Segueing from a Data Category Registry to a Data Concept Registry

Sue Ellen Wright, Menzo Windhouwer, Ineke Schuurman, Daan Broeder

Segueing from a Data Category Registry to a Data Concept Registry

Sue Ellen Wright, Menzo Windhouwer, Ineke Schuurman, Daan Broeder

Kent State University
The Language Archive – DANS
CLARIN-NL – KU Leuven
The Language Archive – MPI

sellenwright@gmail.com
menzo.windhouwer@dans.knaw.nl
ineke@ccl.kuleuven.be
daan.broeder@mpi.nl

Abstract.
The terminology Community of Practice has long standardized data categories in the framework of ISO TC 37. ISO 12620:2009 specifies the data model and procedures for a Data Category Registry (DCR), which has been implemented by the Max Planck Institute for Psycholinguistics as the ISOcat DCR. The DCR has been used by not only ISO TC 37, but also by the CLARIN research infrastructure. This paper describes how the needs of these communities have started to diverge and the process of segueing from a DCR to a Data Concept Registry in order to meet the needs of both communities.

Keywords: data categories, concepts, semantic registries, communities of practice

1 Introduction

For more than a decade now ISO TC 37 has been transitioning from a static paper-based list of data categories for terminology management (ISO 12620:1999) to a more dynamic Data Category Registry (DCR, i.e., http://www.ISOcat.org) designed to serve a broad range of language resource communities [1]. This paper describes these communities and their needs and how they are leading to a new vision.

Terminology management and concept registries have been developed by a variety of Communities of Practice (CoP). Efforts have been made to characterize these CoPs over the last decade and to create a taxonomy of knowledge organization resources [2, 3, 4, 5, 6, 7]. In the context of the ISOcat repository, we distinguish the following:

Discourse-purposed terminology (and concept) management: Lexicographers document the many definitions associated with *words* and special language *terms*, using the head word as their core element, to produce monolingual dictionaries and multilingual glossaries. In contrast, terminologists write careful definitions to document

concepts in special fields and link them to the many designations (terms, synonyms, multilingual equivalents, formulae, symbols, etc.) associated with each concept. They often produce multilingual resources where lexical approaches pose problems for semantic mapping of concept systems and interlingual equivalencies. Despite differences, lexicographers and terminologists provide linguistic and semantic information *for humans using language to speak and write. These terms are elements of discourse.*

Subject-purposed terminology (index languages): Librarians and archivists use terms and definitions to create controlled vocabularies, classification systems, subject catalogues, and thesauri in order to document knowledge and objects in collections and archives (hence the term: *documentary languages*). *Terms* are used to retrieve known objects, such as books or art works in a collection, but they are now also used for information retrieval from open heterogeneous archives. Discourse-purposed term/concept pairs function differently from subject-purposed terms. Svenonius writes: "In an index language the naming function of terms works somewhat differently from the same function in ordinary language. In ordinary language, the word 'butterflies' has as its denotational … meaning, that is, its referent, the set of all butterflies, past, present, and future, real or imaginary. In an index 'butterflies' names a subject and its denotational meaning is the set of all documents about butterflies" [8]. Here, *terms* are identifiers used to retrieve objects or information.

Data dictionaries: Here *terms* are called *data element names* designating elements in data or metadata models, e.g., database schemas or tagging systems for tagged corpora. Together with their underlying *data element concepts*, they are defined in conjunction with conceptual domains, i.e., their permissible values. The combination of data model and data category (DC) specifications is used to create, retrieve, and map elements for developing and sharing compatible resources. Data dictionaries (DDs) vary in scope and purpose, from very specific DDs that describe shared application models and elements to Metadata Registries (MDRs). At their most rigorous, DDs prescribe data types, data elements, and enumerated values, in order to facilitate precise data interchange and interoperability. At their most flexible, DDs focus on semantic content in order to retrieve and integrate data from heterogeneous sources.

Semantic Web and Linked (Open) Data (LOD): the LOD approach connects distributed data sets over the web by sharing URIs. Data are represented using RDF/ RDFS-based languages. Semantic Web technologies, such as OWL and SKOS, are used to represent knowledge and/or thesauri and other controlled vocabularies, potentially enabling automated reasoning on top of LOD. Here *terms* act as classes and properties in knowledge and/or data representation systems.

