CLARIN Requirements for a Semantic Registry

Menzo Windhouwer – Knowledge Systems – The Language Archive – DANS

Introduction

Resources in CLARIN have a very diverse nature, i.e., range from (semi)structured data, glossed text aligned with multimedia timelines to large (free) text corpora. Even common data formats, e.g., EAF or LMF, allow the user freedom in the use of glosses, tier names and container or field names. This is also true for CMDI metadata, which might have any structure and carry any semantics.

Still researchers should be able to find (possible) relevant data even when their research question is phrased using different terminology and based on a different data structure. A first step towards interoperability is to semantically annotate the resources and make the used semantics explicit.

The CLARIN infrastructure advocates this approach, especially in the metadata domain where the use of concept links is an integral part of CMDI to deal with the diversity of data structures and semantics. In the CLARIN-NL national initiative projects have also been obliged to semantically annotate the contents of the resources used/created. But where CMDI already uses the concept links in the (facetted) search facilities the use of the semantic annotations on the resource level is in its infancy. Still the aim is that higher levels CLARIN federated search, i.e., beyond level 0 that supports full text search, could make use of this information.

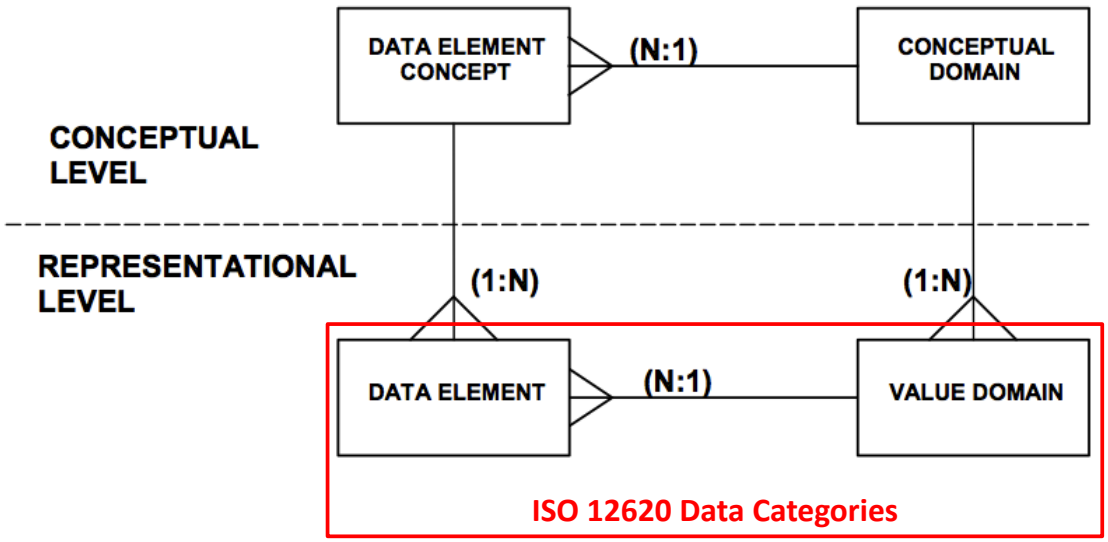
In CLARIN, and especially CMDI, the ISOcat DCR developed and hosted by TLA has been adopted as the Concept Registry. This DCR implementation is the result of a 10 year history of DCR pilots and requirements specifications lead by ISO TC37 and there is even longer history of using data categories in the terminology community of practice. TLA took on development of ISOcat as the TC37 DCR in 2006, when CLARIN was just starting up. Due to its ISO embedding this implementation has been mostly steered by requirements set by TC37 in the development of ISO 12620:2009. However, CLARIN has been using the ISOcat DCR now extensively for some years and its own requirements have surfaced.

The purpose of this requirements analysis is to have a look at these experiences and see what CLARIN actually needs of a Semantic Registry to support semantic annotation for metadata, i.e., CMD records, and resources. The analysis looks at both the data model of the registry and the (community) processes around it, which should ideally discourage proliferation but still be agile enough to timely adapt to the ongoing changes in the semantics of the research domain, e.g., the rise and fall of theories.

