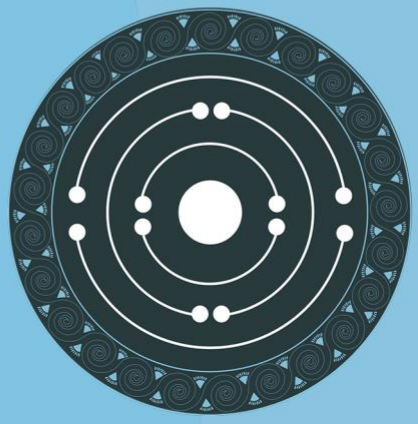


PARTHENOS

Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies

D5.6 Report on Mappings (Final)

CNR

PIN

CLARIN

FORTH

OEAW

MIBACT-ICCU

SISMEL

OVI

CNRS/Huma-Num

28 February 2019

PARTHENOS is a Horizon 2020 project funded by the European Commission. The views and opinions expressed in this publication are the sole responsibility of the author and do not necessarily reflect the views of the European Commission.

HORIZON 2020 - INFRADEV-4-2014/2015:

Grant Agreement No. 654119

PARTHENOS

Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies

REPORT ON MAPPINGS (FINAL)

Deliverable Number	D5.6
Dissemination Level	PUBLIC
Delivery date	28 February 2019
Status	Final
Author(s)	Alessia Bardi George Bruseker Matteo Lorenzini Maria Theodoridou Matej Durco

Project Acronym	PARTHENOS
Project Full title	Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies
Grant Agreement nr.	654119

Deliverable/Document Information

Deliverable nr./title	D5.6
Document title	Report on mappings (final)
Author(s)	Alessia Bardi, George Bruseker, Matteo Lorenzini, Maria Theodoridou
Dissemination level/distribution	PUBLIC / CC-BY

Document History

Version/date	Changes/approval	Author/Approved by
V 0.1 12.12.18	Front matters and copy of the previous deliverable version (D5.3)	Alessia Bardi
V 0.2 04.01.19	Updated figures and sections about the mappings	George Bruseker
V0.3 08.01.19	Added new mapping of the PARTHENOS Reference Resource Datasets and a new descriptor for each mapping called “PARTHENOS Entities Generated”	George Bruseker
V0.4 08.01.19	Added screenshots of transformation tests using the RDF Viewer	George Bruseker
V 0.5 22.01.19	Updated info about CulturaItalia mapping	George Bruseker
V 0.6 25.01.19	Clean up – Draft sent to preliminary review	Alessia Bardi

V 0.7 29.01.19	Updated Introduction Updated figures in Section 4.2 Added missing captions to all figures Updated Conclusion and Executive Summary Renamed URI to BASE URI each in mapping information Integrated formatting and clean up as suggested by Sheena	Alessia Bardi
V 0.8 29.01.19	Updated mapping descriptions	George Bruseker
V 0.9 30.01.19	Added CNRS/Huma-Num on the cover page Minor changes to Huma-Num related section	Alessia Bardi
V 0.11 5.02.19	New screenshot for PARTHENOS top level entities. Updated mapping table. Added placeholder for section on metadata quality check.	Alessia Bardi, George Bruseker
V 1.0 18.02.19	Mappings ids synchronised to those in the table. WP5 mapping description moved before WP8.	Alessia Bardi
V 1.1 26.02.19	Added section on metadata quality checking. Updated description of CENDARI mappings Added link to mappings published on Zenodo, adjusted figures, added details on Virtuoso provenance graph. Updated conclusion with paragraph on data quality.	Matej Durco George Bruseker Alessia Bardi
Final 28.02.19	Final review	Sheena Bassett

Table of Contents

1. Executive Summary

In order to create a Joint Resource Registry for the purpose of supporting resource discovery and data integration, a homogenous information space of metadata must be constructed. To this aim, a process for homogenization at different levels is required. The PARTHENOS infrastructure implements this process by applying structural and semantic mappings to the metadata records offered by the Research Infrastructures (RIs) in the consortium.

Mappings are created with state-of-the-art tools and user-friendly graphical user interfaces that support data experts during the definition of mappings from their local data model to the PARTHENOS Entities Model. As of December 2018, the PARTHENOS consortium created twenty-eight official mappings (out of a total of one hundred and fifty-two mappings, mostly created for testing, sampling and getting familiar with the 3M Editor). Sixteen RI users have collaborated in their creation with the support of FORTH. All mappings have been published on Zenodo and are available at https://doi.org/10.5281/zenodo.2574523 under CC-BY license International v4.0.
The analysis of the twenty-eight mappings shows that there is a growing attention in curation and preservation practices for research data, while research software, although considered an important research product that enables research reproducibility and supports the daily work of all researchers in Digital Humanities, is not typically hosted, preserved and curated via services offered by Research Infrastructures.

This deliverable extends and finalises D5.3 “Report on mappings (interim)” that was delivered in October 2017.

2. Introduction

Describing metadata and the digital resource, i.e. creating a registry of digital resources and their metadata, is fundamental to the possibility of resource discovery of basic assets. On the base of such a minimal registry, the process of deep dataset integration can be carried out for particular topics and research themes. In the context of the PARTHENOS Project, Research Infrastructures (RIs) have built and maintain their own registries, customized to their user base and requirements, where resources are described according to different data models, metadata schemas and at different levels of granularity.

In order to build a common Joint Resource Registry across RIs in order to support cross-disciplinary resource discovery and data integration, a certain level of homogenous information space of metadata must be constructed. To this aim, a process for homogenization at different levels is required. In particular, it is possible to identify the following interoperability challenges to be addressed [1]:

Mediation interoperability: Data sources may export metadata and files according to different standard protocols.

Encoding interoperability: Metadata records can be encoded in different ways, e.g. json, XML, CSV, Turtle, etc.

