Semantic mapping from CMDI to PARTHENOS

Matej Ďurčo
ACDH-OEAW
Vienna, Austria
matej.durco
@oeaw.ac.at

Matteo Lorenzini
ACDH-OEAW
Vienna, Austria
matteo.lorenzini
@oeaw.ac.at

Go Sugimoto
ACDH-OEAW
Vienna, Austria
go.sugimoto
@oeaw.ac.at

Abstract

The Parthenos project aims at pooling resources from existing infrastructures of the broad cultural heritage and humanities cluster. Central to this effort is the common semantic framework -Parthenos Entities -that shall serve as a target model for mapping of information about resources from participating infrastructures. As a representative of linguistic domain, CLARIN will deliver metadata about language resources. Within the Parthenos project separate provisions are foreseen for the mapping task. However, given the complexity of the CLARIN's underlying metadata model (CMDI), traditional one-to-one schema mapping is not applicable and alternative conceptual and technical approach is required. This paper presents the ongoing work on mapping CMDI to the Parthenos model and points out a number of issues identified during the process, partly notorious from the ongoing metadata quality discussion within CLARIN.

1 Introduction

Parthenos (Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies) is a project funded by the European Commission's Horizon 2020 framework programme that started May 2015 running for four years. The project empowers digital research in the fields of history, language studies, cultural heritage, archaeology, and related fields across the humanities. It aims to establish interoperability between several existing research infrastructures, allowing to find, use and combine information from different domains. Consequently a central endeavour is the harmonisation and aggregation of heterogeneous data coming from the participating research infrastructures into a common semantic framework called Parthenos Entities (PE).

CLARIN is a major partner in Parthenos with regard to language resources and language studies in general. It has operated one of the biggest catalogues of language resources in Europe, Virtual Language Observatory (VLO), since 2010. It aggregates the metadata about the resources from over 60 data providers, containing more than 900.000 records. The backbone of CLARIN and the VLO is CMDI (Component Metadata Infrastructure) (Broeder et al., 2011). This paper aims to present an approach adopted for the mapping between the CMDI and the PE model.

2 Underlying components

2.1 CMDI

The CMDI provides a framework to create and (re)use self-defined metadata formats. It relies on a modular model of reusable components, which are assembled together to define profiles serving as blueprint custom schemas which can be used for metadata authoring. The Component Registry is a central place for creation and discovery of metadata components and profiles to promote their reuse and sharing. The registry contains all CMD components and profiles used to describe all metadata in the VLO. Currently, it contains around 1.000 components and around 200 profiles. Fields in the components are linked for semantic grounding to concepts defined in the CLARIN Concept Registry (CCR).

2.2 Parthenos entities (PE) and tools

In order to fulfil the central goal of capturing and representing the research process, Parthenos proposes an ontological model and RDF schema which is able to describe the entities involved in knowledge creation and made available through the research infrastructure on a high level of abstraction. The model is composed of four main entities, *dataset*, *actor*, *service*, and *software*, and defines their categorical descriptions through a minimal metadata set. It is important to note that the mappings and transformations to the PE are by design lossy, i.e. it does not aim at representing all aspects of the source data in the target model, but rather establish the identities of the main entities and relations between them.

The PE model is formalised based on CIDOC-CRM[1] and its extension, CRMdig. The former is able to capture the knowledge of cultural heritage and the latter to describe the provenance of information and digitisation process. The PE model defines 33 classes and 37 properties as the specialisations of the entities defined in the base ontology, although both these additional entities and selected classes and properties of the base ontology are used for the mapping

Within Parthenos, the 3M mapping tool[2] (Minadakis et al. 2015) collaboratively defines mappings from different data models encountered in the participating research infrastructures into one common model, the PEs. The mappings are expressed in the X3ML mapping definition language[3]. These mappings serve as input for the customisable aggregation infrastructure, D-Net[4], which powers a number of aggregation setups, among others, the large-scale research publication portal OpenAire. The D-Net and 3M are being integrated into a hybrid data infrastructure, d4science, running on gCube, to be used as the central content and service provisioning platform of Parthenos.