DCR data model

The DCR data model of ISO 12620:2009 is based on the ISO 11179 family of standards for Metadata Registries. This family of standards describes, among many other aspects, the specification of data elements in a registry, i.e., to foster

reuse by compliant metadata schemas. Metadata elements are depicted as follows:

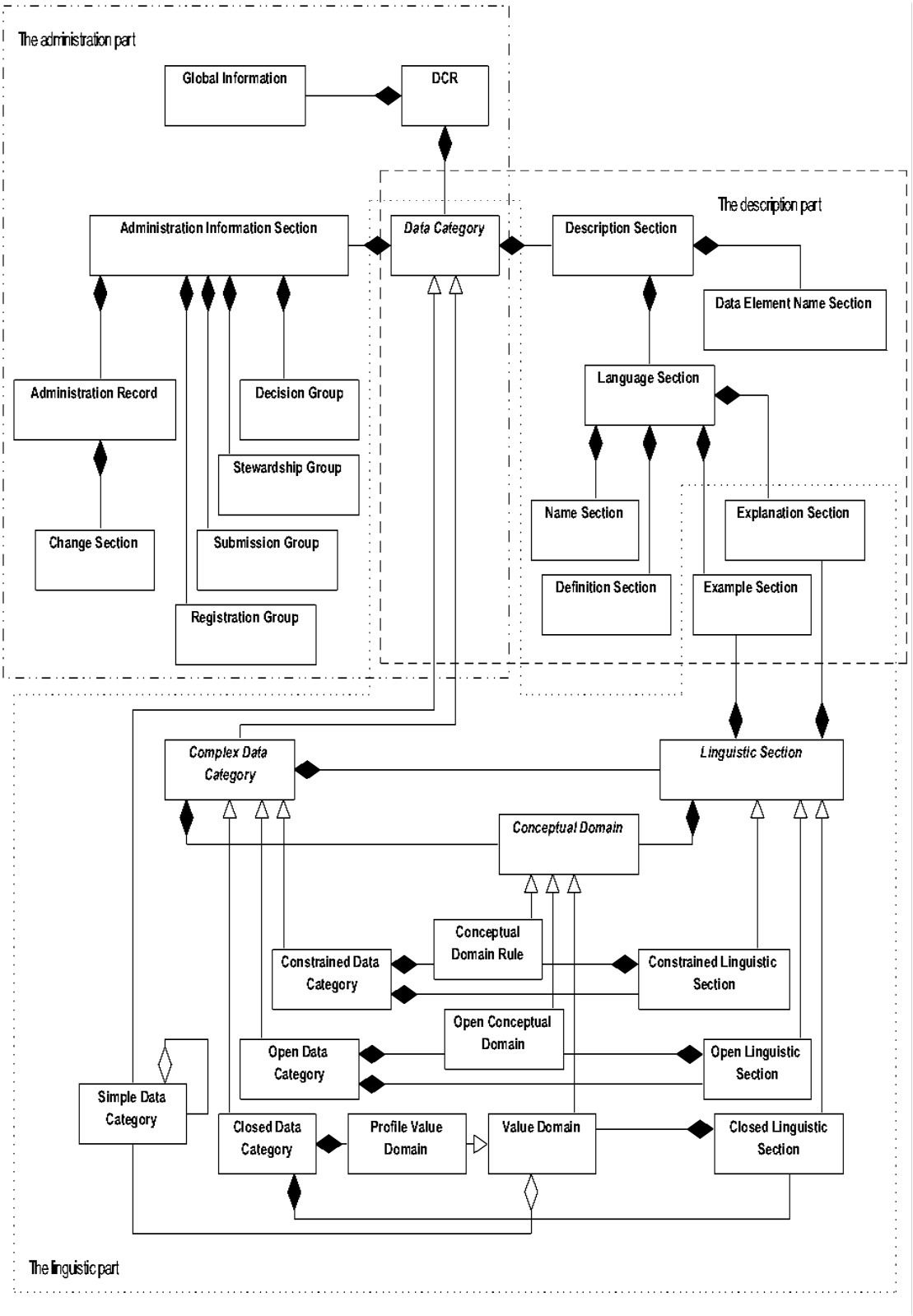


There are two levels: conceptual and representational. While the conceptual level semantically defines the data element the representation level specifies how they are represented, e.g., a string or a date. This model has been adapted a bit by ISO 12620:2009: data elements are called data categories, and values in a value domain are also described as data categories.

