Just for the record, CMDI should be about semantic interoperability

Thorsten Trippel and Claus Zinn
Seminar für Sprachwissenschaft
Universität Tübingen
firstName.lastName@uni-tuebingen.de

Abstract

The Component MetaData Infrastructure (CMDI) provides a lego-brick framework for the creation, use and re-use of self-defined metadata formats. The design of CMDI can be a force for good, but history shows that it has often been misunderstood or badly executed. Consequently, it has led the community towards the dark ages of metadata clutter rather than the bright side of semantic interoperability. In this abstract, we report on the condition of CMDI but also outline an agenda to make the CMDI world a better place to use, share and profit from metadata.

1 Introduction

With a broad range of language-related resources, there cannot be a single metadata schema to properly address all the needs of the heterogeneous community of Humanities and Social Sciences researchers. So says the elders, and so goes the folklore [4]. But rather than having a multitude of different metadata schemas to cater for the various needs of different sub-communities, and the varying nature of their research data, there shall be single, extensible metadata framework that provides a common syntactic and semantic basis. And so the Component MetaData Infrastructure was born. CMDI (ISO 24622-1) is a meta-model that allows users to define metadata schemas to fit their needs. Using a modular approach, metadata modelers can define basic building blocks (“terms”, “data categories”, “concepts”), and with them, they can assemble more complex components, which in turn can be combined to metadata schemas.

CMDI aims at promoting the re-use of existing concepts and components, and hence, fostering the semantic interoperability among metadata formats. For this, it provides two registries: the CLARIN concept registry (CCR) for the definition of basic metadata descriptors [URL-1], and the Component Registry for the definition of components and schemas [URL-2]. Both registries can be accessed via standard web browsers to query existing entries, or to add new ones.

The CMDI communities have not always used the CMDI framework wisely, though. With more than 3000 entries in the CCR, more than 1000 public components, and more than 180 public profiles, it is clear that the aim for reusing metadata descriptors has not been achieved, and hence, that semantic interoperability has fallen behind. In this paper, we give an account of the issues the CMDI community faces, and then outline a healthy set of good practice guidelines to improve CMDI’s semantic interoperability across the metadata universe.

2 CMDI Issues

It is clear that CMDI delivers on the grounds of syntactic interoperability. Based on XML, a CMDI instance documenting a resource is linked to a CMDI metadata schema, and XML validation can be used to check whether the instance adheres to the schema. The main issue to address is the interpretation of the resulting syntactical structure. We highlight four CMDI use cases.

2.1 CMDI for Searching Across Catalogues

The Virtual Language Observatory (VLO) offers a metadata-based search on over 900,000 language-related resources described with CMDI metadata [URL-3]. To explore the aggregated dataset, VLO users
can currently use twelve facets and a full-text search on the metadata.

The VLO metadata is harvested from 32 centres [URL-4]. While the centres have their metadata ready in a dozen of different formats, they mostly provide their metadata in a CMDI-based format, making use of many different CMDI-based schemas. Given the large number of different CMDI schemas in use, it is not feasible to map each schema separately to the facet space. For indexing, the VLO developers have devised a mapping between elementary data categories (DatCats) and the facets [URL-5]. Fig. 1 shows the mapping of some of the data categories to VLO’s language facet. The first eight entries orig-

<table>
<thead>
<tr>
<th>CMD Element ConceptLink</th>
<th>CMD Element ConceptLink</th>
<th>CMD Element ConceptLink</th>
<th>CMD Element ConceptLink</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.isocat.org/_datcat/DC-2462">http://www.isocat.org/_datcat/DC-2462</a></td>
<td><a href="http://hdl.handle.net/11459/CCR_C_2462_08edded2-4086-7a3f-88e6-c0807f01e17">http://hdl.handle.net/11459/CCR_C_2462_08edded2-4086-7a3f-88e6-c0807f01e17</a></td>
<td><a href="http://www.isocat.org/_datcat/DC-2484">http://www.isocat.org/_datcat/DC-2484</a></td>
<td><a href="http://hdl.handle.net/11459/CCR_C_2484_669644e7-cf9e-a9f6-59cb-a25fe89b9b9d">http://hdl.handle.net/11459/CCR_C_2484_669644e7-cf9e-a9f6-59cb-a25fe89b9b9d</a></td>
</tr>
<tr>
<td><a href="http://publ.org/dcr/terms/language">http://publ.org/dcr/terms/language</a></td>
<td>An empty context leads to acceptance.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any context, which is not rejected, leads to acceptance.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rejection context

| Rejected CMD Container ConceptLink | http://www.isocat.org/_datcat/DC-4146 |
| Rejected CMD Container ConceptLink | http://hdl.handle.net/11459/CCR_C_4146_5ccc4c5c-e729-c180-2bf1-fcc56ddc24d |