All these approaches used by the various CoPs share the need to describe the semantics of *terms* so users can determine whether a *term* applies to a given use case. The more data-oriented approaches also provide representation information, e.g.: does one *term* have values (a conceptual domain) or is it a value in such a domain? For instance, does */grammatical gender/* have *masculine* and *feminine* as values, or *neuter* as well? This paper describes the data-oriented ISOcat DCR and how its use by various CoPs steers it towards becoming a Data Concept Registry.

2 The ISOcat Data Category Registry

DCs have a long history especially in the ISO TC 37 community [9]. This section describes this and more recent history revolving around the development of the ISOcat DCR [10] at the Max Planck Institute for Psycholinguistics (MPI-PL) and its use in the wider community, especially CLARIN.[1]

2.1 ISO TC 37, Terminology and other language and content resources

The evolution of the DCR reflects the convergence of multiple purposes and subsets of experts within the framework of the broader community of practice represented by resource and application developers in linguistics and the social sciences.

ISO TC 37 specifies DCs for use in terminology databases and (as expanded in ISOcat) as tags for marking up language resources. The documentation of standardized DC names (and originally, their abbreviations) began when terminologists were still recording information on paper *fiches*. Computerized DDs led to initial efforts to collect and document data element concepts associated with terminology management as a part of the development of a terminology interchange format (originally called MARTIF, then XLT, and now known as the TermBase eXchange format or TBX) during the SALT project (Standards-based Access service to multilingual Lexicons and Terminologies; see [11, 12]). After evolving through the SYNTAX pilot project [13, 14], this effort emerged as a Metadata Registry in the sense of the ISO 11179 family of metadata standards [15], called in TC 37 parlance a *Data Category Registry*.

The primary focus of this effort was originally the definition of DCs representing data element concepts used as semantic units in terminological databases, such as *term, part of speech, definition.* These elements are used in modeling and creating databases, and in manipulating data in exchange environments requiring interoperability, not only in terminology management, but also in a variety of text and corpus annotation frameworks, such as syntactic or semantic information. As a consequence, they are rigorously defined and generally conform to a variety of metamodels ([11, 16, 17, 18], etc.). Given the metamodels used in the various environments, definitions created for use with these resources are ideally rigorously linked to their respective metamodels and reflect relationships, particularly between parent and child DCs, expressed in the DCR as open, closed, simple, and constrained DCs (see Section 3.1).

2.2 MPI-PL and CLARIN

The Max Planck Institute for Psycholinguistics has a long history in cooperation with ISO TC 37. During the LIRICS project they developed a web-based lexicon tool to support the Lexical Markup Framework (LMF), while INRIA created the SYNTAX DCR [19, 20]. When LIRICS ended, the MPI-PL started developing ISOcat as the successor to SYNTAX. Around the same time, the preparatory phase of the CLARIN

[1] A research infrastructure for scholars in the human and social sciences, cf. http://www.clarin.eu

infrastructure started and the ISOcat DCR was introduced as a foundation for semantic interoperability. One of the aims of the European CLARIN infrastructure is to allow scholars to easily find and integrate language resources (LR) and language technology (LT) from a wide range of sources. For this purpose CLARIN set out to develop (1) a joint domain of LR & LT metadata and (2) a federated content search domain allowing users to perform queries on corpora of annotated texts or media housed at different sites in parallel. Differences between sub-community descriptive terminologies dictate that CLARIN address semantic interoperability.

In the description of terminology and corpus management models cited above, interoperability involved adherence to shared metamodels, but in the CLARIN context, interoperability is not so much a function of compatible data design, but rather of data retrieval from potentially heterogeneous resources. In this environment, differences in data description require reinterpretation when retrieved data from different sources are to be processed as one set or when they have to be semantically 'normalized' for a specific tool, although the community is encouraged to use one of the various available description standards, cf. above. The ISOcat DCR has been used in this infrastructure as a recommended resource for the purpose of providing linkage between the heterogeneous (meta)data models in order to enable integrated data retrieval.