Differences between data elements and their values lead to the 2 basic classes of data categories: 1) complex data categories (corresponding to the data elements) and 2) simple data categories (corresponding to the values). Every kind of value can appear in an instance, but they cannot always be enumerated in the semantic registry. The DCR data model caters for 3 different kinds of value domain (a.k.a. the conceptual domain) specifications: 1.1) closed data categories (value domain is enumerated as simple data categories, which should fit also a data type[1]), 1.2) open data categories (value domain is not enumerated, but constrained by a data type), and 1.3) constrained data category (value domain is not enumerated, but constrained by a data type plus additional rules, e.g., a date in the 21st century). Later on this model was extended with 3) container data categories, which can be used to semantically described grouping constructs, e.g., tables, classes, inner nodes, of a data structure. In the ISO TC 37 context these containers, e.g., LMF UML classes, are usually described in a standard, e.g., ISO 24613:2008 for LMF, but in CLARIN a wider variety of models exists and also the grouping levels of these need to be described.

General information and type specific information can be stored, which ultimately leads to the following (high level overview of the) DCR data model:

[1] The DCR model and also the ISOcat do not enforce this in any way, i.e., a closed data category can have a data type integer, while all the simple data categories in its value domain have string



This is a very elaborate model which allows rich specification of the semantics and possible use of data categories. However, the CLARIN experience has revealed various problems:

Proliferation due to types: although a concept is the same if the representation differs another data category has to be created (e.g., in *part-of-speech = "verb" /verb/* is a simple data category, while in *verb = "to walk" /verb/*

is an open data category), i.e., the model does not cater for sharing the same concept across data categories with different representations, which leads to proliferation of the registry.

Conflicts with actual use: data categories are used in a resource context, i.e., in a schema or in an instance, and in this context the representation information is already available, e.g., an XML schema tells if an element can have a value or not. The DCR has no way to enforce a proper match between the representation in the resource and the specification of the representation of the associated data category in the registry. Thus there is a potential for conflicts, which gets higher due to the next problem encountered within CLARIN.

A rare blend of expertise: although there are persons that combine linguistic and technical expertise they are rare, but for a proper specification of a data category exactly this combination of expertise is needed. This turns out to be problematic leading to data categories of types that conflict with the actual representation information in the resources (e.g., simple data categories associated with CMDI components).

Disambiguation of terms used in definitions: it is the aim of data category specifications to be rather context insensitive, i.e., so it can be reused within a different context. However, some data categories are defined within a specific theoretical context and although data categories might share the same name their definition might not share the same theoretical background. Mixing data categories can thus potentially lead to semantic inconsistencies. Within CLARIN it is thus advised to link important concepts in the definition explicitly to the right data categories. It can be hard to determine the right data category type for these important concepts, as they might not occur explicitly in the data model.

These problems together with the insight that the representation information is actually available in the resources leads to the insight that dropping the typing of data categories[2] would make the model already simpler, and would require mainly domain expertise and less technical insight of the data category creators.

Looking at the DCR data model this removes basically the linguistic part. That includes the linguistic section, which is tightly bound to the typing of data categories. However, some of the descriptive information of these sections, e.g., examples and explanations of language specific use of a data category, could be maintained.

This culls all explicit data category relationships based on data category types from the current DCR. Model. In the original design during the editing of ISO 12620:2009 it was already decided to prune away ontological relationships (although the possibility for is-a relations partially survived) as they are too domain/application/user specific. The same is actually true for value domain relationships. CLARINs experience has been that many times owners of closed data categories had to be requested to extend their value domains. This could not always be accommodated and the proposed solution has been to add functionality to allow users to clone data categories owned by other users and then to adapt the value domain of the clone. Out of fear of further proliferation of the DCR this feature has been stalled.