Structural and semantic interoperability of metadata: Metadata records come with different structures, i.e. different encodings (e.g. XML, json) and schemas (e.g. EAD, Dublin Core, CRM), and semantics (e.g. vocabularies, value formats) which differ from data source to data source. Semantics and structure depend on the data source data model, i.e. the entities and relationships used to describe or contextualise the digital objects at hand, but also on the underlying storage platform.

Granularity interoperability of metadata: By *granularity* we mean the level of data model detail represented by one metadata record. In some cases, each record represents one entity of the model (e.g. a Dublin Core record represents and describes one publication

entity); in other cases, it may represent more entities possibly with relationships between them. For example, an EAD[1] record may represent a hierarchy of entities.

The homogenization process proposed and implemented by PARTHENOS is that of:

- Defining a core, cross-disciplinary semantic model / ontology for information management at the Research Infrastructure level. The ontology, named PARTHENOS Entities Model (PE model) is defined as an extension of CIDOC-CRM, as reported in D5.1 “Report on the common semantic framework - Draft” [2] and also in D5.5 “Report on the common semantic framework - Final” [15];
- Supporting the definition of mappings from RIs models to the PE model (T5.2);
- Realization of an Aggregative Data Infrastructure (ADI) [3] (T6.2) capable of:
- addressing interoperability challenges due to the heterogeneity of metadata exposed by RIs in order to construct a homogenous information space of metadata records conforming to a common data model [4]
- expose the generated homogenous information space via different interfaces for the construction of advanced discovery services. In the context of PARTHENOS, the ADI must be able to populate the Joint Resource Registry realised in T6.4, and expose the generated homogenous information space via a SPARQL endpoint and a Solr Index.

The process of homogenization (see Fig. 1) is implemented by two main components of the PARTHENOS technical infrastructure: the 3M Editor and the PARTHENOS aggregator. Data experts use the 3M Editor to define and test mappings from the native data model to the PE model. Mappings are then applied by the PARTHENOS aggregator powered by the D-Net software toolkit, which integrates the X3ML transformation engine. The resulting RDF/XML records are further cleaned by the Metadata Cleaner Service (see Section 4.2 and Deliverable D6.2 [16] for details) and published on Virtuoso, Solr and the Joint Resource Registry.

[1] Library of Congress (2002) *Encoded Archival Description*. Available online at: http://www.loc.gov/ead/

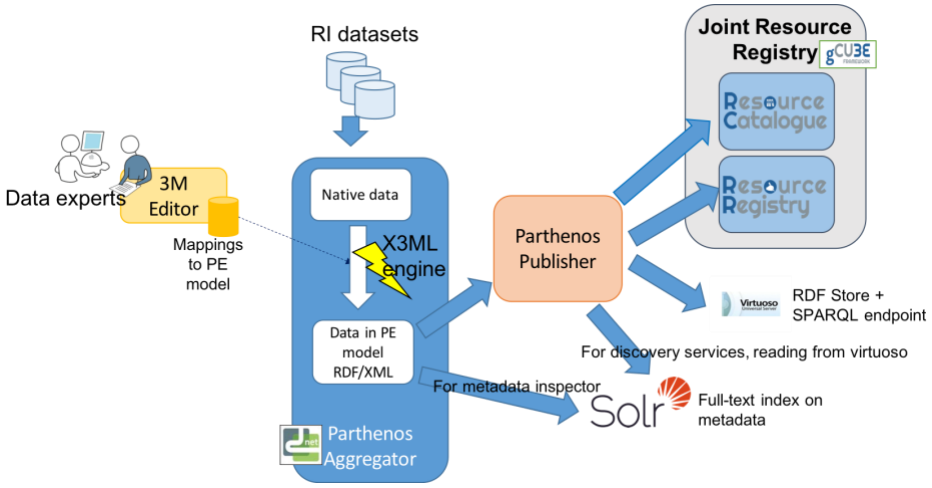


Figure 1. The PARTHENOS harmonization process

2.1 Outline of the report

This deliverable describes the homogenization process implemented by the PARTHENOS technical infrastructure. Section 3 presents the X3ML mapping language and the features it offers to address the interoperability challenges mentioned above. Section 4 describes how data experts can create mappings with the 3M Editor and how those mappings are included and executed by the PARTHENOS aggregator. An overview of the mappings created for PARTHENOS is given in Section 5. All mappings are available on the 3MEditor integrated into the PARTHENOS Virtual Research Environment and their final version is published on Zenodo at https://doi.org/10.5281/zenodo.2574523.

3. The X3ML mapping definition language

The X3ML mapping definition language [5] is an XML based language which describes schema mappings in such a way that they can be collaboratively created and discussed by experts. It is a declarative, human readable language that supports the cognitive process of mapping from one data structure to another. Unlike XSLT that is only intended to be comprehensible by IT technicians, the X3ML mapping definition language can be understood by non-technical actors as well. Thus, a domain expert is capable of both verifying the semantics and reading and validating the schema matching. This model carefully distinguishes between mapping activities carried out by the domain experts, who know and provide the data, and from the IT technicians, who actually implement data translation and integration solutions.

Usually, schema matching is used to describe the process of identifying that two different concepts are semantically related. This allows the definition of the appropriate mappings to be used as rules for the transformation process. However, a common problem is that the IT experts do not fully understand the semantics of the schema matching - it's not their data - and the domain experts do not understand how to use the technical solutions for creating mappings. For this reason, the X3ML Toolkit relies on two distinct components which separate task of schema matching from the more technical processes of URI generation. Schema matching can be fully and independently performed by the domain expert without any specific IT skills and the URI generation by the IT expert, thereby solving the bottleneck that requires the IT expert to fully understand the mapping. Furthermore, this approach keeps the schema mappings between different systems harmonized since their definitions do not change, in contrast to the URIs that may change between different institutions and are independent of the semantics. Moreover, this approach completely separates the definition of the schema matching from the actual execution. This is important because different processes might have different life cycles; in particular, the schema matching definition has a different life cycle compared to the URI generation process. The former is subject to changes more rarely compared to the latter.

