Something will be connected
- Semantic mapping from CMDI to Parthenos Entities

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

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

Go Sugimoto
ACDH-OEAW
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 data model for mapping of metadata about resources from participating infrastructures. Acting as a representative of the 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 CLARIN's underlying metadata model (CMDI), traditional one-to-one schema mapping is not applicable and an alternative conceptual and technical approach is required. This paper presents the current mapping solution and points out a number of issues identified during the process partly perpetuated from the ongoing metadata quality discussion within CLARIN.

1 Introduction

Parthenos[1] (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 and runs for four years. The project empowers digital research in the fields of history, language studies, cultural heritage, archaeology, and related fields across the (digital) humanities. It brings together several existing research infrastructures to make it easier to find, use and combine information about main entities involved in the research process from different domains, such as datasets, services or actors. The project aims to establish interoperability in humanities domain, building a bridge between the existing European Research Infrastructure Consortiums including CLARIN[2], DARIAH[3], EHRI[4], ARIADNE[5], CENDARI[6], CHARISMA[7], and IPERION-CH[8]. One of the biggest challenges is the aggregation of heterogeneous data coming from such different research infrastructures into a common semantic framework called Parthenos Entities model (PE).

[1] http://www.parthenos-project.eu/
[2] https://www.clarin.eu/
[3] http://www.dariah.eu/
[4] https://ehri-project.eu/
[5] http://www.ariadne-infrastructure.eu/
[6] http://www.cendari.eu/
[7] http://www.charismaproject.eu/
[8] http://www.iperionch.eu/home

Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017.
Conference Proceedings published by Linköping University Electronic Press
at www.ep.liu.se/ecp/contents.asp?issue=147. © The Author(s).

CLARIN is a major partner in Parthenos with regard to language resources and language studies in general. It has been operating one of the biggest catalogues of language resources in Europe, Virtual Language Observatory (VLO)[9], since 2010 (Van Uytvanck et al., 2012; Eckart et al., 2015). It aggregates the metadata about the resources from over 60 data providers, containing more than 1.6 million records. The backbone of the VLO is CMDI[10] (Component Metadata Infrastructure) (Broeder et al., 2011; Goosen et al., 2014) which offers a flexible standardised framework to facilitate formalised descriptions for a wide range of resources, aimed at fostering resource discovery within the linguistic domain and beyond. In order to deliver the information about CLARIN resources to Parthenos, it is required to map the metadata schemas defined in CMDI to PE. This paper presents an approach adopted for this mapping and highlights the encountered problems.

2 Underlying standards and components

In the following, we introduce the standards and components that play a role in the mapping task.

2.1 Component Metadata Infrastructures (CMDI)

CMDI provides a framework for creating and (re)using self-defined metadata schemas in order to meet various needs of data providers, and yet to set a mechanism to aggregate and unify heterogeneous metadata of language resources. It relies on a modular model of reusable components, which are assembled together to define profiles serving as a blueprint for custom schemas to be used for new metadata creation. The CMDI Component Registry[11] (Broeder et al., 2010) was created as a central online environment for the 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 metadata in VLO, currently holding around 1.000 components and around 200 profiles. To enable semantic interoperability between the various profiles, fields in the components are linked to well-defined concepts, primarily drawn from the CLARIN Concept Registry (CCR[12]) (Schuurman et al., 2015), a separate module of CMDI, which allows to openly specify stable definitions of semantic concepts.

2.2 Common Semantic Model – Parthenos Entities Model (PE)

Parthenos proposes a common ontological model, Parthenos Entities, to be able to describe, in a generic manner, basic characteristics of all main entities involved in the knowledge generation process as encountered in the source metadata records, irrespective of the peculiarities of individual source formats. The model is composed of four main entities:

- *PE18 Dataset*: defined in PE model, sets or collections of data, records or information (provided by participating infrastructures) that constitute distinct units of information in the knowledge generation process.
- *E39 Actor*: defined in main CIDOC CRM ontology, is an individual or a group that exercises agency in the knowledge generation process, for which they are responsible.
- *PE 1 Service*: defined in PE model, represents the ability and willingness of an actor to execute on demand by a client certain activities of specific benefit to the client. The service includes all auxiliary abilities of the same actor to execute the respective activities, but not services provided by third parties in the course of their service provisioning.
- *D14 Software:* defined in CRMdig extension, represents an artefact that can be executed on a computer to perform specific operations. In particular, software is the necessary information to process datasets algorithmically and to integrate datasets in a collaborative infrastructure.