---

[2] This means that the data category specification doesn't contain representation information anymore, and following the ISO 11179 model this means that actually the registry would store data element/category concept specifications. For the sake of continuity the term data category will keep on being used in the remainder of this document.

On the other hand CLARIN experience shows that full context insensitive definitions of data categories are also not possible. So it should become possible to link to other data categories from a definition to disambiguate certain concepts.

Although no typed relations between data categories would exist in the DCR anymore they are still interesting information, which can be exploited for resource discovery or otherwise. With the removal of ontological relationships the idea of a Relation Registry has also been floating around for some years[3]. This Relation Registry would allow to store sets of project/community/user specific relations, i.e., not one specific ontological perspective. As the relationship types of the Relation Registry is not a closed set additional types can be added for value domain relationships.

The DCR data model also contains classes tightly bound to the standardization process. These can be pruned away from the core model and be left to the DCR implementation. This makes it better possible to design a data model in tandem with a workflow that is more community driven.

The Data Element Name Section is an useful, although also a bit confusing class in the DCR data model. This is a place intended to store the more technical 'names' for a data category. In practice this can be URLs (e.g., in the context of RDF), but also XPaths (e.g., possibly in the context of CMDI). In the OpenPHACTS ConceptWiki[4] approach this kind of information is stored in so called *Also Referred To As* (ARTA) tables. A similar approach could resolve some of the confusion in the DCR data model and make the class more generic and usable.

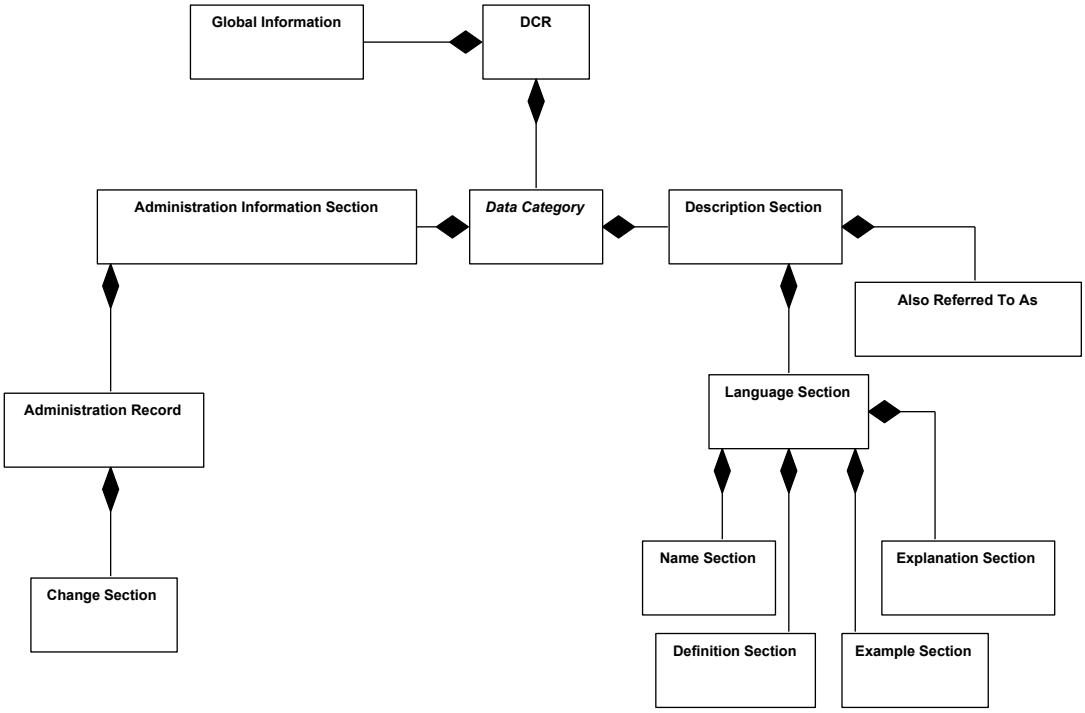
Some other problems with the current data model should also be addressed:

