Originally published in: Simperl, Elena/Cimiano, Philipp/Polleres, Axel/Corcho, Oscar/Presutti, Valentina (Eds.): The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings. – Berlin/Heidelberg: Springer, 2012. Pp. 285-299. (Lecture Notes in Computer Science 7295)
DOI: *https://doi.org/10.1007/978-3-642-30284-8_26*

POSTPRINT

Claus Zinn
Christina Hoppermann
Thorsten Trippel

Department of Linguistics, University of Tübingen, Germany
{claus.zinn,christina.hoppermann,thorsten.trippel}@uni-tuebingen.de

The ISOcat Registry Reloaded

Abstract. The linguistics community is building a metadata-based infrastructure for the description of its research data and tools. At its core is the ISOcat registry, a collaborative platform to hold a (to be standardized) set of data categories (*i.e.*, field descriptors). Descriptors have definitions in natural language and little explicit interrelations. With the registry growing to many hundred entries, authored by many, it is becoming increasingly apparent that the rather informal definitions and their glossary-like design make it hard for users to grasp, exploit and manage the registry's content. In this paper, we take a large subset of the ISOcat term set and reconstruct from it a tree structure following the footsteps of `schema.org`. Our ontological re-engineering yields a representation that gives users a hierarchical view of linguistic, metadata-related terminology. The new representation adds to the precision of all definitions by making explicit information which is only implicitly given in the ISOcat registry. It also helps uncovering and addressing potential inconsistencies in term definitions as well as gaps and redundancies in the overall ISOcat term set. The new representation can serve as a complement to the existing ISOcat model, providing additional support for authors and users in browsing, (re-)using, maintaining, and further extending the community's terminological metadata repertoire.

1 Introduction

The linguistics community has accumulated a tremendous amount of research data over the past decades. It has also realized that the data, being the back-bone of published research findings, deserves equal treatment in terms of archiving and accessibility. For the sustainable management of research data, archiving infrastructures are being built, with metadata-based issues taking center stage. Metadata schemes need to be defined to adequately describe the large variety of research data. The construction of schemas and the archiving of resources will be conducted locally, usually in the place where the research data originated.

To ensure the interoperability of all descriptional means, the ISOcat data category registry has been constructed (see `http://www.isocat.org`). The registry, implementing ISO12620:2009 [4], aims at providing a set of data categories for the description of concepts and resources in various linguistic disciplines (syntax, semantics, *etc.*), but also features a section on metadata terms, which is our primary concern in this paper. Linguists giving a metadata-based description of their research data are solicited to only use metadata descriptors from the

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-108624

registry. When the registry lacks an entry, researchers are encouraged to extend it by defining new data categories. The registry has a governing body to ensure the quality and merit of all entries submitted for standardization. It is hoped that its grass-root nature helps defining a sufficiently large set of metadata descriptors of which a standardized subset reflects a consensus in a large user base. While the grass-roots approach is appealing, the organization of the registry's content as a glossary of descriptors with little structure is problematic. With the metadata term set now containing 450+ entries, with new entries added regularly, it becomes increasingly hard to browse and manage its content. To address this issue, we propose to re-organise the rather flat knowledge structure into a more systematic, formal and hierarchical representation, following the footsteps of `schema.org`. The new structure can serve as a complement to the existing one, giving users an alternative and more accessible entry point to the registry.

The remainder of this paper is structured as follows: Sect. 2 gives an account of the ISOcat registry. In Sect. 3, we describe our ontological reconstruction and re-engineering to represent the contents of the glossary by a hierarchically-structured concept scheme. Sect. 4 discusses ontology engineering issues and sketches future work, and Sect. 5 concludes.

2 The ISOcat Data Category Registry

2.1 Specification of Data Categories

The ISOcat data category registry is an implementation of ISO 12620:2009 [4] and accessible by a web-based interface (see `http://www.isocat.org`). The registry's content is currently partitioned into 14 thematic domain groups (TDGs) such as "Metadata", "Morphosyntax", "Terminology", and "Lexical Semantics", as well as additional project-related work spaces. Each TDG is governed by a decision-making body that considers all requests for the standardization of a data category. Note that the work reported herein is exclusively concerned with the TDG "Metadata". This group has 458 data categories (DC).[1]

All users have read access to the public parts of the registry. Creators of metadata schemes can make reference to an ISOcat DC by using the entry's persistent identifier. Registered users gain write access to the registry; they can define new entries (becoming owner of the entry) and also modify them at a later stage. DCs owned by other users cannot be modified. New entries get the registration status "private", but can be proposed for standardization. The ownership of standardized entries is transferred to the TDG's governing body. This policy concentrates curation efforts on the original creators, or the governing bodies; users wanting changes to data categories they do not own have to contact the DC's owners, or add their own, suitable defined, data category to the registry.