3 Converging and Diverging Communities of Practice

3.1 The ISO TC37 CoP

As noted, in ISOcat DCR practice, DCs specify a data element name assigned to the definition of a data element concept. As such, they play an important role in terminology management and the creation of annotation schemes used to mark up text and speech corpora. The evolution of ISOcat coincided with an expansion of ISO TC 37 to include a range of 'other language resources', many of which share DCs across sub-communities of practice. In its original configuration, the data and organizational model of the DCR was designed to comply with then-current ISO directives pertaining to the standardization of concept-related items cited in ISO standards. This approach dictated the strict identification of so-called Thematic Domain Groups (TDGs). Only a few of these established expert groups have become active:

- Terminology – ISO 12620 and 30042
- Morphosyntax – LMF [20]
- Metadata – CMDI (see Section 3.2; [21])

The ISO requirements cited above imposed a complex standardization process on both the theoretical framework of the DCR and (perhaps more importantly or even unfortunately) on the actual data model and instantiation of the resource. In practice, these structures have proven not only unworkable in terms of human computing conventions, but also unwanted because no actual DC standardization has taken place inside the DCR. Instead, Data Category Selections (DCSs) specified for any given sub-CoP are being simply listed in the related standards [17]. Within the DCR itself,

consensus-based recommendations have proven more effective than formal balloting procedures as prescribed in the now-rejected cumbersome ISO approach.

The original ISOcat design was wedded to the terminological view of linguistic data and categorizes DCs based on their function(s) in various metamodels:

- Open DCs that can take values that conform to the abstract definition of the DC (example: /*writtenForm*/ (isocat.org/datcat/DC-1836));
- Complex DCs, subdivided into:
    - Closed DCs whose values are constrained to an enumerated set of values (examples: /*part of speech*/ (isocat.org/datcat/DC-1345));
    - Simple DCs, which serve as those enumerated values (example: /*adposition*/ (isocat.org/datcat/DC-1231));
- Constrained DCs whose values are defined by automatically parsable rules (example: /*breath alcohol concentration*/ (isocat.org/datcat/DC-4359) specifies a regular expression to limit the value domain);
- Container DCs, which can be used as high-level container components in compliance with various metamodels (example: /*descrip*/ (isocat.org/datcat/DC-3868)), whereby *descrip* can contain multiple other DCs, such as */definition/*, */context/*, */source/*, */note/*, etc.

For the terminology community in particular, the relationships between closed and declared simple DCs is critical to ensure rigorous interoperability in industrial environments. The DCR was originally designed to allow for multiple sets of enumerated values depending on the requirements of different sub-communities, but the need to declare data types and data element categories imposes unwanted constraints for users who may want to use specific data concepts in a variety of ways. For instance, *noun* can function as a simple DC dependent on the parent *part of speech* in one environment, but in another it might have its own sub-categories, e.g., *proper, common, count, mass, etc.* These concerns suggest that the relations currently expressed in the DCR be moved outside the system to external Relation Registries (RRs [22]), so that DCs within the DCR would be unconstrained by these relations (see Section 4.2).

3.2 The CLARIN(-NL) CoP

With respect to the current recommended use of the ISOcat DCR for LRs and LT in the CLARIN domain, we must distinguish between instances of LR & LT metadata and DC use within LR content such as annotations. For the CLARIN joint metadata domain, the Component Metadata Infrastructure (CMDI [21]) actually references ISOcat using links to ISOcat Persistent Identifiers (PIDs) [23]. CMDI allows CLARIN to deal with the wide variety of metadata needs within the LR and LT domain. DC references have been used there to indicate semantic overlap between metadata components, elements and values. Tools like the Virtual Language Observatory (VLO) [24, 25] metadata catalogue use such references to do semantic mapping for kindred metadata attributes. However, similar use of such references for LR content schemas has not progressed far, partially because of the problems of exhaustively describing all LR content schemas used, and partly because providing accurate DC specifications is a hard task, as explained below.

The core of CMDI consists of reusable components. These components group metadata elements and possibly other components, which are managed by a Component Registry (CR). To describe a resource type, a metadata modeler combines components from the CR into a metadata profile. Due to the flexibility of this model, the metadata structures can be very specific to an organization, project or resource type. Although structures can thus vary considerably, they are still within the domain of metadata for linguistic resources and thus share many key semantics. To deal with this variety, general CMDI tools, e.g., the VLO, operate on this shared semantics layer. To establish these shared semantics, CMDI components, elements and values can be linked to concept registries. The major concept registries currently used by CMDI are the Dublin Core metadata elements and terms [26] and the ISOcat DCR. While Dublin Core is closed, ISOcat is an open registry, which means that anyone can register new concepts as needed. Recent visualization experiments have shown an increasing amount of semantic overlap between various sub-communities in the CLARIN joint metadata domain [27, 28].