[9] https://vlo.clarin.eu

[10] https://www.clarin.eu/content/component-metadata

[11] https://catalog.clarin.eu/ds/ComponentRegistry/

[12] https://www.clarin.eu/ccr

The categorical description of these entities is defined by a minimal metadata set. The minimal metadata set is not meant to represent all the information present in the source metadata, but solely to establish an identity for any entity mapped from the graph, i.e. if it is the same or different from another aggregated entity. Thus the mappings and transformations to the PE are lossy by design. The PE model is not intended to capture all the structure and semantics of CMDI, let alone to replace CMDI or any other of the source formats. The goal of Parthenos is merely harmonisation of basic information about resources aggregated from different research infrastructures to enable resource discovery in a unified manner.

The PE model is formalised based on CIDOC CRM and its extension CRMdig. The former serves to capture the information about cultural heritage and the latter to describe the provenance of the information and digitisation process.

The CIDOC CRM, which became an ISO standard in 2006, is an ontology comprising 86 classes and 138 properties which provides definitions and a formal structure for describing the implicit and explicit concepts and relationship used in cultural heritage documentation. It is also intended to be used as a top-level ontology to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information (ICOM/CIDOC CRM Special Interest Group 2017). It does this by defining very general concepts like space, time, object, event, activity, etc., which are independent of a particular problem or domain, while providing also cultural heritage specific properties such as “curated”, “used specific technique” and “has current keeper”. The CRMdig[13], developed as compatible extension of CIDOC CRM, is an ontology and a RDF Schema for encoding metadata about the steps and methods of production ("provenance") of digitization products and synthetic digital representations such as 2D, 3D or even animated models. The PE model additionally defines 33 classes and 37 properties as specialisations of entities defined in the base ontology, though in the target model both the additional entities and selected entities and properties from CIDOC CRM and CRMdig are used. The adoption of CIDOC CRM and CRMdig as a baseline of the PE enables us to maximise the data interoperability and thus support resource discovery across different cultural heritage and humanities domains.

2.3 Parthenos infrastructure components

Within Parthenos, the 3M mapping tool (Minadakis et al. 2015) is employed to collaboratively define mappings from different data models encountered in the participating research infrastructures into one common model, the PE. 3M is an online open source tool for managing the mapping definition files expressed in X3ML[14], an XML-based schema for describing schema mappings from XML to RDF (see Listing 1 for a sample). 3M assists the users during the mapping definition process with a human-friendly user interface that suggests and validates the user input against the source and target schemas. The structure of an X3ML file consists of: 1) a header with basic provenance information such as the date of creation and the author of the mapping file; 2) a series of mappings, each containing a domain and a number of ‘link’ elements with a ‘path’ and a ‘range’ to map the source values to. Each link describes the relation (path) of the domain entity to the corresponding range entity.

Listing 1. Sample mapping in X3ML format

```
<mapping>
   <domain>
      <source_node>/cmd:CMD/cmd:Resources/cmd:ResourceProxyList/cmd:Re-
sourceProxy/cmd:ResourceRef</source_node>
        <target_node>
           <entity>
              <type>crmpe:PE29_Access_Point</type>
              <instance_generator name="UUID"/>
           </entity>
        </target_node>
   </domain>
   <link>
      <path>
         <source_relation>
            <relation>/cmd:ResourceType</relation>
         </source_relation>
```