A specification of a data category consists of three parts: an administrative part, a descriptive part, and a conceptual domain. The administrative part includes, among others, the DC's registration status (private, candidate, standard),

[1] Accessed December 12, 2011 at `http://www.isocat.org`

origin (name of creator), versioning information (creation date, last change), an English mnemonic identifier, and a persistent identifier. The description section gives a natural language definition of the DC in English. Optionally, it can be complemented by other language sections to give for instance, a DC a French name and a French *definiens*. When more than one DC name is given, one of the names needs to be identified as "preferred name". Moreover, it is also possible to complement an existing definition section with another one. Thus a DC can be associated with multiple, presumably semantically similar, definitions. The third part gives the conceptual domain of a (non-simple) data category. A DC must take one of four types: complex/closed, complex/open, complex/constrained, and simple.[2] DCs of type complex/closed have a conceptual domain entirely defined in terms of an enumerated set of values, where each value must be defined as a DC of type simple. DCs of type complex/constrained have a conceptual domain restricted by a constraint given in some constraint language, and a DC of type complex/open can take any value. DCs of type simple are values, and thus do not have a conceptual domain. Each conceptual domain has a mandatory data type, in accordance with those defined by W3C XML Schema. The default data type is "string", which is also the datatype that is used most in the TDG metadata. Complex/open DCs specify only the datatype as conceptual domain.

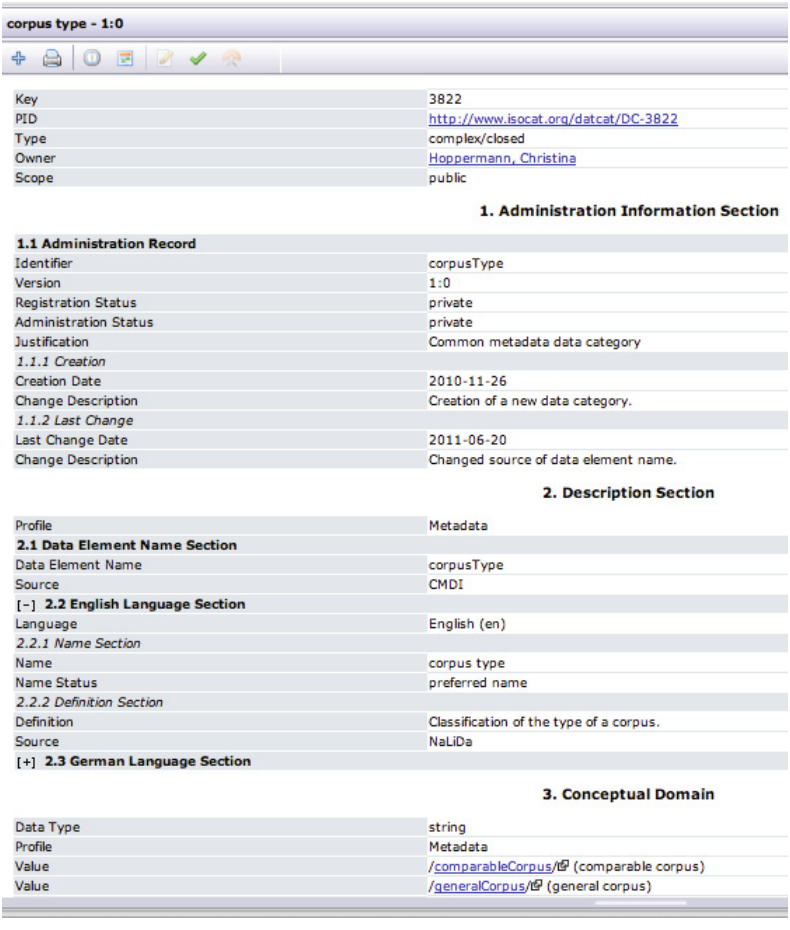
Fig. 1 shows an excerpt of the specification of the complex/open DC `/corpusType/`. Its description section shows the DC's data element name "corpusType", its English name "corpus type" and its natural-language definition ("Classification of the type of a corpus."); the DC's conceptual domain, or value range, is given as a closed set of simple DCs. While `/corpusType/` belongs only to TDG Metadata, a data category can, in general, belong to more than one profile.[3]

2.2 On Data Category Relationships and Definitions

There is a limited notion of "authorized" relationship in the ISOcat registry. A simple data category (*e.g.*, `/specialisedCorpus/`) can be member of the value domain of a complex/closed data category (*e.g.*, `/corpusType/`). ISO12620:2009 also specifies the possibility that simple data categories can be related to each other via a (single-inheritance) IS-A subsumption relation.[4] Example entries in the ISOcat registry include, for instance, `/technicalTranslation/` IS-A `/translation/`, and `/translation/` IS-A `/languageMediation/`.

Relationships between complex data categories, however, are not stored in the DCR. In [2, Slide 26], it is argued that "[R]elation types and modeling strategies for a given data category may differ from application to application" and that the "[M]otivation to agree on relation and modeling strategies will be stronger at individual application level". It is concluded that the "[I]ntegration of multiple relation structures in DCR itself" (in addition to the ones already present) could lead to "endless ontological clutter". For the expression of rich

[2] A fifth type of DCs called "container" is rarely used.
[3] In this case, a conceptual domain can be specified for each profile.
[4] Also see diagram at `http://www.isocat.org/12620/model/DCR_data_model_3.svg`



corpus type - 1:0

Key	3822
PID	http://www.isocat.org/datcat/DC-3822
Type	complex/closed
Owner	Hoppermann, Christina
Scope	public

1. Administration Information Section

1.1 Administration Record	
Identifier	corpusType
Version	1:0
Registration Status	private
Administration Status	private
Justification	Common metadata data category
1.1.1 Creation	
Creation Date	2010-11-26
Change Description	Creation of a new data category.
1.1.2 Last Change	
Last Change Date	2011-06-20
Change Description	Changed source of data element name.