In principal mapping capability is good between the building blocks of CMDI and Data Category types:

- Components can be linked to container DCs;
- Elements can be linked to complex DCs;
- Values can be linked to simple DCs.

The CR edit utility attempts to adhere to this mapping if one uses the CR's ISOcat search interface, but it has always been possible to override this feature and include any concept or DC reference, which has resulted in a growing type mismatch between the content of the CR (components, elements and values) and their referenced DCs:

- 165 elements and 72 components are linked to simple DCs;
- 778 components are linked to complex DCs;
- 4 elements are linked to container DCs. [2]

These data indicate that the metadata modelers assessed the applicability of a DC based on the semantic specification only and did not take the associated representation information, i.e., the data category type, into account. To map totally compatible DCs to the content of the CR, in some cases it becomes necessary to create DCs that are semantically redundant, but that are assigned to different DC types (e.g., *noun* as an open DC, or *noun* as a simple DC that is a value of *part of speech*. This practice can lead to significant proliferation in the DCR, which would also make it harder for users to select the proper DC. Current practice makes it impossible to rely on the typing info in the DCR; instead generic tools (e.g., VLO) rely on inspecting the CMDI profile metadata schema to determine the actual status in a given use case. Hence, the insight has been growing that this typing is, for CLARIN's purposes, counterproductive in the registry, as it can always be, and can better be, gleaned from the actual application involved. This circumstance supports the notion of removing DC typing

[2] Statistics from December 2013. Thanks to Matej Durco.

from the DCR proper and moving this information to specialized RRs for those communities that rely on type categorizations in compliance with ISO 11179.

In addition to metadata DCs, CLARIN-NL has also created DCs for resource content, which is even more diverse, meaning that DC typing in this area can lead to even more proliferation. In many cases non-technical domain experts are asked to create or select relevant DCs, and for them the more technical details of a good specification, e.g., DC type and data type, are very hard. Within CLARIN-NL, and increasingly throughout the broader CLARIN community, these users are supported by an ISOcat content coordinator. She informs them about good patterns, reviews specifications and selections, and recommends DCs for reuse. Nevertheless, the complexity of the current DCR data model and its management processes has become a burden [29].

4 A new focus for ISOcat

The previous sections have shown that there are many problems with the current ISOcat setup. This section describes a leaner focus for ISOcat, while still providing modalities for expressing the additional information that some users need.

4.1 Towards a Data Concept Registry

In general, for communities that are not able to provide expert terminologists, using a complicated model such as ISO 12620 for DCs has proven unusable. Currently CLARIN is investigating alternatives based on a simpler data-model, which is more focused on specifying Data Concepts and leaves the representation of these concepts to the data models. To summarize developments up to this point, we have seen a divergence between the needs and applications of sub-communities using the DCR, specifically between the ISO TC 37 and the CLARIN communities. Where TC 37 experts may need specifically constrained data categories with strictly specified data types and DC types, CLARIN now realizes the need for a repository of data concepts that are unencumbered by the constraints of any specific data modeling environment. Instead of using the DCR as a prescriptive tool for data modelers who need rigorous data definitions, CLARIN users are better served by more semantically suggestive information units. This means that for each data element concept, we need to create a concept specification with a reusable definition, but without the constraints of declaring data type and DC type. This transition in needs also dictates an evolution in the criteria required for writing adequate definitions, which means not only that definitions must be well-conceived (which is not always the case in the current DCR), but they should also be less dependent on any one view of individual data concepts, thus making them "reusable" across applications. Adding to this, the CLARIN community has only fully realized its requirements and also the limitations of the CLARIN community involvement in the last few years. While the current configuration has seemed clear for the terminology community, it has not been truly integrated into the work of other sub-groups within TC 37. So the divergence between the various groups has only come to light with the coming of age of CLARIN and the evolution of TC 37.

The CLARIN CoP (see Section 3.2) clearly sees the need to focus its efforts on describing the concepts underlying the (meta)data of LRs and LT, and hence the need to relieve the registry of the complexity in the data model associated with the assignment of DC types. The future plan is to create an optional open or free area in the data concept specification where it is still possible to retain and add this kind of information. The core registry will not interpret this optional information, which is left to the communities that need to use it.