The structure of X3ML is quite easy to understand consisting of: (a) *a header* that contains basic information (e.g., title, description, contact persons, the source and target schemata, sample records etc.) and (b) a series of *mappings*, each containing a domain (the main

entity that is being mapped), and a number of links which consist of a path and a range. Each link describes the relation (path) of the domain entity to the corresponding range entity. The basic mapping scheme and the XML representation of an X3ML mapping are shown in Figure 2.

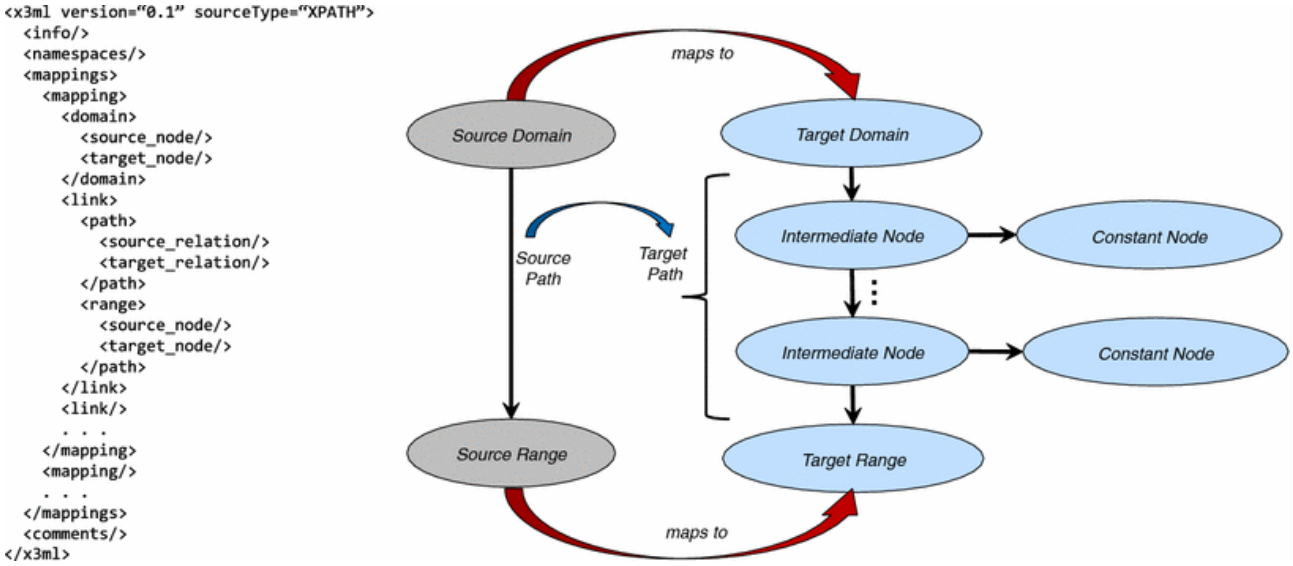


Figure 2. The structure and the XML representation of an X3ML mapping

An X3ML structure consists of:

- the mapping between the source domain and the target domain;
- the mapping between the source range and the target range;
- the proper source path;
- the proper target path;
- the mapping between source path and target path.

The main concepts of the X3ML mapping definition language are the following:

Info and comment: intended to support communications among data experts and technicians and to record provenance information such as the date of creation and the author of the mapping file.

Mapping: Each mapping element consists of a domain element and a number of links. It is quite common to have a single domain mapping and some range mappings, so by using this particular format for mappings, the single domain mapping does not have to be declared again. This is an ergonomic choice good for tree-dominated source schemata or

sets of relational tables, which helps user orientation. It further provides an intuitive default local scope to define that the same instance of a class in a mapping rule appears as domain value in multiple target propositions.

Domain: The domain element is used to specify the mappings between a source (source_node) entity (table, class, non-leaf element) that can be regarded as domain of a source proposition and an equivalent target (target_node) entity. The source_node provides information on how to navigate to the source record and in case of XML it is an XPATH expression. The target_node defines an entity element that will lead to the generation of resource URIs or datatype values for the output graph. It may also contain conditions (described below) upon which the mapping depends.

Link: Inside the link element there is a path element. It allows mapping a source relation from the above defined source domain to a target relation to the above defined target domain. The path element must be followed by a range element, which is used to map the source and the target entities that are the equivalent range of the respective paths. The target relation might contain if conditions as well. A source/target range pair may reappear as a subsequent domain in an X3ML mapping.

Conditional: The conditional expressions in *X3ML mapping definition language* can check for existence, equality and narrowness of values. They are expressed in the form of if statements and can be combined into Boolean expressions.

Intermediate node: Sometimes, a path in the source schema needs to be further analysed to a sequence of paths in the output with respect to the target schema. For this reason, the user can define the generation of an intermediate node (or intermediate entity).

Additional: Regularly constant properties and entities are needed to be added to a target entity, either from background knowledge (provenance information) or to characterize the meaning of a classification by the source schema rather than by data. For instance, a database about museum objects may not mention at all the museum as current keeper. In a database of coins, each coin may be mapped to a “physical object”, of type “coin”. For that purpose, an additional element can be used, containing the entity which will be

attached to the target entity, the relationship describing the link, and the respective constant values.

Variables: Sometimes it is necessary to generate an instance in X3ML only once in the scope of a given domain entity and then re-use it in a number of links of this domain. This is most frequently the case for intermediate target nodes. For example, a description of a museum object may reuse the same production event for mapping its “creator” link and its “date” link. In these cases, a variable is assigned to the re-used entity.