[13] http://www.ics.forth.gr/isl/index_main.php?l=e&c=656
[14] https://github.com/delving/x3ml

```
        <target_relation>
           <relationship>crm:P28_has_type</relationship>
         </target_relation>
      </path>
      <range>
         <source_node>/cmd:ResourceType</source_node>
         <target_node>
            <entity>
               <type>crm:E55_Type</type>
               <instance_generator name="ConceptURI_2step"> ...
            </entity>
         </target_node>
      </range>
   </link>
</mapping>
```

These mappings serve as input for the customisable aggregation infrastructure, D-Net[15], which allows to select and configure the needed services and easily combine them to form complex automated data processing workflows. Its scalability and reliability are proven as it powers a number of aggregation platform, for example, the huge research publication portal OpenAire[16]. For the Parthenos project, the 3M engine has been integrated into D-Net infrastructure to support the aggregation of metadata records from the source research infrastructures based on mappings expressed in X3ML language. D-Net itself is integrated into the hybrid data infrastructure d4science[17], Parthenos' central content and service provisioning infrastructure based on the software solution gCube[18]. It forms the Parthenos Content Cloud Framework (CCF), the component responsible for the whole aggregation, transformation, storage and indexing workflow. In this framework aggregated and transformed metadata records are transformed into different formats and ingested into multiple storage and indexing components which serve as endpoints for resource discovery applications: a) as RDF adhering to PE model into a Virtuoso[19] triple store, allowing full-fledged complex SPARQL[20] queries on the whole RDF graph, b) flattened into indices of an Apache Solr instance for full-text search systems and c) as serialized RDF available via an OAI-PMH[21] endpoint. Figure 1 depicts the whole metadata aggregation and provisioning infrastructure employed in Parthenos.

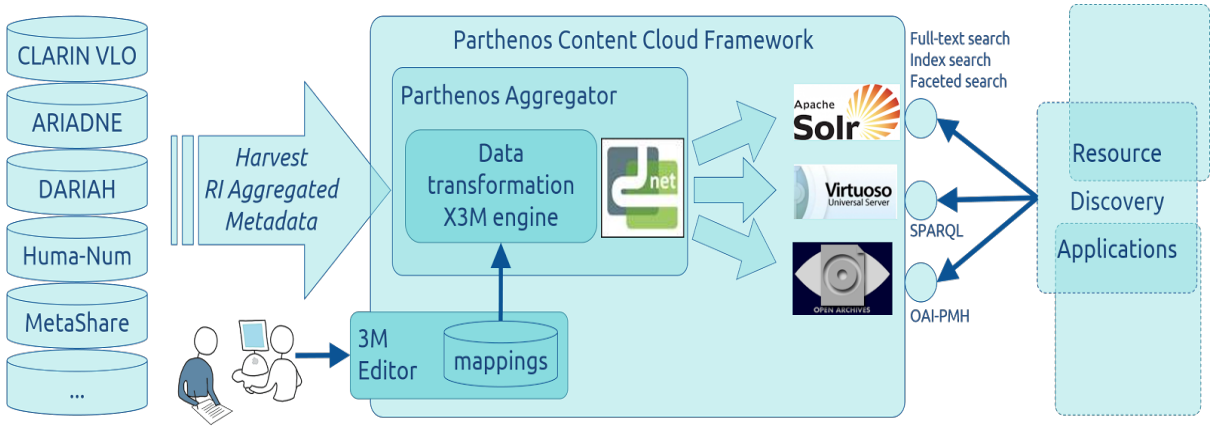

Figure 1. Diagram of the Parthenos metadata aggregation and provisioning infrastructure

[15] http://d-net.research-infrastructures.eu/
[16] https://www.openaire.eu/search/find
[17] https://www.d4science.org/
[18] http://gcube-system.org/
[19] https://virtuoso.openlinksw.com/
[20] https://www.w3.org/TR/rdf-sparql-query/
[21] https://www.openarchives.org/pmh/