Furthermore, as the ISO standardization process has stalled, a community-based recommendation system has already been put into place, which is seen as an easier way to help users select or create data concepts appropriate for their resources or tools. The system provides the ability for multiple sub-communities, including ISO TC 37, to designate individual DCs as "recommended".

4.2 Relation Registry

Ontological relationships between DCs had already been banned from the DCR data model in its early design stages. This was due to the fact that these relationships are heavily context dependent, i.e., they change with regard to application context or domain. Despite this rule, relationships between closed Data Categories and simple Data Categories have always been a core part of the data model. However, the CLARIN experience has shown that even these relations are also very context-dependent, i.e., different applications need different value domains, and in the current system it is hard to extend these domains due to DC ownership or the fear of further proliferation within the registry.

The Relation Registry (RR) was originally envisioned as a way to align multiple DCRs (once those would start to appear), but due to the increasing amount of proliferation in ISOcat itself, such a Relation Registry is even necessary when there is only one DCR, be it a Data Category Registry or a Data Concept Registry. In addition to (loose) equivalence ((quasi-) same-as) relationships, it has been clear that other ontological relationships could be stored as well, e.g., generic and partitive relationships [22, 30]. But the RR can also be a place to store the value domain relationships, which are currently stored in ISOcat. The combination of information from both the Data Concept Registry and the RR can thus result in a complete DC specification. They can even be broader, covering full taxonomies [31, 32], and may be even be configured as ontologies [33]. Placing these resources outside the DCR proper also accommodates the reality that different users may wish to produce different ontological systems using the DCs.

5 Conclusion and future work

The first stage in developing a Data Concept Registry appropriate for our needs involves (hopefully) finding the ideal off-the-shelf, semantically oriented software platform that can be used to meet our requirements or can be modified with minimal investment. In parallel, we must complete the development of a rich, user-friendly RR

utility to support the supplemental definition requirements of those who want to use the DCs for more rigorous data modeling. Conceptually, RR software could run in multiple specialized environments in the periphery of the Data Concept Registry under the control of the individual sub-CoPs who need this capability. Also in parallel, CoPs should be encouraged to expand their recommendations with the Data Concept Registry and to improve DC definitions in order to enhance the perceived value of the resource. Finally, when the new configuration is clearly defined and in place, the old ISO 12620:2009 must be revised accordingly.

References

1. ISO 12620: Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources (2009)
2. Souza, R. R., Tudhope, D., and Barcellos Almeida, M.: Towards a taxonomy of KOS: Dimensions for classifying Knowledge Organization Systems. In: Proceedings of the ISKO Conference 2010, http://mba.eci.ufamg.br/downloads/Souza_Tudhope_Almeida_KOS_Taxonomy.Submitted.pdf (2010)
   Hlava, M. M. K.: Insuring Compatibility and Crosswalks, http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2007/presentations/NKOS%202007-HLava.ppt (2007)
3. Hodge, G.: Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. http://www.clir.org/pubs/abstract/pub91abst.html (2000)
4. Tudhope, D. A.: tentative typology of KOS: towards a KOS of KOS? In: NKOS Workshop, ECDL. (2006)
5. Wright, S.E.: Typology for Knowledge Representation Resources. In: NKOS-CENDI http://nkos.slis.kent.edu/2008workshop/SueEllenWright.pdf (2008)
6. Zeng, M. L.: Knowledge Organization Systems. In: Knowledge Organization. vol. 35(2-3), pp. 160-182 (2008)
7. Svenonius, E.: Access to Nonbook Materials: The Limits of Subject Indexing for Visual and Aural Languages. In: Journal of the American Society for Information Science, vol 45(8), pp. 600-606 (1994)
8. Wright, S.E.: Data Categories for Terminology Management. In: The Handbook of Terminology Management, Amsterdam and Philadelphia: John Benjamins Publishing Company, pp. 552-571 (2001)
10. ISOcat: Data Category Registry, Defining widely accepted linguistic concepts. http://www.isocat.org
11. ISO 30042: Systems to manage terminology, knowledge and content – TermBase eXchange (TBX) (2008)
12. Melby, A.K.: SALT: Standards-based Access service to multilingual Lexicons and Terminologies. http://www.ttt.org/salt/description.html
13 Ide, N. and Romary, L.: A Registry of Standard Data Categories for Linguistic Annotation. In: Proceedings of the 2004 LREC Conference (2004)
14. Wright, S.E.: A Global Data Category Registry for Interoperable Language Resources. In: Proceedings of the 2004 LREC Conference (2004)
15. ISO/IEC 11179: Information Technology – Metadata registries (MDR). http://metadata-standards.org/11179/
16. ISO 24613: Language resource management – Lexical markup framework (LMF) (2008)