Fallback xpath: /c:CMD/c:Components/c:OLAC-DcmTerms/c:language/@olac-language
Fallback xpath: /c:CMD/c:Components/c:OLAC-DcmTerms/c:subject/@olac-language
Fallback xpath: /c:CMD/c:Components/c:OLAC-DcmTerms/reit/language/@olac-language
Fallback xpath: /c:CMD/c:Components/c:OLAC-DcmTerms-ref-DWR/c:language/@olac-language
Blacklist xpath: /c:Actor_Languages/c:Actor_Language/
Blacklist xpath: /c:LanguageName
Blacklist xpath: /c:CMD/c:Components/c:DSAGD_Corpus/c:Collection/c:DocumentationLanguages
Blacklist xpath: /c:CMD/c:Components/c:TextCorpusProfile/c:DocumentationLanguages
Blacklist xpath: /c:CMD/c:Components/c:TextCorpusProfile/c:DocumentationLanguages
Blacklist xpath: /c:CMD/c:Components/c:ToolProfile/c:ToolContext/c:DocumentationLanguages
Blacklist xpath: /c:CMD/c:Components/c:LexicalResourceProfile/c:DocumentationLanguages
Blacklist xpath: /c:CMD/c:Components/c:MultimediaCorpus/c:DocumentationLanguages

Figure 1: The complexity of mapping concept terms to a VLO facet.

iniate from the data categories /language ID/, /language name/, /language usage/, and /language/ originally defined in the ISOCat registry isocat.org, and which were transferred to the CCR. The ninth entry checks for the presence of dc:language. If the given DatCats are not used, a number of fall-back XPath expressions are being tried to find language information in specific subtrees. To exclude certain subtrees from contributing to the language facet, contextual information is used, here, to not confound the documentation language of the resource with its language.

Semantic interoperability is partially achieved by making use of schema elements being grounded in the CCR. This task is complex, in part because of the CCR having often multiple entries for the same concept (e.g., /language/), also see [5], and because certain contexts need to be excluded for concepts that have a quite general meaning. The mapping between DatCats and facets also has to be adjusted whenever a CMDI instance is found using a profile that uses new DatCats or embeds them in unseen contexts. It is unclear how much information is being missed by the mapping being potentially incomplete, erroneous or outdated.

### 2.2 CMDI for OAI-PMH providers

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a widely-used protocol to distribute metadata records. According to the specification, data providers must at least offer their data using Dublin Core (DC). Around 25 CLARIN centres make available their data in CMDI, but overall, 1

---

1. The facets are: Language, Collection, Resource type, Modality, Format, Keyword, Genre, Subject, Country, Organisation, Data provider, and Availability.
over twenty different formats can be harvested, including OLAC (15 times), and the bibliographic formats MODS (7 times) and METS (5 times).

The conversion between CMDI and DC is done by the data providers who supposedly know their CMDI-based schemata better than any centralized service such as the VLO. It seems more feasible for data providers to define detailed mappings between their native CMDI schema and DC, and hence to deliver high quality DC. While most CLARIN partners deliver DC, the richness of their DC export is, however, sometimes disappointing. Some exports only feature dc:title and dc:description tags, or only generate the dc:identifier descriptor; some even distribute syntactically incorrect DC. On the CLARIN website https://www.clarin.eu/content/oai-pmh-cmdi, we find the quote: “Or, in human words, how should I map my CMDI descriptions to the Dublin core format that is compulsory when using OAI-PMH? The answer is: you probably know this the best, there is no single answer to this. It’s probably a good idea to use common sense and to refer (minimally to the described resources with a DC:identifier element)”. This implies a problem in semantic interoperability.

2.3 CMDI for Library Cataloguing

Libraries and archives hardly know about CMDI. However, it may be in their interest to ingest CMDI-based metadata into library catalogues so that a library catalogue search not only yields the traditional publications for scientists, but also the research data they have created [2], and this may be appropriate for enhanced publications. For this, CMDI-based metadata must be converted to bibliographic standards such as DC and MARC 21, yielding metadata that adheres to the high level quality standard of libraries.

In [2], Kaminski et al. have identified a number of issues that need to be addressed to convert CMDI-based metadata to DC and MARC21. A simple “concept-to-facet” mapping was shown too weak to achieve those high standards. Consider the CMDI component /Person/(clarin.eu:cr1:c_1447674760335), see Fig. 2. It consists of three data descriptors, /firstName/, /lastName/ and /Role/. While the first two are anchored to the CCR, the last is not. In a generic fashion, it seems possible to build a full name using the values for the data elements /firstName/ and /lastName/, but without a semantic anchor for /Role/ it is impossible to assign the person’s full name to, say, the value of dc:creator or dc:contributor, or dc:rightsHolder.

Of course, knowing a specific schema in advance, it is always possible to perform schema-specific mappings, and check for whether /Person/ occurs inside /Creation/, /Project/ or /License/, but why not making use of the respective DatCats in the CMDI profile in the first place?