Join operator: Sometimes, it is required to combine values from different tables in the source and produce new values in the output. This is the definition of the relational join operation. X3ML contains a specific operator to support (n-ary) join operation between different tables. The operator that is used is ‘==’ and is being used inside a link element. More specifically, it is being expressed inside the path element and expresses the equality of the value of the left-hand side (its table is the one defined in the domain of the corresponding mapping) with the value of the right-hand side (its table is the one defined on the range of the corresponding link).

Instance generation policy: The definition of the URI generation policy is a process that begins after the schema matching has been accomplished and is usually performed by an IT expert who must ensure that the generated URIs match certain criteria such as consistency and uniqueness. A set of predefined URI generators (UUIDs, literals) and templates are available but any URI-generating function can be implemented and incorporated in the system. In the X3ML definition, the target domain and range contain the functions that generate URIs or literals.

The PARTHENOS instance generation policy is based on the following principles:

- The PARTHENOS base namespace is http://parthenos.d4science.org/handle/ and the registry namespace is http://parthenos.d4science.org/handle/Parthenos/REG/.
- Each RI uses the base namespace to create its own base namespace, which is http://parthenos.d4science.org/handle/{RI-name}/{DB-BeingMapped}

- The generator policy defines instance generation functions for each of the major entities defined in the ontology, at the top level. These include: Project, Service, Dataset, Software, Actor, Place, Thing, Dimension. The namespace path is then generated by using the base namespace of the RI and adding an additional path corresponding to the type of reference entity. It is important to distinguish between entities themselves and their names and, for this reason, each entity also has a corresponding appellation generator. For example:
- PE22 Persistent Dataset from DARIAH GR/DYAS will get the URI: http://parthenos.d4science.org/handle/DariahGR/Dyas/Dataset/595
- PE24 Volatile Dataset from CulturaItalia MUSEID will get the URI: http://parthenos.d4science.org/handle/Culturaitalia/MUSEID/Dataset/work_63550
- The instance generators for Concepts follow the same logic but they use the registry namespace in order to provide a common terminology, for example: http://parthenos.d4science.org/handle/Parthenos/REG/Concept/datasettype/metadata

The latest version of the PARTHENOS generation policy file is version 1.5 available at https://goo.gl/M4yjXV.

4. Using the mappings

4.1 Mapping Memory Manager (3M)

3M is an online open source tool for managing the X3ML mapping definition files. It provides a number of administrative actions that assist the experts to manage their mapping definition files such as create, edit, delete, export, import etc. as well as some basic user registration, authentication and rights management.

The heart of 3M is the 3M Editor component, a Web application suite aimed to assist users during the mapping definition process, using a human-friendly user interface and a set of sub-components that either suggest or validate the user input. The main task of the editor is to support the creation of a complete X3ML file and check how the actual source data are mapped to the defined target output.

To create a complete X3ML file, the user has to fill in information in three tabs:

Info: Contains a general mapping description (Figure 3), the source schema (Figure 3), the target schemas (Figure 4) and the corresponding namespaces (Figure 5). Additionally, it contains some provenance information about the mapping (who did it and how they can be contacted), sample source and target data and the generator policy file (Figure 6).

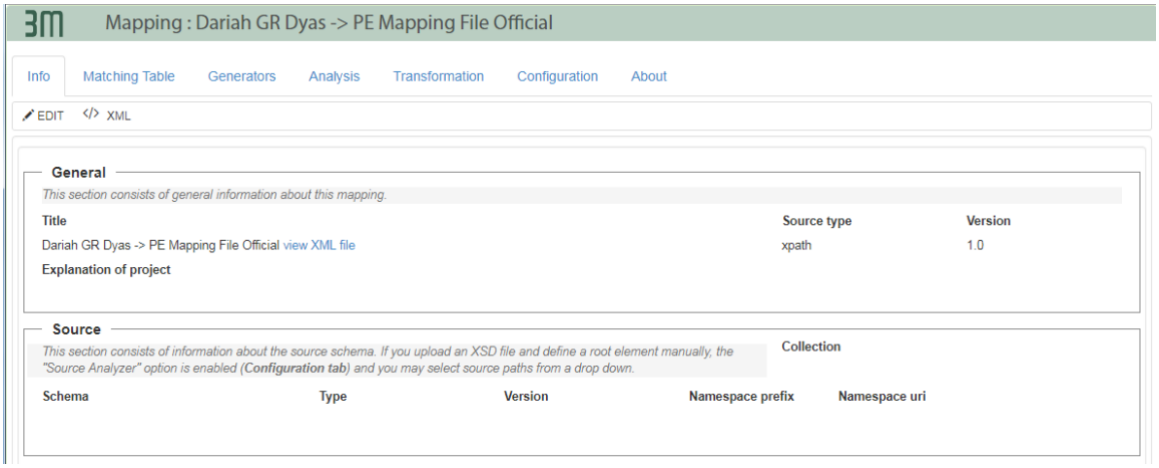
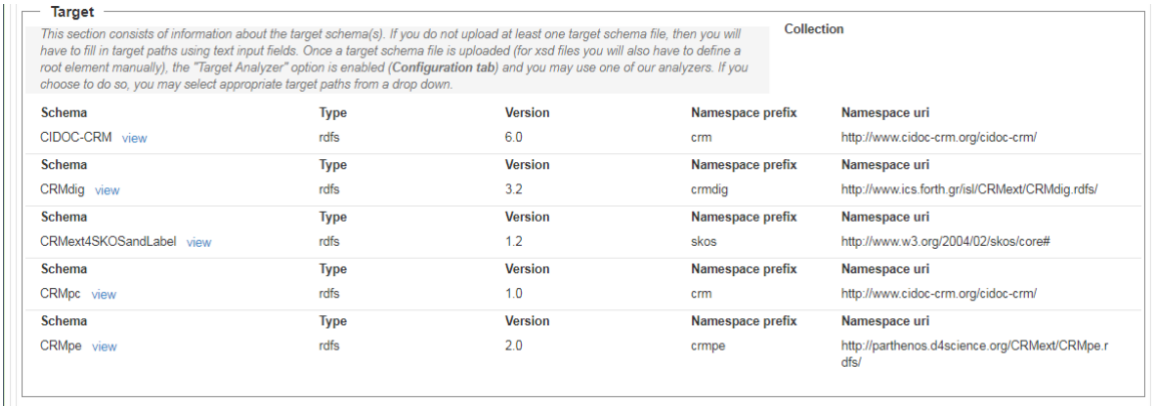