2. Description Section

Profile	Metadata
2.1 Data Element Name Section	
Data Element Name	corpusType
Source	CMDI
[-] **2.2 English Language Section**	
Language	English (en)
2.2.1 Name Section	
Name	corpus type
Name Status	preferred name
2.2.2 Definition Section	
Definition	Classification of the type of a corpus.
Source	NaLiDa
[+] **2.3 German Language Section**	

3. Conceptual Domain

Data Type	string
Profile	Metadata
Value	/comparableCorpus/ (comparable corpus)
Value	/generalCorpus/ (general corpus)

Fig. 1. Excerpt of the ISOcat entry for `/corpusType/`

relational structures, a *relation registry* should be used. To provide evidence for these claims, the authors of [8] give a modeling example where the data category "noun" in placed in two different conceptualizations.

Naturally, a definition establishes a relation between the term being defined, *i.e.*, its *definiendum*, and its *definiens*. Users of the ISOcat registry are encouraged to follow guidelines when defining new data categories. The DCR Style Guidelines [1, page 3] make reference to ISO-704 [5], and list the following:

- *They should consist of a single sentence fragment;*
- *They should begin with the superordinate concept, either immediately above or at a higher level of the data category concept being defined;*
- *They should list critical and delimiting characteristic(s) that distinguish the concept from other related concepts.*

Since this encodes the notion of *genus-differentia* definitions, it is clear that any set of (related) definitions induces a concept system. Notwithstanding, the DCR Style Guidelines also point out that "concept systems, such as are implied here by the reference to broader and related concepts, should be modeled in Relation Registries outside the DCR." In line with the policy to disallow (formal) relationships between complex data categories, the DCR guidelines continue saying

> *"Furthermore, different domains and communities of practice may differ in their choice of the immediate broader concept, depending upon any given ontological perspective. Harmonized definitions for shared DCs should attempt to choose generic references insofar as possible."*

This policy can induce quite some tension or confusion. While the definition of a DC must reference a superordinate concept, it should reference a rather generic than a rather specific superordinate concept. Moreover, superordinate concepts in the *definiens* are referenced with natural-language expressions rather than with formal references (say, by pointing to existing terms of the registry).

In the sequel, we will present a reconstruction of a concept system from the many hundred data category entries and their definitions. This concept system then makes formally explicit the relationships between ISOcat terms and the concepts they denote. The concept system could then be seen as a more formal (and complementary) account of the ISOcat registry; the system could, in fact, be understood as the possible content of a relation registry making explicit all relations between the ISOCat Metadata data categories.

3 Expression of ISOcat.org Using Schema.org

3.1 Building the Skeleton

By December 2011, the TDG Metadata consisted of more than 200 simple DCs, more than 200 complex/open DCs, and less than 50 complex/closed entries. To construct the ontology's skeleton, we studied the explicitly given relation structures present between complex/closed DCs and its members of the value range, as well as the existing IS-A relations present between simple DCs. We use OWL for all subsequent modeling.

The use of IS-A constructs in the TDG Metadata gives a mixed picture. Fig. 2 shows the IS-A context of the DC `/translation/`. While most of the relationships are subsumptions (*e.g.*, relating `/gisting/` with `/translation/`), others are not (*e.g.*, relating `/projectManagement/` or `/postProjectReview/` with `/translation/`). For the construction of our skeleton, we have only taken the correct use of IS-A relations into account, anticipating that the standardization process of the ISOcat registry will address the issue of its incorrect uses.

The modeling of the relationships between a complex/closed DC and its value range can be modeled by an OWL class description of the "enumeration" type. Reconsider the ISOcat entry `/corpusType/`, which in OWL can be expressed as:

1. sub DC: /adaptation-1:0/
2. sub DC: /alignedText-1:0/
3. sub DC: /backTranslation-1:0/
4. sub DC: /computerAssistedTranslation-1:0/
5. sub DC: /editedTranslation-1:0/
6. sub DC: /gisting-1:0/
7. sub DC: /globalization-1:0/
8. sub DC: /internationalization-1:0/
9. sub DC: /literaryTranslation-1:0/
10. sub DC: /localization-1:0/
11. sub DC: /machineTranslation-1:0/
12. sub DC: /medicalTranslation-1:0/
13. sub DC: /pivotLanguageTranslation-1:0/
14. sub DC: /postProjectReview-1:0/
15. sub DC: /pre-translation-1:0/
16. sub DC: /projectManagement-1:0/
17. sub DC: /proofreading-1:0/
18. sub DC: /scientificTranslation-1:0/
19. sub DC: /sightTranslation-1:0/
20. sub DC: /sourceText-1:0/
21. sub DC: /specialLanguage-1:0/
22. sub DC: /targetText-1:0/
23. sub DC: /technicalTranslation-1:0/
24. sub DC: /terminography-1:0/
25. sub DC: /transcreation-1:0/
26. sub DC: /translationEditing-1:0/
27. sub DC: /translationMemory-1:0/
28. sub DC: /translationMemoryTool-1:0/
29. sub DC: /translationQualityAssessment-1:0/
30. super DC: /languageMediation-1:0/

