TEI and CMD

Matej Ďurčo, Karlheinz Mörth
Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences, Vienna, Austria

As is well-known the world of digital humanities abounds in diverging systems to encode metadata. To achieve a higher degree of interoperability, CLARIN has developed the Component MetaData Infrastructure (CMDI), a system designed to allow flexible reuse of different metadata blueprints. Our presentation intends to report on efforts to integrate TEI conformant metadata into this new infrastructure.

CMDI

CMDI – CLARIN/Component Metadata Infrastructure, first conceived within the CLARIN initiative in 2008 – is a distributed system consisting of multiple interconnected applications designed to create and provide metadata for language resources in a coherent harmonized way. It is built on top of the Component Metadata Framework – a flexible meta-model for defining metadata schemas – and the Data Category Registry (DCR) – a community-based registry for linguistic concepts.

The CMD (Component Metadata Framework) allows to define so called profiles made up of reusable components – collections of metadata fields. The components can contain other components and can be reused in multiple profiles. These profiles are automatically translated into XML schemas, that are used to generate and validate actual CMD records, i.e. metadata instances.

Conceptually, CMD is grounded in the Data Category Registry (DCR) a central registry meant to enable the community to collectively maintain definitions for a set of relevant linguistic concepts. The resulting controlled vocabulary is the cornerstone for grounding the semantic interpretation within the CMD framework. The data model and the procedures of the DCR are defined by an ISO standard [ISO12620:2009], and is implemented in *ISOcat*[1].

The Component Metadata Framework (CMD) is built on top of the DCR and complements it. While the DCR defines the atomic concepts, CMD allows to define schemas made up of fields that refer via a PID to exactly one data category in the ISO DCR, thus indicating unambiguously how the content of the field in a metadata description should be interpreted" [Broeder et al. 2010]. This allows to easily infer equivalencies between metadata fields in different CMD-based schemas. While the primary registry used in CMD is the ISOcat DCR, other authoritative sources for data categories ("trusted registries") are accepted, especially those of the Dublin Core Metadata Initiative [Powell et al. 2005].

Currently, the joint metadata domain of CMDI contains more than half a million records in 60 different CMD profiles collected from 69 data providers.

[1] http://www.isocat.org

TEI

The TEI guidelines is a de-facto standard for encoding textual resources. One of the most widely used set of elements contained in these guidelines is the so called teiHeader, a complex element which allows encoders to describe digital resources, i.e. to store metadata.

The widespread use of TEI for encoding textual resources has created a strong need (mostly on part of research teams in the CLARIN community) to integrate TEI with CMDI.To achieve this end, it is necessary to express the teiHeader as a CMD profile, to convert existing teiHeaders into CMD records complying with the created profile and to make these records available for harvesting via OAI-PMH.

There have been a number of attempts at expressing teiHeaders as CMD profiles. The first such profile (created by the team of ICLTT in 2010) models the recommended `teiHeader`[2], encoding the `fileDesc` and `profileDesc` components (however leaving out `encodingDesc` and `revisionDesc`). The leaf elements were bound to the most prominent *dublincore* or *isocat* data categories.

Nederlab is a large-scale project starting in Netherlands in 2013 working on the digitisation of historic Dutch language material. In this project, another set of CMD profiles was created, reusing existing components from the 2010 profile. The components `fileDesc` and `profileDesc` were reused, while the components `encodingDesc` and `revisionDesc`, left out in the original profile, were added.

The *Deutsches Text Archiv (DTA)*, a digitising project aiming at a German reference corpus for the period 1650 - 1900, also uses TEI to encode their data and metadata. So far 857 TEI headers have been converted into CMD records and fed into the CMDI. Here again a separate CMD profile has been created.

Yet another approach was applied in the context of other CLARIN-NL projects: Windhouwer (2012) generated (based on an ODD-file) a data category for every element of the `teiHeader` (135 items). In a subsequent step, an enriched schema was generated, that remodells the original `teiHeader` schema, annotating the individual elements with the new data categories (using the `dcr:datcat` attribute). This schema is now maintained in the *SCHEMAcat*, a specialized registry for semantically annotated schemas.

On the basis of Windhouwer's work, a new profile could be defined, binding all the contained components and elements to the corresponding data categories. This would yield a more complex, but also a more systematic and flexible setup, with a clean separation of the semantic space of TEI and the possibility to map the TEI elements (via their data categories) to different data categories needed for specific research questions.

Further work

As can be seen, there are a number of stakeholders concerned with the issue under discussion. Therefore, CLARIN's Centre Committee's metadata curation taskforce coordinated by a member of the ICLTT team will start the work on a common harmonized comprehensive CMD profile for a teiHeader based on the above

[2] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD

mentioned previous work, accompanied by appropriate scripts for conversion of teiHeaders into proper CMD records.

References

D. Broeder, O. Schonefeld, T. Trippel, D. Van Uytvanck, and A. Witt. A pragmatic approach to XML interoperability - the Component Metadata Infrastructure (CMDI), in Balisage: The Markup Conference 2011, vol. 7, 2011. citeulike:9861691.

ISO12620:2009 Computer Applications in Terminology – Data Categories – Specification of Data Categories and Management of a Data Category Registry for Language Resources 2009

D. Broeder, M. Kemps-Snijders, D. V. Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. A data category registry- and component-based metadata framework, in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10) (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, eds.), (Valletta, Malta), European Language Resources Association (ELRA), May 2010.

A. Powell, M. Nilsson, A. Naeve, and P. Johnston. DCMI Abstract Model. tech. rep., Mar. 2005.