Figure 3. Info tab – General description, Source schema



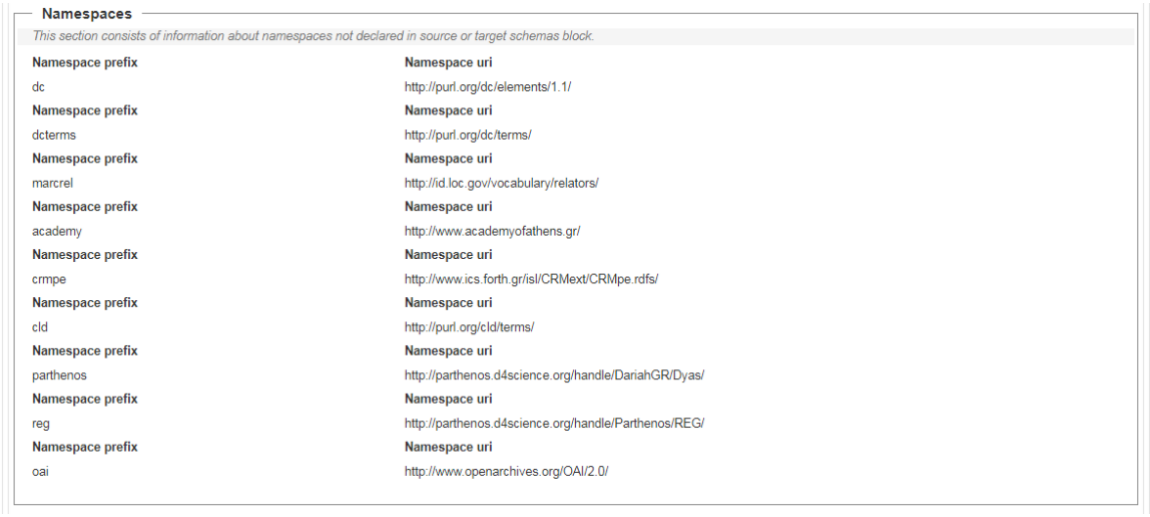
Target

*This section consists of information about the target schema(s). If you do not upload at least one target schema file, then you will have to fill in target paths using text input fields. Once a target schema file is uploaded (for xsd files you will also have to define a root element manually), the "Target Analyzer" option is enabled (**Configuration tab**) and you may use one of our analyzers. If you choose to do so, you may select appropriate target paths from a drop down.*

Collection

Schema	Type	Version	Namespace prefix	Namespace uri
CIDOC-CRM view	rdfs	6.0	crm	http://www.cidoc-crm.org/cidoc-crm/
Schema	**Type**	**Version**	**Namespace prefix**	**Namespace uri**
CRMdig view	rdfs	3.2	crmdig	http://www.ics.forth.gr/isl/CRMext/CRMdig.rdfs/
Schema	**Type**	**Version**	**Namespace prefix**	**Namespace uri**
CRMext4SKOSandLabel view	rdfs	1.2	skos	http://www.w3.org/2004/02/skos/core#
Schema	**Type**	**Version**	**Namespace prefix**	**Namespace uri**
CRMpc view	rdfs	1.0	crm	http://www.cidoc-crm.org/cidoc-crm/
Schema	**Type**	**Version**	**Namespace prefix**	**Namespace uri**
CRMpe view	rdfs	2.0	crmpe	http://parthenos.d4science.org/CRMext/CRMpe.rdfs/

Figure 4. Target schemas

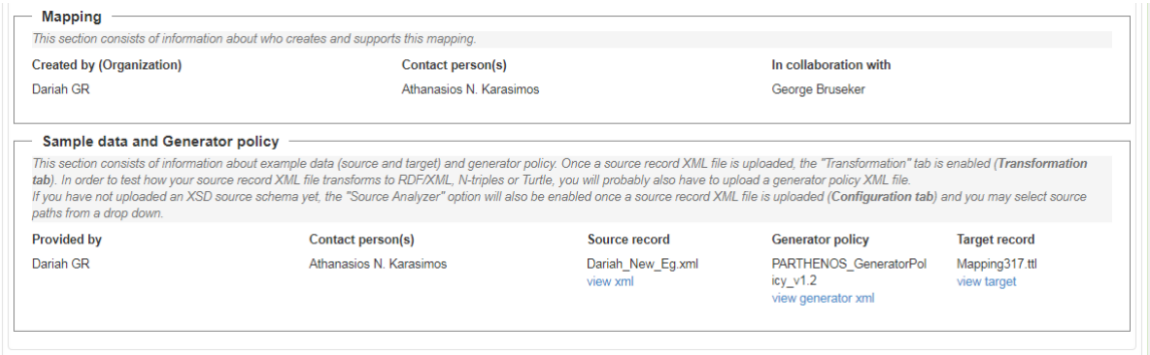


Namespaces

This section consists of information about namespaces not declared in source or target schemas block.

