



ISO TC37/SC4 Infrastructure Note on Registry Databases

Peter Wittenburg, Sue Ellen Wright (Ed.)¹

It is widely understood that one of the major challenges of our time is to achieve semantic interoperability at various levels. Ontologies of different types will play the primary role in meeting these needs. ISO TC37/SC4, together with representatives of other TC 37 SCs, has made a wise choice by deciding to separate the concept definition component of an ontology from the component or components that store relations between and among these concepts. The main reason for this decision is that it is very difficult indeed for even closely related communities of practice to achieve wide-spread consensus on the definition of domain-specific concepts that are used for tagging linguistic resources for semantic knowledge. Adding mechanisms and structures for storing relational information would add another layer of complexity. Nevertheless, as has been shown by TC37's accomplishments in elaborating specifications for data element concepts (data categories) in the framework of the Syntax Data Category Registry (DCR) project, the definition task seems to be doable if kept within manageable bounds.

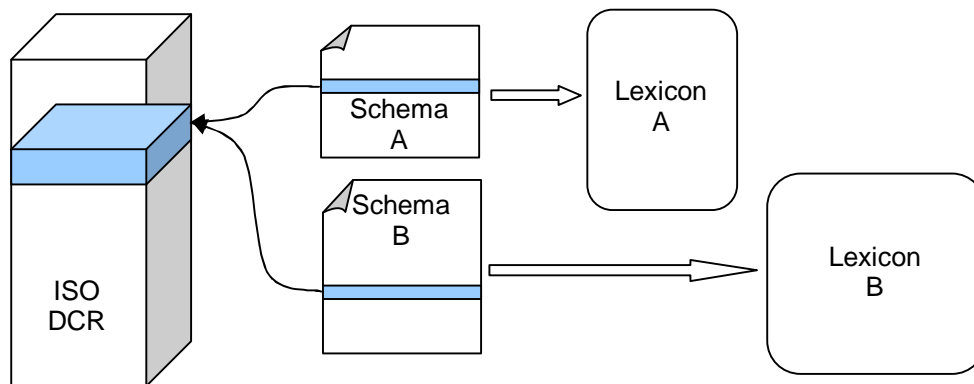
In the absence of a mature set of well-defined, widely accepted concept entries, the problems involved in relating concepts to one another, including classifying them and ordering them in semantic networks that can be used for inferencing seem insurmountable at this moment due to differences in language, differences in linguistic theories and the concrete practical purposes for which inferencing is required.

Using Data Categories

By positioning the relational aspects of the DCR outside the definitional core, it becomes possible to leave it to the users to determine how they want to include and use the semantic knowledge that is manifested in the DCR, with the expectation that when two linguists refer to the same entry in the DCR, they want to express a certain semantic relation with this referenced entry. This scenario could, for instance, involve two lexical schemas defining classes of lexica that make use of the entry as depicted in the following figure, i.e., to add a lexical attribute that is taken from the DCR. This kind of coordinated use of the same linguistic data category in both lexica enables dynamic searching across the two resources without additional effort. The search engine just needs to "know" to exploit the reference contained in the schemas to achieve the effect of a dynamically federated resource.

Figure 1: Federated lexicon searching via shared, schema-enabled metadata reference