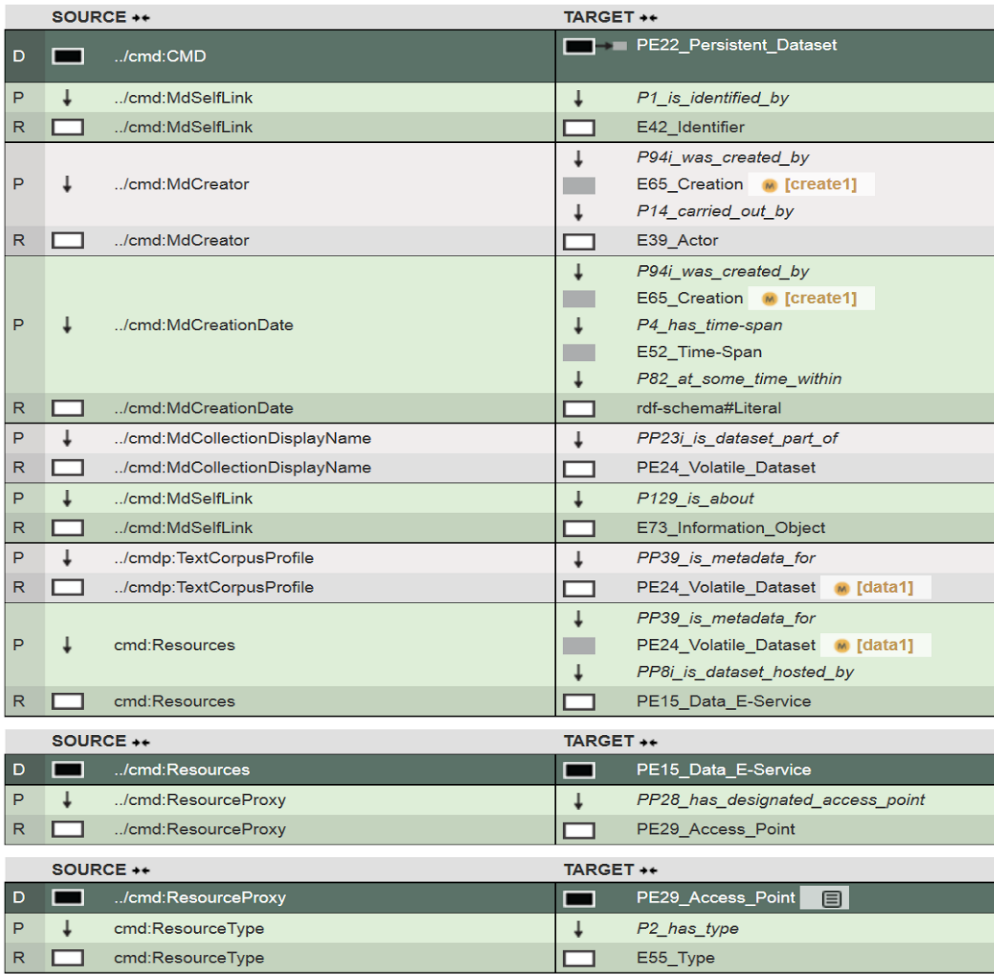

Figure 2. Screenshot of the 3M mapping tool

3 Mapping

3.1 Method

The default mapping approach in the Parthenos project is a 1:1 crosswalk between a "local" source schema specific to individual research infrastructure and the target schema (PE). However, as outlined in the previous section, CMDI is not just one schema but a framework for creating and reusing schemas. In fact, currently more than 200 different schemas have been defined. It is, therefore, not feasible to define the mapping in this traditional way. Instead we take advantage of the mechanism already employed in the VLO, which is a mapping relying on the built-in semantic interoperability layer, that is, the semantic binding of the structural elements of CMD profiles to well-defined concepts. The developed mapping solution aims to identify PE properties which are (near) equivalent to the concepts of CCR (Figure 3. Mapping Definition Phase), to derive XPath[22] patterns for any profile by matching concepts in the corresponding XML schema (Figure 3. Profile Pre-processing Phase), and finally to use the XPaths to extract values from actual CMD instances (records) to generate a corresponding entity description adhering to the PE model (Figure 3. Aggregation Phase).