Namespace prefix	Namespace uri
dc	http://purl.org/dc/elements/1.1/
Namespace prefix	**Namespace uri**
dcterms	http://purl.org/dc/terms/
Namespace prefix	**Namespace uri**
marcrel	http://id.loc.gov/vocabulary/relators/
Namespace prefix	**Namespace uri**
academy	http://www.academyofathens.gr/
Namespace prefix	**Namespace uri**
crmpe	http://www.ics.forth.gr/isl/CRMext/CRMpe.rdfs/
Namespace prefix	**Namespace uri**
cld	http://purl.org/cld/terms/
Namespace prefix	**Namespace uri**
parthenos	http://parthenos.d4science.org/handle/DariahGR/Dyas/
Namespace prefix	**Namespace uri**
reg	http://parthenos.d4science.org/handle/Parthenos/REG/
Namespace prefix	**Namespace uri**
oai	http://www.openarchives.org/OAI/2.0/

Figure 5. Namespaces



Mapping

This section consists of information about who creates and supports this mapping.

Created by (Organization)	Contact person(s)	In collaboration with
Dariah GR	Athanasios N. Karasimos	George Bruseker

Sample data and Generator policy

*This section consists of information about example data (source and target) and generator policy. Once a source record XML file is uploaded, the "Transformation" tab is enabled (**Transformation tab**). In order to test how your source record XML file transforms to RDF/XML, N-triples or Turtle, you will probably also have to upload a generator policy XML file.*
*If you have not uploaded an XSD source schema yet, the "Source Analyzer" option will also be enabled once a source record XML file is uploaded (**Configuration tab**) and you may select source paths from a drop down.*

Provided by	Contact person(s)	Source record	Generator policy	Target record
Dariah GR	Athanasios N. Karasimos	Dariah_New_Eg.xml view xml	PARTHENOS_GeneratorPolicy_v1.2 view generator xml	Mapping317.ttl view target

Figure 6. Provenance info, sample data, generator policy

Matching Table: Contains the actual mappings. Mappings are presented to the user in a tabular format. Since most users are accustomed to maintaining mappings of their own in spreadsheets, this form of presentation is designed to be familiar to the user an easy to read. Each map is represented as a table (see Figure 7). The header of the table represents the domain of the mapping, and the rows represent the links. Since each link

contains two elements, one path and one range, the rows are double in size and contain both these elements. The columns of the table are used to separate the expression in the source schema (i.e., the provider's schema), from the expression in the target schema (i.e., the aggregator's schema), as well as conditional expressions or comments (for humans).

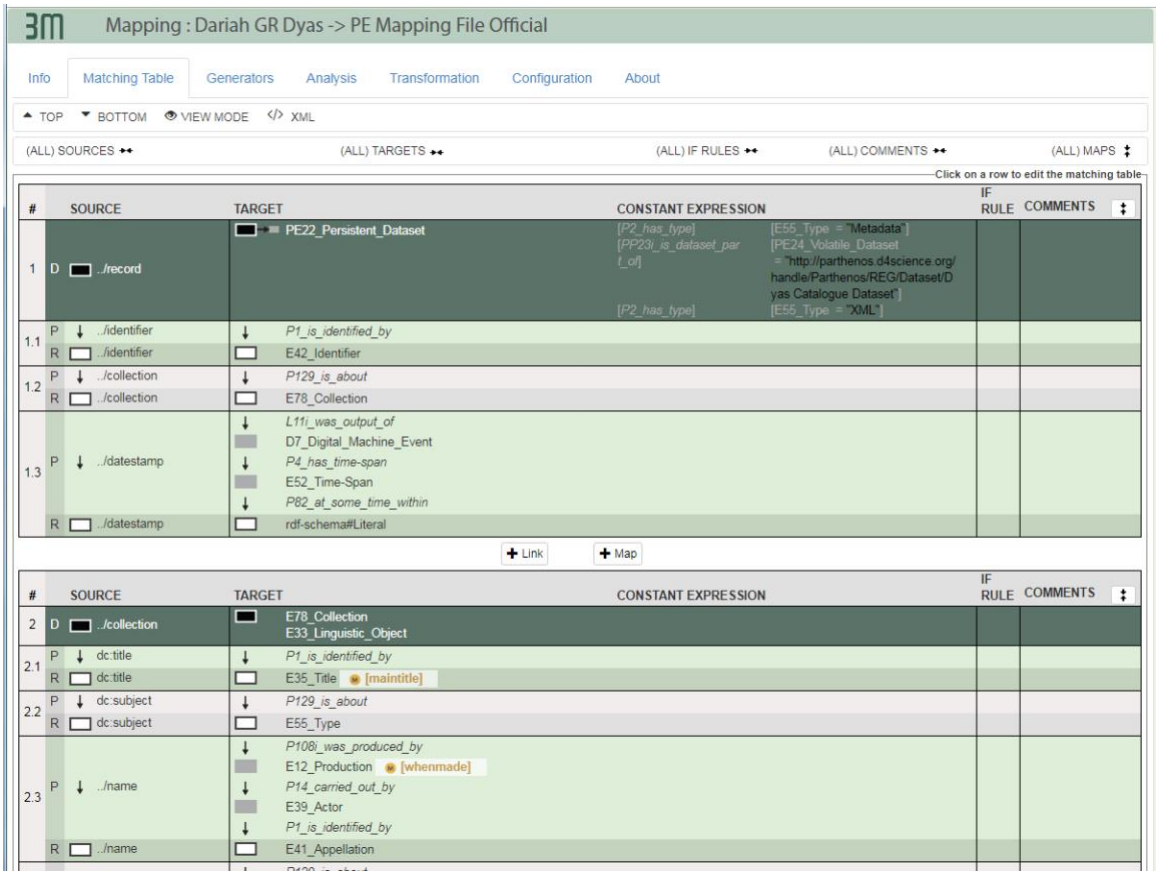


Figure 7. Matching table

The mapping process is facilitated with the Source (Figure 8) and Target (Figure 9) Schema Visualizers, two components responsible for assisting users in selecting the appropriate source and target paths for the definition of the mapping, minimizing typing errors and inconsistencies in the target.