Fig. 2. Subsumption hierarchy for the simple DC `/translation/`

```
<owl:Class rdf:ID="Corpus">
  <rdfs:subClassOf rdf:resource="#Resource"/>
  <owl:oneOf rdf:parseType="Collection">
    <owl:Thing rdf:about="#ComparableCorpus"/>
    <owl:Thing rdf:about="#ParallelCorpus"/>
    <owl:Thing rdf:about="#Treebank"/>   [...]
  </owl:oneOf>
</owl:Class>
```

with the individuals `ComparableCorpus`, `ParallelCorpus`, `Treebank` all being instances of the class. The class `Corpus`, thus, is defined by exhaustively enumerating its instances (and its subclass relationship with `Resource`).

Alternatively, we can model complex/closed DCs using a union construct:

```
<owl:Class rdf:ID="Corpus">
  <rdfs:subClassOf rdf:resource="#Resource"/>
  <owl:unionOf rdf:parseType="Collection">
    <owl:Class rdf:about="#ComparableCorpus"/>
    <owl:Class rdf:about="#ParallelCorpus"/>
    <owl:Class rdf:about="#Treebank"/>      [...]
  </owl:unionOf>
</owl:Class>
```

where `ComparableCorpus`, `ParallelCorpus`, `Treebank` are all to be modeled as classes. The latter option is suited when there are relations specific to subclasses of `Corpus`, *i.e.*, with subclasses of `Corpus` as domain or range.

To account for relations implicit in DC definitions, we have been using ISOcat's search functionality to return all DCs where "type" or "class" occurred in either a DC's name (26 DCs returned) or its natural-language definition (55 DCs). From those, the DCs given in Table 1 are a good starting point to bootstrap the

Table 1. Central data categories indicating class hierarchy

Data Category	Definition	Example Section
DC-3806 `/resourceClass/`	Indication of the class, i.e. the type, of a resource.	corpus, lexicon, experiment, tool, grammar, etc.
DC-3822 `/corpusType/`	Classification of the type of corpus.	Value range: `/comparableCorpus/`, `/generalCorpus/`, `/subcorpus/`, `/specialisedCorpus/`, `/other/`, `/learnerCorpus/`, `/monitorCorpus/`, `/unknown/`, `/parallelCorpus/`, `/treebank/`, `/referenceCorpus/`
DC-2487 `/lexiconType/`	A description of the type of the lexicon.	word list, monolingual dictionary, thesaurus, bilingual dictionary, glossary term base
DC-3871 `/experimentType/`	Specification of the design type used for the elicitation of experimental data within a research study, especially in the field of psychology.	experimental design, quasi-experimental design, within-subjects design, between-subjects design, mixed design, pretest-posttest design, laboratory experiment, field experiment, etc.
DC-3810 `/toolType/`	Indication of the type of a tool.	annotation tool, lemmatizer, chunker, segmentation tool, corpus manager, editor, concordancer, tagger, etc.
DC-3900 `/writtenResource-Type/`	The type of the written resource.	primary text, annotation, etnography, study, etc.
DC-3901 `/writtenResource-SubType/`	The subtype of the written resource.	dictionary, terminology, wordlist, lexicon, etc. (if written resource type is LexicalAnalysis).

class hierarchy of linguistic resources, despite the fact that their definitions fail to follow the DCR Style Guidelines promoting *genus-differentia* definitions.[5]

Fig. 3 depicts our initial class skeleton that we derived from the DCs given in the table. Its top class (just below `Thing`) stems from the complex/open DC `/resourceClass/`. The elements cited in its example section, however, should

[5] Two entries have explanation sections adding to their definitions. The explanation section of DC-3900 mentions "type[s] of written resource such as Text, Annotation, Lexical research, Transcription etc", whereas the respective section of DC-3901 mentions that "[d]ifferent types of written resources have different controlled vocabularies for SubType: the type 'Lexical research' has as SubType vocabulary {dictionary, terminology, wordlist, lexicon,...}. In case the Written Resource Type is Annotation the SubType specifies the type of annotation such as phonetic, morphosyntax etc."

