closely related resources of different types.

5 Conclusion and Future Work

Although the ISOcat data model is expressed in a more conventional UML data model the registered data categories can actually easily be used in the context of Linked Data due to the use of cool URIs as PIDs. This paper discussed the use of dedicated annotation attributes and properties to annotate existing XML and RDF documents. It also discussed some if the design decisions for RELcat, a Relation Registry, which enabled to specify ontological relationships among ISOcat data categories but also with concepts from other registries. Future work includes further development of RELcat aiming at achieving a higher level of semantic interoperability.

References

Berners-Lee T (1998) Cool URIs don't change. Tech. rep., World Wide Web Consortium, http://www.w3.org/Provider/Style/URI.html

Broeder D, Declerck T, Hinrichs E, Piperidis S, Romary L, Calzolari N, Wittenburg P (2008) Foundation of a component-based flexible registry for language resources and technology. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco

Farrar S, Langendoen DT (2010) An OWL-DL implementation of GOLD: An ontology for the semantic web. In: Witt AW, Metzing D (eds) Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology, Springer

ISO 12620 (2009) Terminology and other language and content resources - Specification of data categories and management of a Data Category Registry for language resources

ISO 24613 (2008) Language resource management - Lexical markup framework (LMF)

Kemps-Snijders M, Windhouwer M, Wittenburg P, Wright SE (2008) ISOcat: Corralling data categories in the wild. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, http://www.lrec-conf.org/proceedings/lrec2008/

Schuurman I, Windhouwer M (2011) Explicit semantics for enriched documents. What do ISOcat, RELcat and SCHEMAcat have to offer? In: Proceedings of the 2nd Supporting Digital Humanities Conference, Copenhagen, Denmark

Simons G, Bird S (2003) The open language archives community: An infrastructure for distributed archiving of language resources. Literary and Linguistic Computing 18(2):117–128

Windhouwer M, Wright SE, Kemps-Snijders M (2010) Referencing ISOcat data categories. In: Budin G, Declerck T, Romary L, Wittenburg P (eds) Proceedings of the LREC 2010 LRT standards workshop, Malta, http://www.lrec-conf.org/proceedings/lrec2010/workshops/W4.pdf