While the basic mechanism is similar to the one applied for populating the VLO, the specific context is quite different, requiring a new custom solution. In particular, the question is how to integrate the automatic mapping step, i.e. the resolution of concepts to appropriate XPaths, into the foreseen aggregation pipeline, aimed at extracting values from source metadata and generating the target structured records. Following scenarios were considered: a) the VLO software component responsible for data

[22] https://www.w3.org/TR/xpath/

transformation and ingestion can become a part of the D-Net aggregation infrastructure (with some adjustments), b) custom XSL stylesheets (natively supported by D-Net) can be generated, or c) the generated mapping is converted to a format required by X3ML, pushing all processing logic to the Parthenos side. We chose the third option and developed a simple java application[23] that does not do the actual transformation of the records, but only generates the specific X3ML-mapping files, based on a mapping file template containing multiple concepts and fall-back XPaths (as is the case in the concepts to facets file serving as input for VLO-importer) in specific locations to be resolved against a given individual CMD profile. The entire procedure is depicted in Figure 3.

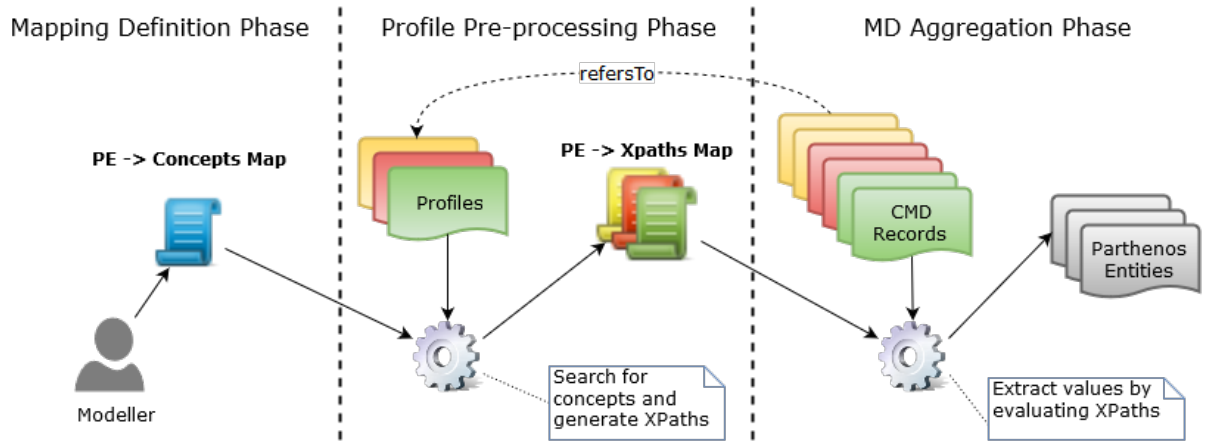

Figure 3. CMDI to Parthenos Entities mapping generator algorithm

3.2 Mapping decisions

There is a broad leeway in how the source data can be expressed in the target model (PE). To ensure conceptually sound mappings as well as a harmonized approach among the infrastructures, a number of modelling decisions were taken. We present some of them below in Table 1.

Based on these general modelling conventions, we defined mappings from CMD schemas to PE in an iterative collaborative process. Following the general model of the CMDI framework, we distinguish between global mappings of the generic CMD envelope applicable to all CMD records (selected examples in Table 2) and 'local' mappings custom to individual CMD profiles (Table 3).