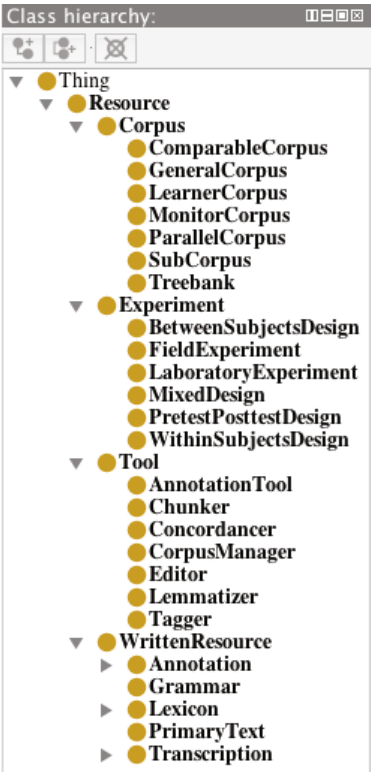


Fig. 3. Class hierarchy for linguistic resources (excerpt)

not always be taken as *direct* subclasses of "Resource", as the definition of `/writtenResourceType/` and `/writtenResourceSubType/` indicate. In fact, all definitions, when taken together, do not give a clear hierarchical picture. The string "lexicon", *e.g.*, is mentioned as type of resource in `/resourceClass/`, and also as a subclass to "WrittenResource" (`/writtenResourceSubType/`). Note that "lexicon" and many other strings appearing in the example sections of the complex DCs are not explicitly defined as a DC. And while the related term "LexicalAnalysis"is mentioned in the definition of `/writtenResourceSubType/` as type of written resource, it not mentioned elsewhere. The situation is similar for the DCs related to experiments. There is no DC with name "experiment", although `/experimentType/` includes a possible definition. None of the strings in its example section has a corresponding DC. Moreover, there is the DC `/blindExperiment/`, but for technical reasons it is not part of the enumeration.

The situation becomes more complex when attaching properties to the class hierarchy. As the ISOcat registry does not make a distinction between concepts and properties, and given the rather fuzzy definition of many DCs, there are no hard criteria other than modeling expertise to follow (see Sect. 3.3).

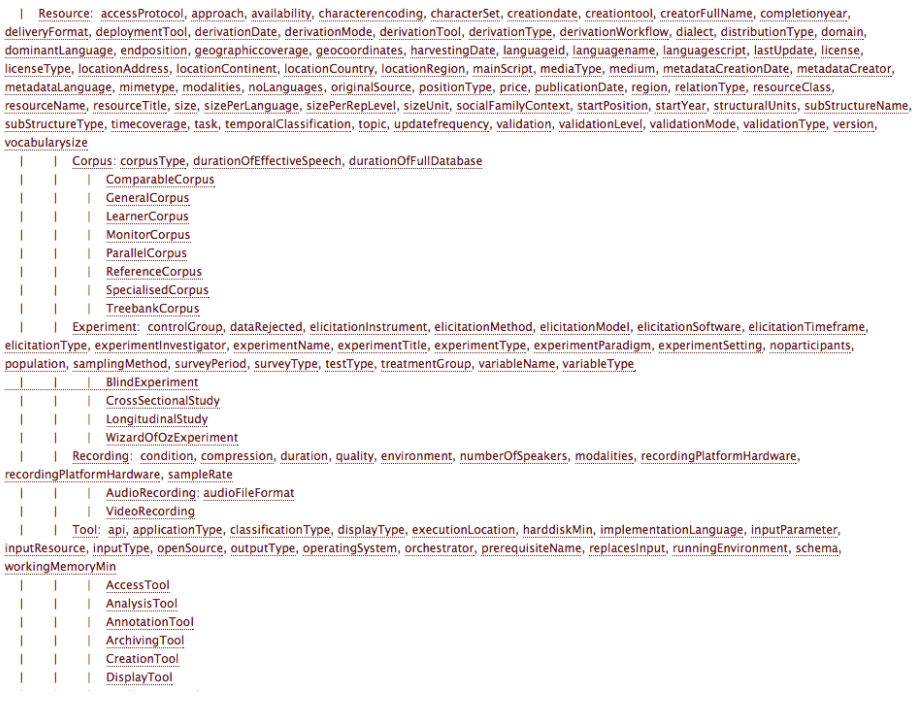


Fig. 4. The ISOcat registry, following the design of `schema.org` (excerpt)

3.2 Re-representation of ISOcat.org (Initial Version)

Fig. 4 depicts one possible hierarchical representation of ISOcat data categories in the proposed new form. It uses the design of `schema.org`, which we found appealing for both its expressive power and simplicity. The new representation gives a structural account of the ISOcat terms. Each concept, relation or instance is linked to the original entry in the ISOcat registry using the DC's persistent identifier. The structure currently accounts for over two thirds of DCs of the TDG Metadata. Those DCs that are not yet included in the hierarchy often have a rather fuzzy definition, not permitting a clear-cut inclusion in the hierarchy.

The benefits of the new representation are rather obvious. Readers get an immediate overview of the different types of linguistic resources and the properties that can be attributed to them. The new representation makes explicit a class hierarchy that is only implicitly present – and scattered – in the ISOcat registry. It differentiates between classes and relations, and attaches the latter to the former. Our re-representation of the TDG Metadata of the ISOcat registry into hierarchical form is preliminary. In the following, we describe the lessons we learned. It is on the maintainers and owners of the data categories to take the lessons into account, and to render their entries more precise and concise.

3.3 Lessons Learned

Rearranging the terms of a glossary into a hierarchy helps to better understand the notion of each term. We hope that our concept scheme informs renewed efforts into achieving a high-quality glossary of linguistic terms.