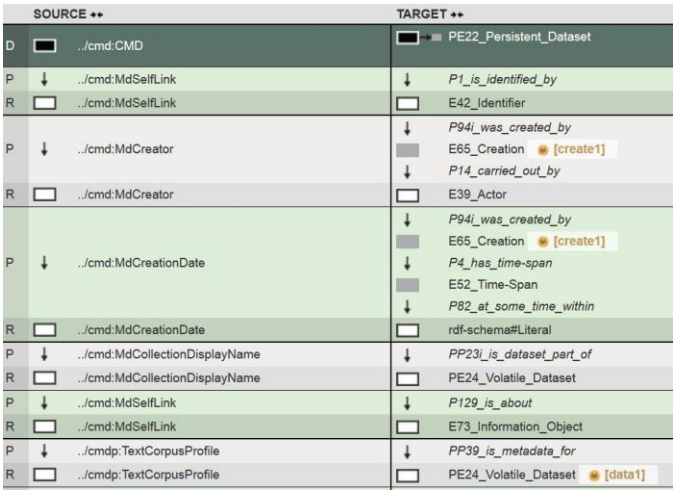

Figure 1. Screenshot of the 3M mapping tool

3 Semantic mapping

3.1 Mapping approach

The default approach to mapping is 1:1 crosswalk between a local source schema and the target schema CIDOC-PE. However the CMDI is not just one schema but a framework for creating and reusing schemas. It is, therefore, not feasible to define the mapping in a traditional way. Instead, we apply the same approach already employed in the VLO, which is a mapping relying on the built-in semantic interoperability layer -semantic binding of the structural elements of the CMDI profiles to well-defined concepts. The developed solution basically adds an indirection layer, identifying concepts used in the CMDI which are (near) equivalent to the individual PE properties, deriving specific XPath patterns for any profile by matching concepts in the corresponding XML schema, to finally tune back into the default workflow using the XPaths to extract values from actual CMD instances (records) to generate a corresponding entity description adhering to the PE model.

[1] http://www.cidoc-crm.org/
[2] https://mapping-d-parthenos.d4science.org/3M/
[3] https://github.com/delving/x3ml
[4] http://d-net.research-infrastructures.eu/

Following scenarios for technical integration are thinkable: a) a dedicated VLO-inspired software component responsible for data transformation and ingestion can become a part of the D-Net aggregation infrastructure, b) custom XSL stylesheets (natively supported by D-Net) can be generated to perform the transformations, or c) mappings are generated in X3MLformat as input for the 3M tool, pushing as much processing logic as possible to the default aggregation pipeline on the Parthenos side. We chose the latter option and developed a simple Java application that does not do the actual transformation of the records, but only generates the X3ML-mapping files specific for individual CMD profile. The development of the automatised procedure was informed by numerous iterations of manual mappings on a few sample records within the 3M tool. Currently, the aggregation machinery is being set up, so that soon a large scale validation of the mappings on actual real data will be possible.

3.2 Global mapping and local mapping

There are many ways how the source data can be expressed in the target model. To ensure conceptually sound mappings and a harmonised approach among the infrastructures, numerous meetings have been held in the project over the last two years and a number of modelling/mapping decisions has been taken. In the following we present a few of these for illustration. Following the general model of CMD framework, we distinguish the global mapping of the generic CMD envelope applicable to all CMD records (selected mapping examples in Table 1) from the local mappings custom to the individual CMD profiles (TextCorpusProfile is presented in Table 2 as an example).

CMDI XPath	***CIDOC-PE relation pattern***	***Note***
/cmd:CMD	crmpe:PE22_Persistent_Dataset	Metadata record itself also represented as first-class citizen
./cmd:Header	PE22 → crmdig:L11i_was_output_for → D7_Digital_Machine_Event	Creation of the record as an Event
…cmd:MdProfile	D7 → crmdig:L23_used_software_... → crmpe:PE38_Schema	CMD schema as the "software" used in the creation event
//cmd:Components /cmdp:*	PE22 → crmpe:pp39_is_metadata_for → crmpe:PE24_Volatile Dataset	Explicit aboutness-relation between record and resource
…cmd:ResourceProxy	→ crmpe:pp39_is_metadata_for → crmpe:PE24_Volatile Dataset→ crmpe:PP8i_is_dataset_hosted_by → crmpe:PE15_Data_E-Service	Relation between the one CMD record to potentially many described resources
…cmd:MdCollectionDisplayName	crmpe:PE24_Volatile_Dataset→ crmpe:PP23i_is_dataset_part_of → crmpe:PE24_Volatile_Dataset → crm:P1_is_identified_by → crm:E41_Appellation	Part of relation between the resource (not the metadata record!) and a collection