[23] https://github.com/acdh-oeaw/parthenos_mapping

Type of information	PE	Note
Values	E40_Legal_Body → P3_has_note → rdfschema#Literal, E35_Title → P1_is_identified_by → E41_Appellation → rdfs:label	If the entity refers to a literal value, the referred data is mapped as rdf:literal. If the entity refers to a value string, the referred data is mapped as rdfs:label
Publication date	PE24_Volatile_Dataset → crm:P94i_was_created_by → **crm:E65_Creation** → crm:P4_has_timespan → crm:E52_Time-Span → crm:P82_at_some_time_within → http://www.w3.org/2000/01/rdfschema#Literal	interpreted as the creation date of the resource PE24_Volatile_Dataset or as the date on which curaton of the dataset begins
Title	crmpe:PE24_Volatile_Dataset → crm:P1_is_identified_by → crm:E41_Appellation → rdfs:label	
Email, phone	E74_Group/crm:E40_Legal_Body/crm:E21_Person → crm:P76_has_contact_point → crm:E51_Contact_Point [crm:E55_Type = "parthenos-type:email"]	Further specify type of contact point with E55_Type
URL, handle	crmpe:PE22_Persistent_Dataset → crm:P1_is_identified_by → crm:E42_Identifier	URL to encode is typed as crm:E42_Identifier

Table 1. Selected general modelling decisions

CMDI XPath	PE	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:Header	**crmdig:D7_Digital_Machine_Event**	
cmd:MdCreationDate	D7 → crm:P4_has_time_span → crm:E52_Time_Span → crm:P82_at_some_time_within → **rdfschema#Literal**	When did the creation event happen
cmd:MdCreator	D7 → crmdig:L23_used_software_... → **crmpe:PE21_Persistent_Software**	Field cmd:MdCreator is very heterogeneous containing references to persons, institutions, projects as well as software

		applied for generation. Proposed mapping reflects the last variant.
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:Header/ cmd:MdCollectionDisplayName	crmpe:PE24_Volatile_Dataset [resource!] → 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 2. Selected global mappings

CMDI	PE
../cmdp:TextCorpusProfile	crmpe:PE24_Volatile_Dataset
→ cmdp:Name	→ crm:P1_is_identified_by → crm:E41_Appellation
→ cmdp:Title	→ crm:P1_is_identified_by → crm:E35_Title
→ cmdp:Owner	→ crm:P105_right_held_by →crm:E40_Legal_Body
→ cmdp:Description	→ crm:P3_has_note → rdfs-schema#Literal
→ cmdp:Project	→ crm:P94i_was_created_by → crm:E65_Creation → crmpe:PP43i_is_project_activity_supported_by → crmpe:PE35_Project
→ cmdp:Availability	→ crm:P129i_is_subject_of → crm:E30_Right → crm:P3_has_note → rdf-schema#Literal
../cmdp:Access	crmpe:PE15_Data_E-Service
→ crmdp:Contact	crmpe:PP2_provided_by → crm:E40_Legal_Body

Table 3. Examples of local mappings

3.3 Current status

Based on the experience we gathered while manually defining mappings in X3ML for 3 sample profiles (teiHeader[24], TextCorpusProfile[25], and OLAC-DcmiTerms[26]), we derived two templates as expected by the mapping generator, one for datasets, the other for services, and furnished these with the most frequently referred concepts to be resolved against the individual schemas. These manually defined mappings were applied on a sample collection of roughly 3.000 CMD records, which were processed through the Content Cloud Framework and made available as PE conformant RDF.

In a next step, we identified all CMD profiles with records in a recent VLO data dump, and based on the template files we automatically generated maps for all these currently employed profiles.

The initial transformation of the small sample dataset is an important milestone demonstrating the feasibility of the approach and established connectivity. However, it also revealed many issues on various levels of the aggregation process, prompting a feedback loop to fine-tune the individual steps of the transformation workflow: a) the generation of profile-specific mappings, b) mapping from CMDI to PE; c) normalisation, harmonisation of values; d) transformation and ingest from PE to a flat index-search engine (Solr). Finally, there is also a possibility that the problem already lies in the source data (CMD records) as delivered by the original service providers (cf. section 4 Issues and challenges).