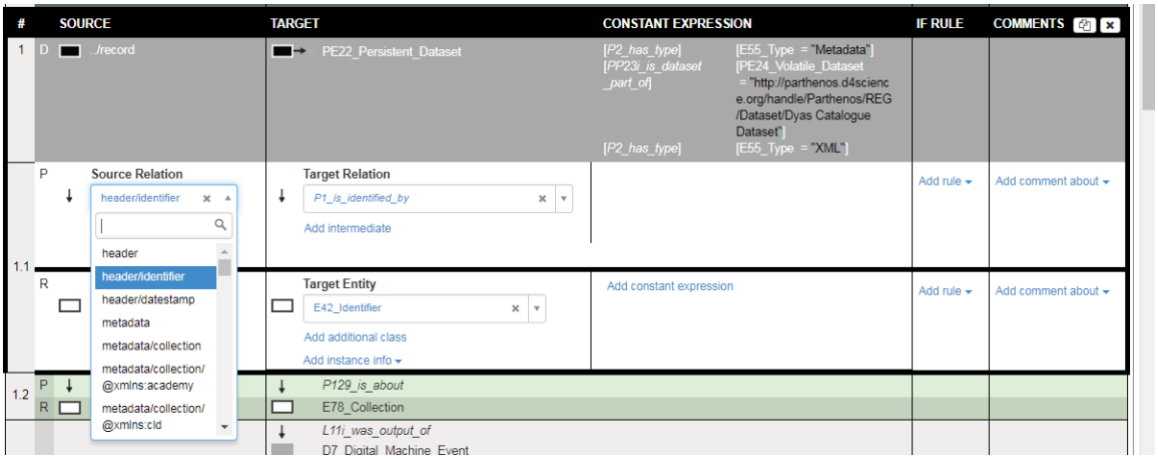


Figure 8. Source Schema Visualizer

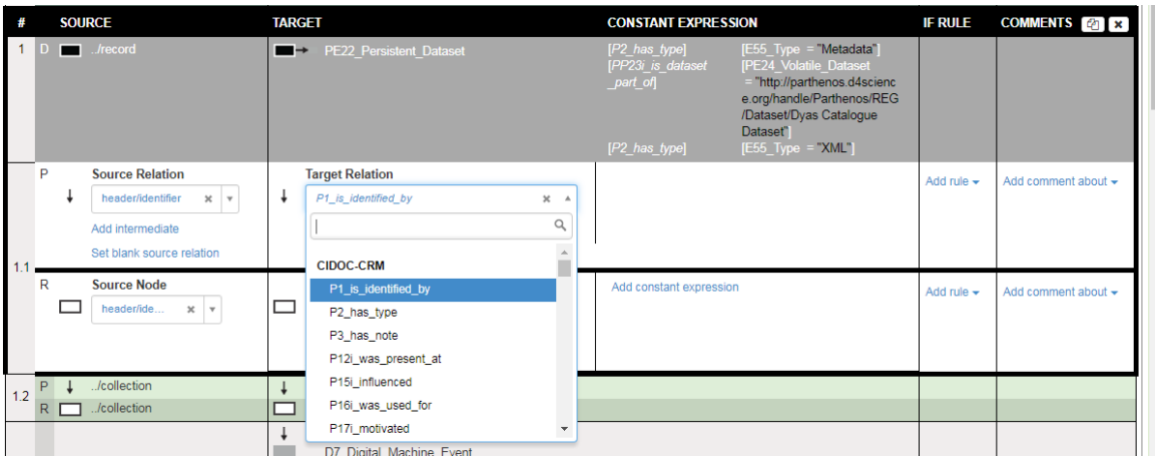


Figure 9. Target Schema Visualizer

Generators: Contains the specification of the instance generation rules which follows the specification of the Schema Matching Definition. The user interface is similar to the Schema Matcher component. However, users can only edit details about the instance and value generators (Figure 10).