Lack of Structure. In knowledge representation systems that make use of class hierarchies and inheritance, attributes should be as generic as possible and as specific as required. With the ISOcat registry being designed as a glossary rather than a concept system, information that is usually inherited from superclassses needs to be encoded by extra data category entries. Consider, for instance, the many data categories defined for naming different kinds of entities:

Data Category	Definition
DC-2536 `/projectName/`	A short name or abbreviation of the project that led to the creation of the resource or tool/service.
DC-2544 `/resourceName/`	A short name to identify the language resource.
DC-2577 `/participantName/`	The name of the person participating in the content of the recording as it is used by others in the transcription.
DC-2512 `/creatorFullName/`	The name of the person who was participating in the creation project.

Similar cases are `/experimentName/` (DC-3861), `/scriptName/` (DC-3809), `/substructureName/` (DC-3820), `/countryName/` (DC-3792), `/continentName/` (DC-3791), `/variableName/` (DC-3880), `/prerequisiteName/` (DC-3805), `/participantFullName/` (DC-2556), `/languageName/` (DC-2484), and `/contactFullName/` (DC-2454). While these entries, and others, bear no formal relation to each other in a glossary, from a "concept system" point of view, they should. Following our re-engineering of the ISOcat glossary into an ontology, we would – similar to `schema.org` – attach a general-purpose relation `name` to the top class. With this modeling, the aforementioned DCs can be formally related to the general-purpose `name` relation, defining a property hierarchy:[6]

```
<owl:DatatypeProperty rdf:id="personName">
    <rdfs:subPropertyOf rdf:resource="#name"/>
</owl:DatatypeProperty>
```

DCs related to dates and sizes can be dealt with similarly. The DC `/creationDate/` could be made a sub-property of a (new) DC `/date/`, and in turn, the existing DCs `/derivationDate/`, `/metadataCreationDate/`, `/publicationDate/` *etc.* can be made sub-properties of `/creationDate/`. All properties of this type should inherit from `/date/` its range `Date` from the XML Schema datatype standard. Similarly, if the DC `/size/` (DC-2580) is regarded as general-purpose relation that can be associated with any type of resource, then `/vocabularySize/` (DC-2504), `/sizePerLanguage/` (DC-2581) and `/sizePerRepresentativeLevel/` (DC-2582) could be defined as its sub-properties.

[6] The DC `/description/` (DC-2520) is similar in generality to `name`; it is defined as "a description in general prose text of the issues that are indicated by the context. The description field can occur at many different places in a component and profile."

Lack of Precision. An analysis of the ISOcat content shows a sloppy use of language when defining DC entries.

Naming Policies. The TDG Metadata of the ISOcat registry has many DCs to name entities. While some of these DC terms carry "name" in their name (*e.g.*, `/personName/`, `/projectName/`, others do not:

Data Category	Definition
DC-3793 `/cooperationPartner/`	Naming of the cooperation partner of a research project.
DC-2522 `/funder/`	Name of the funder of the project.

Naming is also an issue when considering other DC subsets such as the pair DC-2568 `/environment/` ("description of the environmental conditions under which the recording was created.") and DC-2696 `/recordingenvironment/` ("the environment where the recording took place"). Here, the name for DC-2568 should be made more specific, say "recordingCondition". Also, for usability reasons, a general policy for naming DCs is advisable and needs to be enforced.

Non-adherence to genus-differentia definitions. The entries given in Table 1, and the many other examples we have given, show that many authors fail to give definitions in line with ISOcat's advocated policy. It is advisable that a TDG's governing body sensitize newly registered users to the importance of good definitions. Sometimes, a user will have difficulty to choose from a pair of semantically similar DCs: while the DC `/address/` is given as "the address of an organization that was/is involved in creating, managing and accessing resource or tool/service" the DC `/locationAddress/` is defined as "the address where the resource was created or originated". Both DCs follow the principle of *genus-differentia*, but the language used is too imprecise to draw a distinction.

Typing. The complex/closed DC-2548 `/anonymizationFlag/` is of datatype boolean; however its value range comprises `/true/` (DC-2952), `/false/` (DC-2953), `/unspecified/` (DC-2592), and `/unknown/` (DC-2591), whereas the XML Schema boolean datatype can only take the values "true" and "false".

Incompleteness. With the class hierarchy giving a birds-eye view on the ISOcat glossary, several gaps can be easily spotted. For many of the main classes, there are no corresponding entries in the ISOcat registry. Entries are missing, for instance, for "resource", "lexicon", "corpus", "experiment" *etc.* although references are made to them in the definition and example sections of many DCs.

There are many minor gaps. There is, for instance, the DC-2689 `/audio-fileformat/`, but there are no corresponding DCs for "videoFileFormat", "documentFileFormat" *etc.* Moreover, there are DCs of type complex/open but their type could be complex/closed. DC-2516 `/derivationMode/`, for instance, could easily be closed by adding the simple DC "semi-automatic" to the existing values "manual" and "automatic".