4 Issues and Challenges

During the initial mapping process, we encountered several issues which will have adverse effect on the discovery and exploitation of the aggregated data. A major issue arising in the mapping task is the oftentimes ambiguous or underspecified semantics of numerous structures/expressions used in CMDI. The foremost 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 the same resource. This case is covered by a specific mapping of the cmd:ResourceRef elements as typed PE29_Access_Point entities.
2. The record represents a collection and all ResourceProxies point to other metadata records describing the items of the collection. In this case, the relation between the collection and its members can be expressed using crmpe:PP23i_is_dataset_part_of.
3. The record represents a number of distinct resources. In this case the id-attribute can be referenced from the corresponding XML-elements in the cmd:Components mapping block. This case is not yet fully covered by the mapping provisions.

This setup is by design and is algorithmically distinguishable, but it requires specific provisions in the mapping task, i.e. injection of procedural processing in the mapping process beyond declarative crosswalk definitions. An evaluation on a sample recent VLO data dump with 879.497 CMD records yielded that there are 685.832 records of case 1, 1.421 records of case 2 and 193.662 records of case 3.

Another substantial shortcoming in CMDI semantics is unclear statements about the persistent nature of the described resource (i.e. can the resource change, or is it immutable?), and the mingling of information about a provided web service and the underlying software.

PE makes a clear distinction between Software and Service (D14 vs. PE1 or PE8 for E-Service), but it is partly impossible to derive the difference from CMD records. The PE also distinguishes between a Volatile and a Persistent Dataset (PE24 vs. PE22). While the former is defined as “dataset that are changed without notice or archiving of intermediate states but maintained by an instance of PE12 Data curating Service.” and “are typically whole databases or mash-ups with active data feeds”, the scope of the latter is “datasets that contain collections of data, records or information kept as a persistent unit of information in the knowledge generation process from primary records up to any level of aggregation

[24] clarin.eu:cr1:p_1381926654438
[25] clarin.eu:cr1:p_1271859438164
[26] clarin.eu:cr1:p_1288172614026

or integration". Also in this case, it is sometime impossible to decide to which class given resource belongs, as the original metadata was often created without concerning such difference.

An example of problematic semantics on the instance level is the different values in the cmd:MdCreator element with a mix of around 300 distinct person names, projects, collections, software solutions and scripts involved in the creation of the records[27].

In addition, we have to deal with implicit entities. For example, although there is a lot of information about actors encoded within CMD records (e.g. publisher, organisation responsible for creation of the resource, rights holder etc.), it needs to be extracted to generate the corresponding Actor entities. Here, we are confronted with a long standing issue in CMDI metadata - the variability of descriptors. It is caused inter alia by not using identifiers, but rather just string values to denote entities, like organisations. As a consequence, we are not able to fully identify the same entities described in different variations of vocabularies (e.g "Max Planck Institut" and "MPI" may or may not be the same entity). The normalisation of values is on-going process within the CLARIN's metadata curation taskforce.

In addition, we encountered information gaps. Even if a record contains information about the corresponding actor, in most cases it is not sufficient to build a full description, such as the hierarchy of organisations. It needs to be either collected from other sources, or curated manually. Nonetheless, in the specific case of organisations, we can build on the work done in the CLAVAS project[28], where organisations from the VLO were extracted, manually curated, and published as a vocabulary.

Moreover, the well-known problems of metadata quality under discussion in the context of CLARIN resurface in the mapping task. Of note among these are, in particular the (facet) coverage (King et al., 2015), i.e. missing values for specific aspects of a resource description, and the variability of values, especially those denoting entities like organisations (Ostojic et al., 2016). Both issues have strong influence on the quality of the resulting harmonized metadata and dramatically hamper the recall. The latter is especially problematic given the goal of the overall Parthenos mapping task to establish identities for main entities, and make also actors (e.g. organisations and persons) first-class citizen in the CIDOC-PE data space.