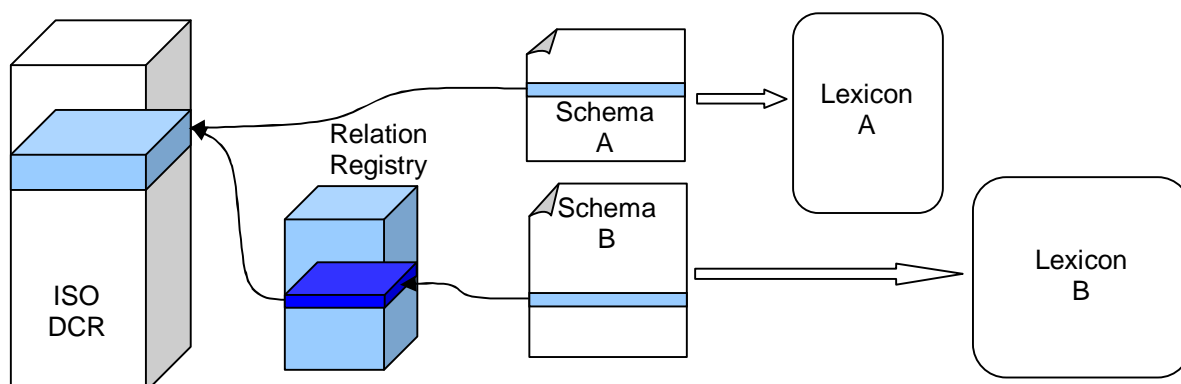
¹ This note emerged as a response to a discussion at the ISO meeting in Provo and due to several CLARIN project meetings. Hasida Koiti, Nicoletta Calzolari, Sue Ellen Wright, Laurent Romary, Andreas Witt, Gerhard Budin, Daan Broeder, Marc Kemps-Snijders and Menzo Windhouwer were closely involved in the discussions and the creation of this note. Sue Ellen Wright's contribution to this version is primarily editorial. Comparison to the latest version of ISO/IEC 11179-3 and some thoughts on a taxonomy of Resource References (Knowledge Representation Resources) will follow.



In general we may expect a more complicated situation than the simpler one where two schemas directly reference the same data category. Linguists often want to define own categories for different reasons, but still are committed to support interoperability. As indicated in the Figure 2, they can do so by inserting a relation such as "schema_element_X" is_subclass_of "datcat_Y". This relation is typically stored in a light weight ontology, which we would like to call a "Relation Registry".

Again, if we inform the search engine that it should make use of the lexical schemas as well as the Relation Registries, it could carry out search operations on both lexica instantiated by the schemas.

Figure 2: Federated lexicon searching via shared, schema-enabled metadata reference and Relation Registry