17. ISO 24611: Language resource management – Morpho-syntactic annotation framework (MAF) (2012)
18. ISO 24612: Language resource management – Linguistic annotation framework (LAF) (2012)
19. LIRICS. http://lirics.loria.fr
20. Francopoulo, G., ed.: LMF Lexical Markup Framework. Hoboken, John Wiley & Sons Inc./ISTE (2013)
21. Broeder, D., Kemps-Snijders M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., Zinn, C.: A Data Category Registry- and Component-based Metadata Framework. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Malta, May 19-21 (2010)
22. Windhouwer, M.: RELcat: a Relation Registry for ISOcat data categories. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, May 23-25 (2012)
23. ISO 24619: Language resource management – Persistent identification and sustainable access (2011)
24. Virtual Language Observatory (VLO). http://www.clarin.eu/vlo
25. Uytvanck, D. Van, Zinn, C., Broeder, D., Wittenburg, P., Gardellini, M.: Virtual language observatory: The portal to the language resources and technology universe. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Malta, May 19-21 (2010)
26. Dublin Core. http://dublincore.org/
27. Durco, M., Windhouwer, M.: Semantic Mapping in CLARIN Component Metadata. In: E. Garoufallou and J. Greenberg (eds.), Metadata and Semantics Research (MTSR 2013), CCIS Vol. 390, Springer, Thessaloniki, Greece, November 20-22 (2013)
28. Durco, M., Windhouwer, M.: CMD Cloud. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, May 28-30 (2014)
29. Broeder, D., Schuurman, I, Windhouwer, M.: Experiences with the ISOcat Data Category Registry. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, May 28-30 (2014)
30. Windhouwer, M., Schuurman, I.: Linguistic resources and cats: how to use ISOcat, RELcat and SCHEMAcat. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, May 28-30 (2014)
31. Nistrup Madsen, B.; Erdman Thomsen, H, Lassen, T., Pram Nielsen, L., Odgaard, A.E., Lyngby Hoffmann, P.: Towards a New Taxonomy of Terminological Data Categories. In: eDITion : Fachzeitschrift für Terminologie, Vol. 9, No. 1, p. 18-24 (2013)
32. Zinn, C., Hoppermann, C., Trippel, T.: The ISOcat Registry Reloaded: a Re-engineering Proposal Following schema.org. In Proceedings of the 9th Extended Semantic Web Conference (ESWC 2012), Lecture Notes in Computer Science (volume 7295/2012), The Semantic Web: Research and Applications, pp. 285-299, DOI: 10.1007/978-3-642-30284-8_26. Berlin: Springer (2012)
33. Nistrup Madsen, B.; Erdman Thomsen, H, In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, May 28-30 (2008)

Abbreviations

CLARIN	Common Language Resources and Technology Infrastructure
CMDI	Component Metadata Infrastructure
CoP	Community of Practice
CR	Component Registry
DC (DCs)	data category, -ies
DCR	Data Category Registry
DCS	Data Category Selection
DD (DDs)	data dictionary, -ies
INRIA	*Institut national de recherche en informatique et en automatique*
ISO	International Organization for Standardization
LIRICS	Linguistic Infrastructure for Interoperable Resources and Systems
LMF	Lexical Markup Framework
LOD	Linked Open Data
LR	language resource
LT	language technology
MARTIF	Machine-Readable Terminology Interchange Format
MPI-PL	Max Plank Institute for Psycholinguistics, Nijmegen
OWL	Web Ontology Language
PID	Persistent Identifier
RDF/RDFS	Resource Description Framework/RDF Schema
SALT	Standards-based Access to multilingual Lexicons and Terminologies
SKOS	Simple Knowledge Organization System
SYNTAX	(DCR pilot project)
TBX	Termbase eXchange Format
TC	Technical Committee (ISO)
TDG	Thematic Domain Group
URI	Uniform Resource Identifier
VLO	Virtual Language Observatory
XLT	eXchange format for Lex/Term data