5 Conclusion

In this paper, we describe the ongoing work on mapping CMDI metadata to Parthenos Entities model. The mapping strategy relies on semantic interoperability mechanisms established in the CLARIN infrastructure. We introduced an intermediate processing step, in which a hand-crafted template file furnished with CCR concepts is expanded by a dedicated small utility Java application into a valid X3ML mapping file with XPaths corresponding to given concepts relative to a specific CMD schema. These generated mapping files are used by the integrated transformation framework D-Net to extract values from CMD instances and to generate an entity description in PE model. After the crafting of the template file based on two profiles, mapping files for all profiles encountered in the VLO were generated.

During our mapping effort, several problems were recognised. One of the major issues was the semantic ambiguity and lack of explicit statements regarding crucial aspects of the described resources in numerous structures in the CMD records, for instance concerning the distinction between a software and a service or between a volatile and a persistent dataset. In addition, well-known metadata quality issues such as missing values and variability of values cause mapping errors.

We strongly believe, that the task of mapping the CLARIN metadata to the PE model is not only an academic exercise and a one-way contribution, but also that CLARIN's metadata infrastructure and community can benefit greatly from expressing the information about the resources in a well-established high-level conceptual model like CIDOC CRM. Conversely, the process of mapping the complex CMDI metadata also allows us to identify potential omissions in the PE model and has proven useful for the modelling work.

[27] https://github.com/acdh-oeaw/parthenos_mapping/blob/master/cmdi_utils/mdCreator_values.txt
[28] https://openskos.meertens.knaw.nl/clavas/

The mappings between PE and other schemas of different infrastructures are in the final phase. When our mapping is completed, Parthenos will be able to harvest and aggregate metadata from all the participating infrastructures, offering the users access to a comprehensive aggregation of datasets and tools in cultural heritage for their research.

References

[Broeder et al. 2010] D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. 2010. A data category registry-and component-based metadata framework. In *Procedings of the Seventh Conference on International Language Resources and Evaluation*. Pp. 43-47.

[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.

[Doerr 2003] M. Doerr. 2003. The CIDOC Conceptual Reference Module: an ontological approach to semantic interoperability of metadata. In *AI magazine*, 24(3):75.

[Eckart et al. 2015] T. Eckart, A. Helwig, and T. Goosen. 2015. Influence of Interface Design on User Behaviour in the VLO. In *CLARIN Annual Conference 2015 Book of Abstracts*. https://www.clarin.eu/sites/default/files/book%20of%20abstracts%202015.pdf

[FORTH-ICS 2017] PARTHENOS Entities: Research Infrastructure Model V2.0.

[Goosen et al.2014] T. Goosen, M. Windhouwer, O. Ohren, A. Herold, T. Eckart, M. Ďurčo and O. Schonefeld. 2014. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. In *Selected Papers from the CLARIN 2014 Conference*. Pp. 36-53.

[ICOM/CIDOC CRM Special Interest Group 2017] Definition of the CIDOC Conceptual Reference Model Version 6.2. 3 October 2017.

http://www.CIDOC CRM.org/sites/default/files/2017-12-30%23CIDOC%20CRM_v6.2.3_esIP.pdf

[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 Con-ference 2015, October 14–16, 2015, Wroclaw, Poland (pp. 25–44) Linköping University Electronic Press. http://www.ep.liu.se/ecp/123/003/ecp15123003.pdf

[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] D. Ostojic, G. Sugimoto, and M. Ďurčo. 2016. Curation module in action - preliminary findings on VLO metadata quality. Retrieved from https://www.clarin.eu/sites/default/files/ostojic-etal-CLARIN2016_paper_22.pdf

[Schuurman et al. 2015] I. Schuurman, M. Windhouwer, O. Ohren, and D. Zeman. 2015. CLARIN concept registry: the new semantic registry replacing ISOcat. In *CLARIN Annual Conference 2015*. Pp. 80–83.

[Van Uytvanck 2012] D. Van Uytvanck, H. Stehouwer, and L. Lampen. 2012. Semantic metadata mapping in practice: The Virtual Language Observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. Pp. 1029-1034.