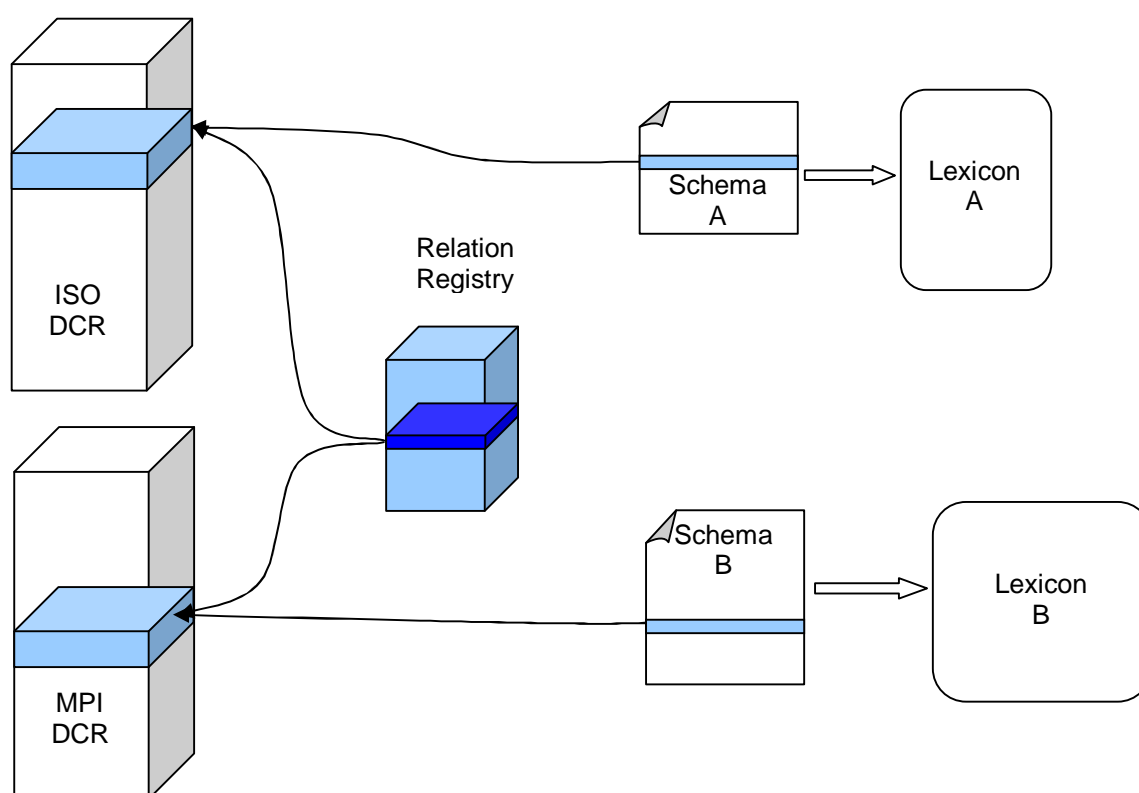


As the DCR matures and the Relation Registries proliferate, we even foresee another slightly more complicated scenario² to emerge step by step. Large institutions or projects want to create their own Data Category Registries for very pragmatic reasons. One of these could be that they have already achieved agreement about concepts as a result of a long discussion process. They should be able to create their own DCRs, however, it is expected that they will use the ISO TC37 data model for interoperability reasons³. In this scenario we would need an ontology that will relate the two data categories embedded in the two DCRs as is shown in the Figure 3.

² This scenario is basically a variation of the second model. The only difference is that this model involves a more standardized reference, i.e., the model holds for any schema that references a specific data category from an organizational DCR as opposed to the ad hoc reference in the model in Figure 2, which only holds for a specific data category for a specific schema

³ There is an enormous opportunity to implement this approach in the near term because the DCR data model will be published very soon, at which point the ISOcat software will be made freely available to everyone. Yet the ISOcat team knows of only a few current initiatives that are dealing with these aspects. We have just learned, for example, the GOLD initiative would be willing to include their definitions in a DCR. The official ISO DCR has room for private definitions, but we expect a great interest from large institutions to remain independent and to create their own form of internal decision processes.

Figure 3: Mediating schema-driven search strategies across multiple DCRs via a Relation Registry



Different mixed scenarios can be envisioned and are bound to emerge pragmatically in the future. In the examples shown here, we have only used lexicon schemas as resources that could utilize data categories. The same functionality, however, holds for all schemas that define structured language resources such as metadata, annotations, knowledge spaces, etc.

Since schemas will mostly be created by state-of-the-art tools, we have to convince tool builders to include an API for DCRs so that DCR services can be offered to the user during schema definition.. These services might take two possible forms:

- 1) When a new lexicon attribute or annotation tier is specified, the tool should offer the DCR datcats (both from the ISO TC 37 DCR and from organization specific DCR(s)) as the favored option
- 2) Again, a new entity is specified as described above, but the user wants to add a new datcat for whatever reasons; in this case, the tool should motivate the user to add it to a private DCR space and to relate it to an ISO Datcat where possible.

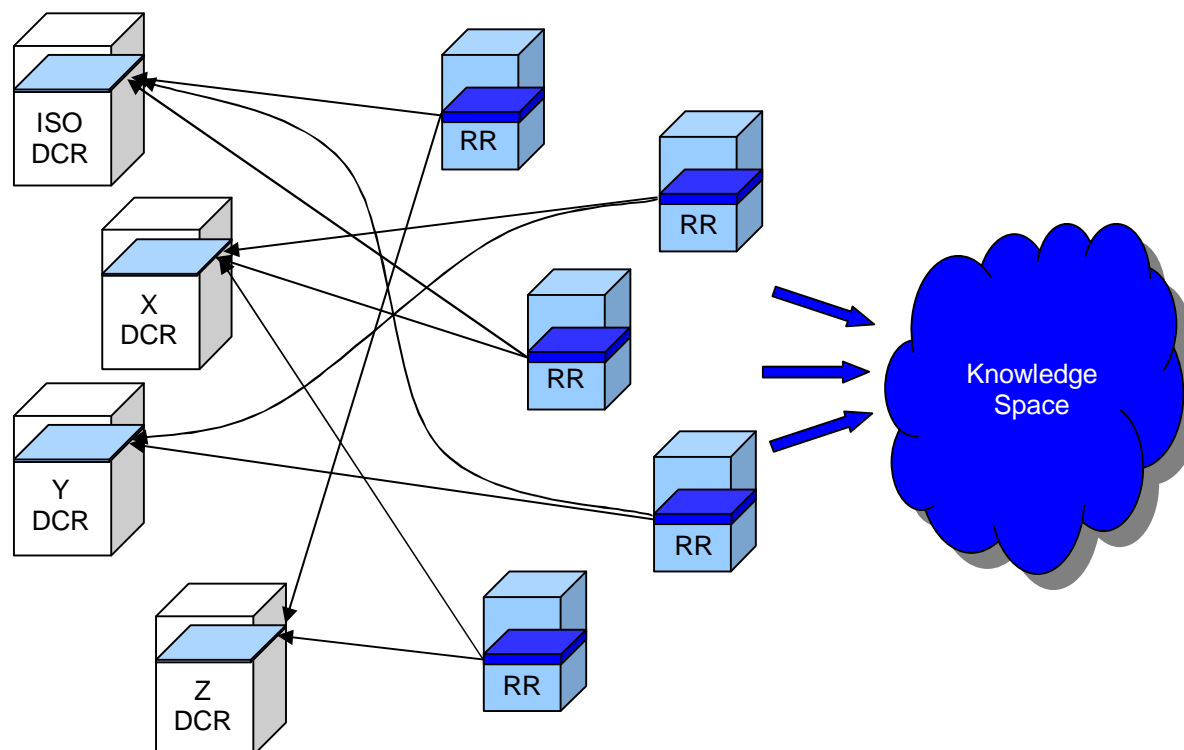
Even if a user does not want to take the trouble to define a relation to an ISO Datcat, linking data categories from user defined schema will become much easier. Currently, we know of two tools that use this kind of API: GATE (from U Sheffield) and LEXUS (from MPI Nijmegen). Many more tools need to follow these examples.