There are cases where a data category's association with its profiles is incomplete. There is, for instance, DC-2008 `/languageCode/`; it is only associated with the TDG Morphosyntax. Instead of also associating this DC with the TDG Metadata, users have created yet another, but conceptually identical DC, namely `/languageId/` (DC-2482), and have associated it with the TDG Metadata.

Usage of Existing Standards. There is a fair share of data categories that refer to general metadata-related concepts rather than specific linguistic ones. Mapping these DCs to a hierarchy shows that it could share substantial parts with `schema.org`. This includes descriptors that are widely used across many domains, such as metadata about persons, organizations, and places, but also software applications.[7] Take the class `http://schema.org/PostalAddress`, for instance. It serves as an anchor point to address-related properties, most of which with near equivalents in the ISOcat registry: `/locationAddress/` (DC-2528), `/locationRegion/` (DC-2533), `/locationCountry/` (DC-2532), `/locationContinent/` (DC-2531), `/email/` (DC-2521) and `/faxNumber/` (DC-2455). The DC `/address/` (DC-2505) could then be seen as relation with domain `Person` or `Organization`, and range `PostalAddress`.

The designers of `schema.org` propagate the usage of ISO standards whenever possible. While the ISOcat registry already advises to use language (ISO 639-X) and country codes (ISO 3166-1), it could also profit from the inclusion of additional, and widely known, standards. For linguistic resources, the ISO standard ISO-8601 [3] on dates and times is particularly interesting for the description of segments and their duration (intervals) from recordings, transcriptions, annotations *etc.* The current term set on experimental data would profit from consulting (and referring to) an existing ontology for scientific experiments [7]. The ISOcat entries related to media types, file formats, programming languages, software platforms should make reference to existing knowledge sources such as the IANA registry for MIME media types[8], the Wikipedia list on file formats[9], programming languages[10], and operation systems[11].

Moreover, the Semantic Web community, including `schema.org`, is encouraged to give an RDF representation for those lists with persistent identification.

Actions to be Taken. The re-representation of the ISOcat registry of metadata terms unveiled a number of issues that need to be addressed. The most pressing issue is the DC authors' ignorance of recommended good practise when defining entries. Naming conventions and the use of *genus-differentia* need to be enforced. An enforcement will prompt users to add those entries to the registry that we highlighted as gaps, and others. Much can be gained by grouping together DCs that are not explicitly linked together by ISOcat's subsumption relation or the

[7] For the description of software in `schema.org`, please see `http://www.google.com/support/webmasters/bin/answer.py?answer=1645432`.
[8] See `http://www.iana.org/assignments/media-types/index.html`
[9] See `http://en.wikipedia.org/wiki/List_of_file_formats`
[10] See `http://en.wikipedia.org/wiki/Lists_of_programming_languages`
[11] See `http://en.wikipedia.org/wiki/List_of_operating_systems`

relationships between complex/closed DCs and the simple DCs of their value range. Moreover, there is ample opportunity to connect to existing ontologies instead of inventing terminology anew.

We believe that our re-representation addresses these issues. It has the potential to serve the goals of the ISOcat user community; it adds to the precision of the ISOcat metadata-related content and groups together entries that are semantically related; its hierarchical structure gives users a birds-eye view to better access and manage a large repertoire of expert terminology.

4 Discussion

4.1 Expert Vocabulary: From Glossary to Ontology

The ISOcat registry is designed as a glossary of terms, and this design can quickly be understood by a large user base without expertise in knowledge representation. Users can easily define a data category whenever they believe such an entry is missing. ISOcat's ease-of-use is also its fundamental shortcoming, however. The definition of a DC is given in natural language, and hence, is inherently vague and often open to multiple interpretations. Also, ISOcat entries vary in style and quality, given the collaborative authoring effort. The increasing size of the TDG Metadata, now containing more than 450 terms, its glossary-like organization, the current data curation policy of the registry – authors can only modify the entries they own – may prompt users to rashly define their own data category instead of identifying and re-using an appropriate existing one. Nevertheless, it is hoped that a standardization process, once set in motion, will lead to an expert vocabulary most linguists agree upon.

It is clear that the definitions of the ISOcat metadata terms spawn a concept system. Simple DCs are related to complex DCs because they appear in the value range of the latter, and it is also possible to define subsumption relations between simple DCs. Moreover, *genus-differentia* definitions relate to each other *definiendum* and *definiens*. The non-adherence of authors to good practise when defining, potentially prompted by a policy that disallows formal relationships between complex DCs, is responsible for many of the weaknesses identified.

It is argued that relationships between complex DCs should be represented in a *relation registry* [8]. DCR authors are encouraged to keep the definitions of their entries deliberately vague so that this vagueness can then be addressed – in varying manner – externally by using the relation registry. While the relation registry is currently used for the mapping of terms from *different* TDGs or vocabularies (using SKOS-like relation types such as *broader* and *narrower*, see [6]), we find it questionable whether this is a viable approach for intra-vocabulary mapping within the TDG Metadata. It would be unclear, *e.g.*, how to draw the line between explicitly and implicitly defined relations in the ISOcat data category registry and those defined in the relation registry, and the possible confusion it creates when the registries' content is in contradiction to each other.