2.4 CMDI for Linked Data

The Semantic Web is built from structured data of unique resource identifiers (URIs) that are highly interlinked. It takes searching across (library or research data) catalogues to the next level (searching across all published data sets). Fig. 3 shows metadata for a scientific publication in Linked Data provided by the German National Library. Note that the publication has a URI, and the property dcterms:
creator has as value a URI, too. The resource http://d-nb.info/gnd/173732410 is given a name, and is also linked to another authority record: http://viaf.org/viaf/210410286. And somewhere in the Linked Data cloud, we may find more information about the person referred to by this VIAF identifier. – For CMDI-based metadata to take part in Linked Data, it is necessary to map the metadata vocabulary to existing Linked Data vocabulary, and it must map the value space of CMDI metadata to existing Linked Data entities.

3 Good Practise Guidelines

From the perspective of semantic interoperability we suggest a reevaluation of priorities in the design of CMDI profiles. The following guidelines aim at maximising semantic interoperability of CMDI-based metadata with the metadata universe “out there”.

Use existing, established terms whenever possible. For the CCR coordinators, (i) consider rejecting all concept definitions that duplicate existing, established terms, (ii) consider deprecating all terms that have a semantically equivalent term in an established metadata standard.

Only describe terms that are needed and have not been defined yet. New terms should only be defined when there is no semantic equivalent metadata term in any of the established bibliographic metadata standards (DC, MARC21, METS, MODS, MADS).

Use controlled vocabulary when assigning values to metadata fields. In particular, use authority identifiers such as VIAF (see viaf.org, also [5]) for identifying persons, ISO 639-3 URIs for languages, ISO-3166-based URIs for countries, IANA-based URIs for mimetypes, and make use of the LoC subject classification for subject and genre fields.

When new terms need to be defined, be specific rather than general. Use /actorLanguage/ and /documentationLanguage/ rather than /language/ to minimize contextual interpretation.

Use CMDI to provide the best possible terms for the description of language-related material. There is ample potential to describe treebanks, linguistically motivated experiments, richly annotated text corpora, and the many other resource types to a good level of detail. For this, new vocabulary is needed, and CMDI is the best suited framework for this. For the “shallow” parts of describing resources, use existing bibliographic terms from established metadata schemes.

Arguably, at least the first three recommendations were known to most people that had to work with the two registries. So why did the CCR grow to a size of a few thousands entries, among of which are so many duplicates? Here, the design of the ISOcat registry, where most DatCats of the CCR originated from, might be to blame. In the old registry, a data category could take one of four types: complex/closed, complex/open, complex/constrained, and simple. Some users may have found these notions too confus-
ing, and defined their own DatCats rather than re-using existing ones of presumably wrong type. The new CLARIN concept registry simplified the notion of data category.

To cope with the 3000+ entries, the CCR complements its full-text search with a faceted based search. But while the new CCR offers facet filters for Status (“Approved”, “Candidate”, “Expired”), Concept Schemes, Collections, and others, the facets’ sparse value space does not help users well enough to identify DatCats relevant for their task. As a result, users still have to scan the search result in a linear rather than focused fashion. Here, curation work is necessary to better assign DatCats to their appropriate facets. In particular, a larger number of DatCats must get approved more quickly, and at the same time, their semantically equivalent or semantically similar counterparts must get deprecated simultaneously.

The CLARIN Component Registry could also be improved. Components might also profit from a status tag (other than “Public” or “Private”) to promote their re-use. Also usage data could be made available to the community to signal to users those CMDI components that are already widely used.

To increase user awareness, a CLARIN newsletter could offer a “data category or component of the month” column to promote good practise, or to communicate a curation effort. In this sense, the grass-roots movement to metadata vocabulary is moving towards a more centralized control, where metadata experts, supported by tools (say, for gathering usage data) provide a feedback loop to the community.

4 Discussion

The CMDI metadata framework has the potential to unite the various communities under a common metadata umbrella. While CMDI delivers on syntactic interoperability, there is much to be done to fully achieve semantic interoperability. While the interoperability issue could be achieved within the CLARIN world (cleaning up the metadata clutter in the CLARIN registries), it seems beneficial to look beyond the CLARIN horizon and to exploit the metadata approaches of other communities. The Library Sciences have an established vocabulary on offer to describe electronic resources to a good level of detail. The CLARIN community should use this vocabulary whenever possible. For a deeper, more linguistic-driven description of resources, new CMDI-based vocabulary should be properly defined and semantically anchored in the CCR. New components can be defined to describe all the deeper aspects of the various resource types, such as treebanks, highly annotated corpora, or experimental set-ups. Here, CMDI’s design could take an unrivaled role in deep metadata descriptions. Having joint metadata vocabulary with the Library Sciences also gives the CLARIN community access to the library catalogueing and hence, the Linked Data world. It brings together the traditional publications of researchers with the research data they create, and with Linked Data, the power of semantic querying to gather information about related scientists, their publications, their research data, or their other properties and activities.

Acknowledgments. We would like to thank the anonymous referees for their comments.

References (All links were accessed on September 28, 2016.)