Given the above scenarios, the ISO DCR has a very good chance to be widely accepted as THE reference concept registry. The criteria to achieve such a status are very obvious:

- The underlying model (i.e., the DCR core metamodel) must remain comparatively simple;
- The concept definitions in the DCR must be excellent and widely acceptable;
- The DCR infrastructure must be simple to extend and must provide enough flexibility for groups of or individual researchers to create their own concept domains;
- Consensus needs to be reached with respect to ontology mechanisms to relate concepts that are used in different DCRs and resource schemas;
- More tools need to support the ISOcat API;
- We need an API for accessing the RR.

Relation Extensions and Knowledge Spaces

As described, we foresee the need for ontologies that establish typed relations on top of the data categories in DCRs and the elements used in schemas. The basis for this configuration is the current model underlying the DCRs, since it will not be changed. Relations are seen as annotations on top of the definitions in the DCRs, whereby the stand-off principles hold. Making this choice of separating definitions and relations allows users to generate several sets of relations that may even include conflicting knowledge. These first layer ontologies could also be called "relation registries", since they store simple relation triples such that they can easily be used by search engines and easily manipulated by simple editors.



These simple Relation Registries need to be distinguished from true ontologies, which form Knowledge Spaces. Knowledge Spaces lend themselves for inferencing, i.e., they need to include definitions eventually extracted from DCRs and they will include relations, properties, etc. to form logically complete systems.

While Relation Registries could be stored in simply structured XML files, Knowledge Spaces need to be represented in knowledge representation schemas such as RDF, SKOS etc. Of course, it will be necessary to ensure that relation types to be used in Relation Registries are compliant with types found in RDF-S, OWL, etc. Smart tools will need to be created to extract definitions and relations to form Knowledge Spaces.

Resource Constraints

As indicated, we believe that only a simple DCR model will lead to the success that is necessary for establishing an interoperable domain of language resources and technology. We know that when using a data category in a special context its usage needs to be constrained beyond the generic constraints defined for the global data category defined in the DCR. Since there are so many possible usage scenarios, it is not feasible to include supplemental constraints within the DCR. Again, they need to be treated in the context of their usage.

Data Categories are used in schemas to define, for example, classes of lexica or annotations. Schemas or even component schemas as specified for example for LMF are excellent places to add local and long range constraints. Local constraints can influence, for example, the value range, i.e. the values taken over from the DCR can be restricted. Furthermore wide range constraints can be implemented if the data model is designed to implement distributed constraints. For instance, LMF allows the association of operations with a certain attribute, which could be used to influence the values of other cells in a given lexicon.

Introducing these options will make it possible to fully implement, for instance, the PAROLE/SIMPLE class of lexica by defining appropriate component schemas and including them in lexicon instances that should have the same characteristics.