- The legacy identifier is confusing for users as it isn't globally unique in the registry, a solution could be the require a DCR specific DEN/ARTA enytry;
- Source has different meanings at different places, e.g., in the DEN or Description Section, a solution would be to rename the DEN source to origin;
- A Language Section should only contain one Definition Section, if multiple definitions are considered they should be discussed in the community, e.g., in a forum or an email list, and result in one clear definition in the specification;
- A Language Section should have only one preferred Name Section;
- Superseded Name Section should refer to the successor, or only deprecation should be allowed.

This gives us the following, more lightweight, model:

[3] An alpha version of the RELcat Relation Registry (http://lux13.mpi.nl/relcat/) has been around for some while and has also seen some use within the CLARIN context. Development of an beta version with an user interface to manage relation sets has been delayed a bit to see the outcome of the workshop/meeting, as RELcat make take over some of the functionality that ISOcat might shed.

[4] http://www.conceptwiki.org/



Alternatives for this trimmed down DCR model can and should be considered. SKOS is used more and more in this area, and general software, e.g., HIVE and OpenSKOS, is coming available. However, they might not have all the functionality we need. HIVE[5] and OpenSKOS[6] do provide APIs to access specific concept schemes, but HIVE doesn't provide an editor while OpenSKOS does. For a more wiki like approach the use of Semantic MediaWiki[7] for SKOS concept schemes might be interesting. I could envision (trimmed down) data categories mapped to concepts and selections to (user) specific concept schemes. RELcat relation sets could then be used to add (user) specific relations to such a concept scheme, i.e., thus creating a derived concept scheme.

Process

Next to the data model there are currently issues with the process that hamper uptake from the DCR. The DCR has rather strict ownership of data category specifications, i.e., for each change the owner has to be contacted unless (s)he shared edit rights on the data category. What could be opened up to make it easier to extend and publish a data category specification?

Adding a new DEN/ARTA entry: it should be possible to add a new entry on how, in technical terms, this data category is addressed in a specific domain/application/…

Adding a new translation: next to the mandatory English language section other language sections are possible, and it should be possible to relatively easily add/edit them (still maybe controlled by users with a translator role?)

Adding new profile memberships: especially the relationships between typed data categories where intricately interwoven with profile membership, i.e.,

[5] https://www.nescent.org/sites/hive/Main_Page
[6] http://www.openskos.org/
[7] http://semantic-mediawiki.org/

to enforce that the thematic domains actually confirmed that these data category specifications were relevant and properly defined for their domain. With removal of the relationships the role of profile membership is less prominent and might be replaced by allowing simple tagging (where a tag could represent a thematic domain) of data categories.

Publishing of data categories by community coordinators: when a data category owners hands over a data category for review, e.g., to a CLARIN national content coordinator, (s)he indicates that (s)he considers the data category ready for general use and if the coordinator agrees it should be possible for the coordinator to then make the data category publically available.

These changes of edit and publishing rights could be taken to an extreme, i.e., take a wiki like approach[8] and allow anyone to edit a data category specification. Each version would have its PID, so one would always refer to a specific one which will stay stable[9]. However, depending on the granularity of change (e.g., major, minor, typo) the PIDs might change too often. This could be remedied with an explicit versioning policy, but this requires again explicit ownership, i.e., someone has to decide when to release or retract a version.

The CLARIN experience shows that users worry about the stability and like to claim ownership. To accommodate this wish the edit rights for the English language section and the versioning policy could be controlled by them. The drawback of this is that a data category specification might become hard to maintain when the owner loses interest. Forced (?) transfer of ownership might be possible or having clone capabilities. But the latter comes with the drawback of possible proliferation.