Table 1. Selected generic mappings

CMDI	***CIDOC-PE relation pattern***
//cmd:Components/cmdp:TextCorpusProfile	crmpe:PE24_Volatile_Dataset
.//cmdp:ProjectName	→ crm:P1_is_identified_by → crm:E41_Appellation
.//cmdp:PublicationDate	crmpe:PE24_Volatile_Dataset → crm:P94i_was_created_by → crm:E65_Creation→ crm:P4_has_time-span → crm:E52_Time-Span → crm:P81_ongoing_throughout → rdfs:Literal
./cmdp:Collection/cmdp:Access/cmdp:Contact	PE24 → crm:P105_right_held_by →crm:E40_Legal_Body
./cmdp:Contact/cmdp:Address	E40 → crm:P76_has_contact_point → crm:E45_Address
./cmdp:Contact/cmdp:Person	E40 → crm:P107_has_current_or_former_member → crm:E21_Person

Table 2. Examples of local mappings

4 Mapping issues and challenges

During the mapping process we encountered several issues which may make further acquisition and aggregation process more difficult. A major issue is the oftentimes ambiguous or underspecified semantics of numerous structures/expressions used in the CMDI. The fore-most example is cmd:ResourceProxy. One metadata record can contain a number of ResourceProxies (cmd:ResourceProxyList{1}/ cmd:ResourceProxy{1...n}) expressing three different semantics:

1. Different access points for or representations of the same resource – raw data (possibly in various formats), landing page, search page, and search service)
2. The record represents a collection and all ResourceProxy elements point to other metadata records describing the items of the collection
3. The record represents a number of distinct resources, each described by a separate structure in the profile-specific part of the CMD record (linked via an id attribute)

This setup is by design and is algorithmically distinguishable, however it requires specific provisions in the mapping process, i.e. the injection of business logic beyond declarative crosswalk definitions. A general shortcoming in the CMDI semantics is the oftentimes unspecified persistent nature of the described resource (e.g. can the resource change, or is it immutable?), and the mingling of information about a provided web service and underlying software. The PE makes a clear distinction between a Volatile (PE24) and a Persistent Dataset (PE22), as well as between Software (D14) and Service (PE1 or PE8 for E-Service), which are partly impossible to deduce from the information present in the CMD records. An example of problematic semantics on the instance level is the different values in the cmd:MdCreator element, denoting person names, projects, collections, and software involved in the creation of the records. Besides these issues, the well-known problems of metadata quality in CLARIN percolate to the mapping. These are mainly the coverage (King et al., 2015), i.e. missing values for aspects of a resource, and the variability of values, especially those denoting entities like organisations (Ostojic et al., 2016). Both issues influence the quality of the resulting harmonised metadata, the latter being especially problematic to establish identities for main entities and to make actors (i.e. organisations and persons) unambiguous first-class citizens in the CIDOC-PE data space.

5 Conclusion

In this paper, we described the ongoing work on mapping the CMDI to the CIDOC Parthenos Entities. The mapping strategy relies on the semantic interoperability and mapping mechanisms established in the CLARIN infrastructure. We believe not only that the mapping of the CLARIN metadata to the PE model is a good academic exercise and a valuable contribution to the CMDI, but also that the CLARIN community can benefit greatly from expressing the information about the resources in a well-established high-level conceptual model, like the CIDOC-CRM. Conversely, the process of mapping of the CMDI metadata to the PE also allows us to identify potential omissions in the PE model and has proven a valuable input for the modelling work.

Reference

[Broeder et al. 2011] D. Broeder, O. Schonefeld, T. Trippel, D. Van Uytvanck, and A. Witt. 2011. A pragmatic approach to XML interoperability—the Component Metadata Infrastructure (CMDI). In *Proceedings of Balisage: The Markup Conference 2011*. Balisage Series on Markup Technologies, volume 7.

[King et al.2016] M. King, D. Ostojic, M. Ďurčo, and G. Sugimoto. 2016. Variability of the Facet Values in the VLO–a Case for Metadata Curation. In Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland (pp. 25–44) Linköping University Electronic Press.

[Minadakis et al. 2015] N. Minadakis, Y. Marketakis, H. Kondylakis, G. Flouris, M. Theodoridou, M. Doerr, and G. de Jong. 2015. X3ML framework: an effective suite for supporting data mappings. In *Workshop for Extending, Mapping and Focusing the CRM*—co-located with TPDL'2015

[Ostojic et al.2016] Ostojic, D., G. Sugimoto, and M. Ďurčo. 2016. Curation module in action -preliminary findings on VLO metadata quality. In *Proceedings of CLARIN Annual Conference*, 2016. Aix-en-Provence