Infrastructure

To summarize, we foresee the following rich landscape related with Data Category Registries:

- There will be a number of equivalent mirror sites for the ISOcat DCR services, including both standardized ISO data category definitions as well as private definitions;
- There will be a number of institutional, project-oriented and perhaps even individual DCRs, all based on the ISO model (by making use of the ISOcat software⁴);
- There will be increasingly more schemas and component schemas that will reference the ISOcat DCR and other DCRs; these schemas will include constraints of different types on the use of data categories taken from a DCR;
- There will be increasingly more light weight ontologies (Relation Registries) that will include relations between concept or schema entries, i.e., these RRs will contain two references and a relation-type, i.e. they will be proper RDF triples independent of whether they are stored according to a simple XML-schema or as RDF files;
- There will be Knowledge Spaces (full-blown ontologies) that include amongst other elements definitions from DCRs and relations from RRs.

It is obvious that we need a suitable infrastructure to allow users to navigate in this landscape and to carry out manipulations that are necessary for their work. The following briefly describes some cornerstones of such an infrastructure:

- A registry is needed that allows users to register all resource types such as ISOcat DCR mirrors, other DCRs that are created by other initiatives, schemas and component schema registries that are used to specify the structure of linguistic resources, Relation Registries that are shared with other users, and various KS. This Resource Registry must be based on a taxonomy of language resources, allowing us to derive a proper hierarchy of nodes with which individual instances can be associated.⁵
- We need editors and registration mechanisms to create and extend the Resource Registry, and we need browsers and search engines that support resource discovery⁶.

⁴ The ISOcat software needs to be open source and freely available in order to achieve an open and rich landscape with a strong potential for interoperability. Indeed: the most important goal of the whole endeavor is interoperability – all other aspects need to be secondary.

⁵ SEW: How broad a selection of items do you want to include in such a resource registry? I've got the beginnings of an exhaustive taxonomy of knowledge representation resources, although I need to add our notion of schemas and reference registries to it.

⁶ The former notion of "resource metadata" is one aspect that will be incorporated in "resource registries". Existing metadata catalogues for resources and tools need to be integrated into such a new registry setup.

- We need a flexible and easy to use DCR editor that allows users to create new data category entries, to manipulate existing ones and to access all definitions⁷. In particular, for the ISO DCR we need to support a decision process as outlined in Annex ST of the ISO Directives for Standards as Databases, while at the same time retaining total freedom of access and freedom to create non-standard data categories and data category selections in private spaces. It must also be easily possible for initiatives to setup their own DCRs using ISOcat software.
- We need simple ontology editors that allow users to create or manipulate Relation Registries, i.e. this editor must allow users to select data categories from registered DCRs, relate them using accepted relation-types, store the results and make this information shareable. Special focus is needed for legacy data, whose schemas may contain many elements that are not included in any DCR.
- It must be possible to create RDF versions that include selected datcats and relations so that they can be integrated in Knowledge Spaces.

It should be mentioned that the CLARIN project is devoted to establishing such an infrastructure during the coming 3 years. Since the creation of such an infrastructure is an international activity, we would like to invite all colleagues to participate in this endeavor. CLARIN will work out the requirements, discuss them with all interested colleagues and offer them to the ISO standardization process where it seems to be feasible.

ISO TC37/SC4

ISO TC37/SC4, in cooperation with the rest of ISO TC37, is seen as the most suitable umbrella to launch widely accepted standards that have the potential of establishing a domain of highly interoperable language resources. We foresee the need to establish standards for

- Relation Registries that are orthogonal to the DCRs and that can be seen as light weight ontologies;
- Relation-types that are accepted to be used within RR and that are compliant with frameworks such as RDF-S and OWL;
- A resource registry mechanism that is based on an accepted taxonomy of language resources in the general sense.

The newly established CLARIN research infrastructure is devoted to working out these aspects within the coming year and to make proposals to ISO.

Note: The scenario described here concerns the definition and full specification of data categories (data element concepts) in a metadata data registry, but research groups modeling terminological data are also designing systems that separate the core definition and documentation of terminological concepts from associated relational references and concept schemes. We should maintain close contact between the two communities of practice because it is highly likely that the same binding strategies can be used in order to facilitate expanded interoperability.

⁷ The ISOcat software currently under development will fulfill this role.