However, due to the open nature of the DCR proliferation is always possible, i.e., as any user can create a data category with the same name and (almost the) same semantics. In the ConceptWiki approach and also in the RDA Data Foundation and Terminology (DFT) working group some interesting clustering approaches are used/discussed. In the ConceptWiki all concepts with the same or nearly the same semantics are grouped (by whom?) in knowlets[10], which is thus is basically a same-as or almost same-as (or near sameness) group. In the RDA DFT a trial with StackExchange was proposed[11]. Normally a StackExchange page has one question and then ranked answers (users with enough credits can promote or demote an answer). Ignoring the question one could envision various alternative data categories which users (or real actual use) could promote or demote and thus make it clearer to the community which alternative is a good candidate to pick, while still allowing the selection of another one with (slightly) different semantics that suit ones needs better.

By having open registries we hope to cater for a grass roots approach where new data categories bubble up from the communities, and also to provide the

---

[8] The ConceptWiki use to have a Semantic MediaWiki setup, they have replaced that (see http://www.conceptwiki.org/) but haven't enable the community features to allow anyone to change the concepts.

[9] There shouldn't be a PID (in my opinion not even an URL as people could misuse it) that always points to the latest version of a data category as that would cause drifting semantics, i.e., one should always point to a version one has inspected and deemed relevant for the resource.

[10] Although conceptually nice it is not (yet) clear to me how they are constructed and how to detect when a concept should fork off a new knowlet when its semantics have drifted too far from the original knowlet.

[11] See http://meta.opendata.stackexchange.com/questions/149/what-is-data-citation

agility to adapt to new trends, theories and research coming up. Still there is a role for authorities here, as many users like to be able to select data categories for core parts of their domain. Sometimes there are many proposals and they like authorities to give them a recommendation. These recommendations can come from various levels, e.g., a project (GOLD), a infrastructure (CLARIN), a community (Athens Core), a standardization organization (ISO TC37). Next to recommending data categories created in the registry it would also be good to represent commonly used vocabularies, e.g., the Dublin Core elements and terms, the GOLD ontology. We did do this partially already, but it could be intertwined with the recommendation functionality, i.e., import Dublin Core, keep the PURLs, and state they are recommended by DCMI. In the ConceptWiki approach they do this, i.e., import well established ontologies and taxonomies from their domain.

In a recent EUDAT workshop it was discussed that it would be nice to have a light weight semantic registry that would allow users to register terms and their definitions quickly and get a PID/URL to use in their resources. This uncurated registry could be used by authorities to see what concepts are bubbling up from the community and could enter the process to be taken up into the carefully curated knowledge bases, e.g., BioPortal. The original URI in the light weight registry would stay working but in due time will be connected to the equivalent concept in the authoritive knowledge base. This is a bit more decoupling than we did in ISOcat, but it might actually create a clearer split for the user: informal semantics in the lightweight (cross domain) semantic registry versus formal semantics in the authoritive (domain specific) knowledge bases. An API would allow users to search both, but curated concepts should be higher ranked matches. If no match is found or deemed relevant by the user there should be a simple process to insert a new (minimal) entry into the light weight registry. This API can be used during creation of resources or metadata, e.g., in ELAN, LEXUS, CMDI Component Registry, but also during semantic annotation, e.g., in DWAN, EUDAT annotator.

User Interface

The current ISOcat user interface uses a desktop-like framework within the browser. This framework is getting outdated and should be replaced by a more modern approach. If a switch is made to an existing framework for a semantic registry, e.g., OpenSKOS, ConceptWiki or MediaWiki, the selected solution will most likely also determine the user interface framework. Notice that it might also determine the (easy) availability of current ISOcat functionality, like, selections, groups and views. In any case a more wiki-like approach (meaning here the more textual and page orientated mode of input and not specifically the openness of and processes around a wiki) might be preferred.

Other

How to assist an user in finding the right data category/concept remains an open issue. Easier extension of the DEN/ARTA entries might help to make an entry easier findable for specific viewpoints. Also integration with a more populated Relation Registry might help, i.e., it can provide an taxonomy for drilling down or a same-as clique to look in a neighborhood.