In fact, the concept scheme we derived from our analysis could be seen as an incarnation of the relation registry. But in light of the previous discussion, it must be an officially sanctioned one, aiming at giving an adequate account of ISOcat metadata-related content.

4.2 Impact on Existing Metadata Infrastructure

The concept scheme can serve as a tool to better browse and manage the ISOcat term registry for metadata. It can inform curation efforts to render precise the definition of existing entries, or to create new entries to fill the gaps made obvious by our ontological reengineering. For this, the concept scheme and the ISOcat registry need to be synchronized. This can be achieved by enforcing the policy that authors of new DCs must somehow provide anchor points that link a DC to a node in the hierarchy. Reconsider the entry `/resourceClass/` (cf. Table 1, page 7). It could be "semantically enriched" by making explicit the class hierarchy that is only implicitly given in the informal language definition of the entry: `Resource is a class. Corpus is a subclass of Resource. Lexicon is a subclass of Resource` *etc.*

The semantic enrichment of the DC's definition could then prompt users to create entries for "corpus", "lexicon", "experiment" *etc.* Alternatively, and more in line with common usage in many dictionaries, users could be encouraged to associate the term being defined with broader, narrower, or related terms.

We hope that our concept scheme serves as a starting, reference and entry point to the content of the ISOcat metadata-related vocabulary. For this, it needs to be "in sync" but also officially sanctioned to better reflect, at any given time, the content of the ISOcat registry. Our concept scheme, when understood as a "relation registry", has the advantage that – by following `schema.org` and its OWL version (see `http://schema.org/docs/schemaorg.owl`) – it is based on existing, open, and widely-used W3C standards. Future work will address how to best profit from this technology in terms of sharing vocabulary with `schema.org` and distributing metadata about linguistic resources using microformats.

5 Conclusion

The ISOcat registry has taken a central role in those parts of the linguistics community that care about metadata. Its low-entry barrier allows users to contribute towards a set of terms for the description of linguistic resources. The ISOcat registry will continue to serve this role, but the registry and its users can profit from the provisions we have outlined. With the re-representation of the ISOcat metadata registry into a hierarchical structure, we have gained a birds-eye view of its content. Our work unveiled current shortcomings of the ISOcat registry from a knowledge representation perspective, where class hierarchies are often constructed centrally and in a systematic and top-down manner.

Many of the problems that we have highlighted are typical for distributed work on a lexicographic resource; here, contributors often take a local stance asking

whether a glossary contains a certain term suitable for some given application of the term, or not. With a glossary growing to many hundred entries, it is not surprising that there will be two or more entries denoting the same concept (synonymy), or two entries sharing the same data category name having different (homonymy) or only partially overlapping (polysemy) meanings, *etc.*

A large part of our critique could be addressed by pointing out the "private" nature of the DCs in the TDG Metadata. Once the standardization process of the data categories gains traction, many of the issues can be addressed and solved. We believe that our ontological approach would greatly support this process. DCs owners are encouraged to consult our formalization and check whether their entries can be improved by the birds-eye view now at their disposition. The standardization body is encouraged to take our ontology to identify "important" DCs and schedule them for standardization. For users of the registry, it serves as efficient access method complementing existing search and browse functionality.

Our hierarchy is one of possibly many interpretations of the ISOcat metadata registry. With on-going work on the registry, it will need to be revised accordingly. Note that we do not seek to replace the TDG Metadata with the ontology we have reconstructed by interpreting its content. It is intended to support existing workflows in order to obtain an ISOcat-based metadata repertoire that progresses towards the completeness and high-quality of its entries.

The url `http://www.sfs.uni-tuebingen.de/nalida/isocat/` points to the current version of the hierarchy. Feedback is most welcome!

References

1. DCR Style Guidelines. Version "2010-05-16", `http://www.isocat.org/manual/DCRGuidelines.pdf` (retrieved December 5, 2011)
2. Data Category specifications. Clarin-NL ISOcat workshop (May 2011), `http://www.isocat.org/manual/tutorial/2011/ISOcat-DC-specifications.pdf` (retrieved December 5, 2011)
3. Int'l Organization of Standardization. Data elements and interchange formats – Information interchange – Representation of dates and times (ISO-8601), Geneva (2009)
4. Int'l Organization of Standardization. Terminology and other language and content resources - Specification of data categories and management of a Data Category Registry for language resources (ISO-12620), Geneva (2009)
5. Int'l Organization of Standardization. Terminology work – Principles and methods (ISO-704), Geneva (2009)
6. Schuurman, I., Windhouwer, M.: Explicit semantics for enriched documents. What do ISOcat, RELcat and SCHEMAcat have to offer? In: Proceedings of Supporting Digital Humanities, SDH 2011 (2011)
7. Soldatova, L.N., King, R.D.: An ontology of scientific experiments. Journal of the Royal Society Interface 3(11), 795–803 (2006)
8. Wright, S.E., Kemps-Snijders, M., Windhouwer, M.A.: The OWL and the ISOcat: Modeling Relations in and around the DCR. In: LRT Standards Workshop at LREC 2010, Malta (May 2010)