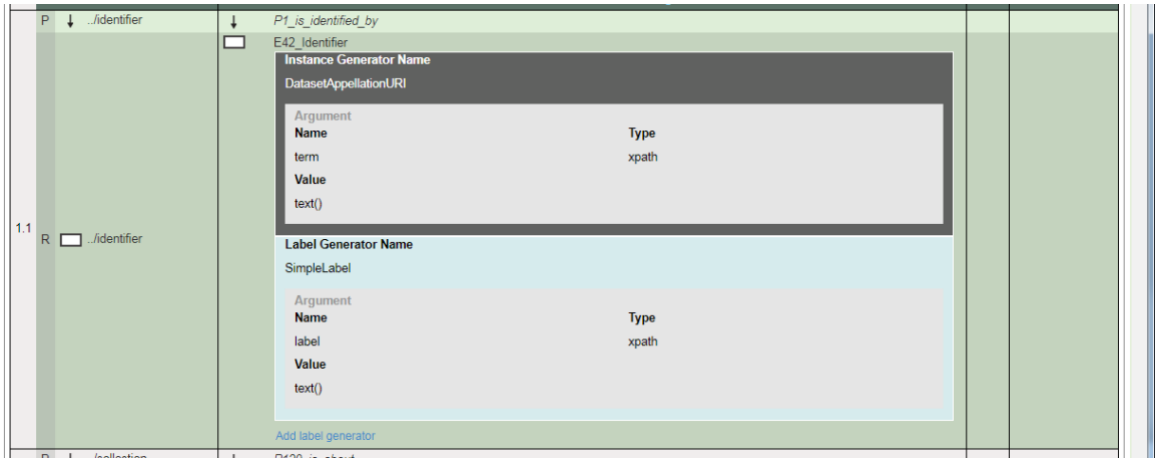


Figure 10. Instance and Label generators

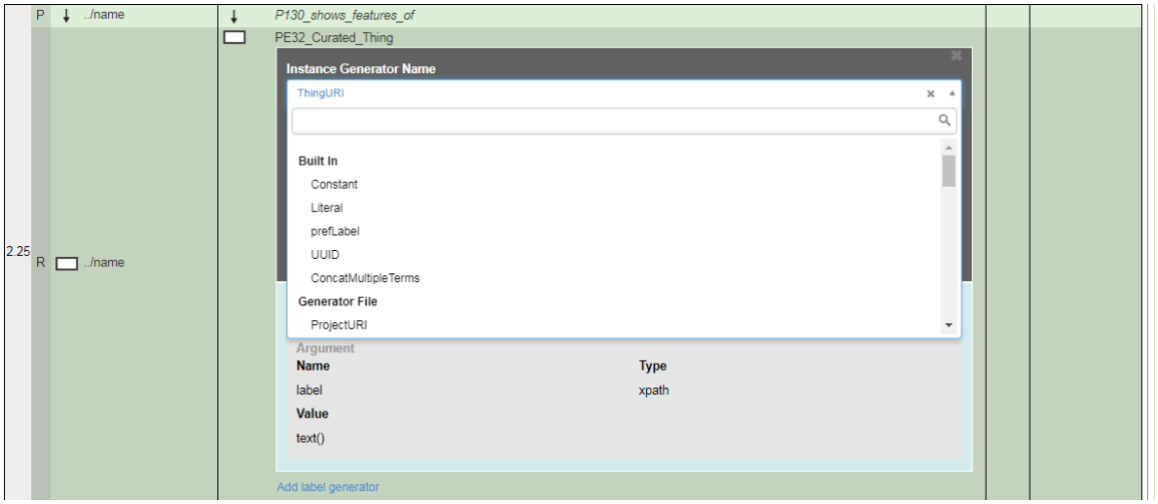


Figure 11. Available generators

The component uses generator functions for specifying how a URI or a label will be created and they can be exploited throughout the mapping. The generator functions are templates. These templates are defined in the generator policy file, which is designed and edited in a simple XML file outside the 3M system. This file is uploaded in the info tab and linked to the particular mapping. The system also supports built in and custom-made generators (Figure 11). Each target entity must have only one instance generator and any number of label generators.

Transformation: In this tab, the user can execute the X3ML engine which takes the sample source records, applies the X3ML mapping definition and produces the sample target records. Thus, it allows the user to inspect how source records will be transformed with respect to the defined mappings.

4.2 Application of mappings in the PARTHENOS aggregator

The PARTHENOS aggregative data infrastructure has been realised with the D-Net Software Toolkit. The D-NET Software Toolkit (D-NET for brevity) [6] proposes a service-oriented framework specifically designed to support developers at constructing custom aggregative infrastructures in a cost-effective way. D-NET offers *data management services* capable of providing access to different kinds of external data sources, storing and processing information objects of any data models, converting them into common formats, and exposing information objects to third-party applications through a number of standard access APIs. D-NET data management services address all types of interoperability issues described in Section 2:

Mediation and encoding interoperability: D-NET features several built-in plugins for the collection of metadata records (in XML, CSV, TSV and similar) via different exchange protocols (e.g. OAI-PMH, SFTP, FTP(S), HTTP(S), local file system). Additional plug-ins can be easily implemented and integrated if needed.

Granularity interoperability of metadata: D-NET features a Packaging Service to merge entities from different metadata records into one and an Unpackaging Service to split entities from one single input record to several metadata records.

Structural and semantic interoperability of metadata: D-NET features a Transformation Service capable of transforming metadata records from one format to another by applying mappings of different kinds (XSLTs, Groovy scripts, Java code, and D-NET transformation rules[2]). For semantic interoperability at the level of values, D-NET offers the Metadata Cleaner Service. The service harmonises values in metadata records based on a set of thesauri. A D-NET thesaurus consists of a *vocabulary* that is a list of authoritative *terms* together with associations between terms and their *synonyms*. Given a metadata format, the metadata cleaner service can be configured to associate the metadata fields to specific vocabularies. The service, provided records conforming to the metadata format, processes the records to clean field values according to the given associations between

[2] D-NET Transformation Rules are expressed in a textual language created by the development team from University of Bielefeld.

fields and vocabularies. Specifically, field values are replaced by a vocabulary term only if the value falls in the synonym list for the term. If no match is found, the field is marked as 'invalid'.

D-NET offers *infrastructure enabling services* that facilitate the construction of domain-specific aggregative infrastructures by selecting and configuring the needed services and easily combining them to form autonomic data processing workflows. The combination of out-of-the box data management services and tools for assembling them into workflows makes the toolkit an appealing starting platform for developers having to face the realisation of aggregative infrastructures. For a more detailed description of D-NET and its features, please refer to D6.1, Section 2.5 and [6].

For the PARTHENOS project, D-NET has been extended to support mappings expressed in X3ML language by integrating the X3ML Engine. The X3ML Engine is a Java library and, in order for the integration to be possible, FORTH and CNR-ISTI collaborated to update the library with new Java methods and to consolidate the code by upgrading some old library dependencies. As of December 2018, D-NET integrated version 1.9.0 of the x3ml-engine library available in the FORTH ISL Maven Repository[3].

Figure 12 shows the D-NET web user interface used by aggregator managers to configure an aggregation workflow.

[3] FORTH ISL Maven Repository: http://www.ics.forth.gr/isl/maven

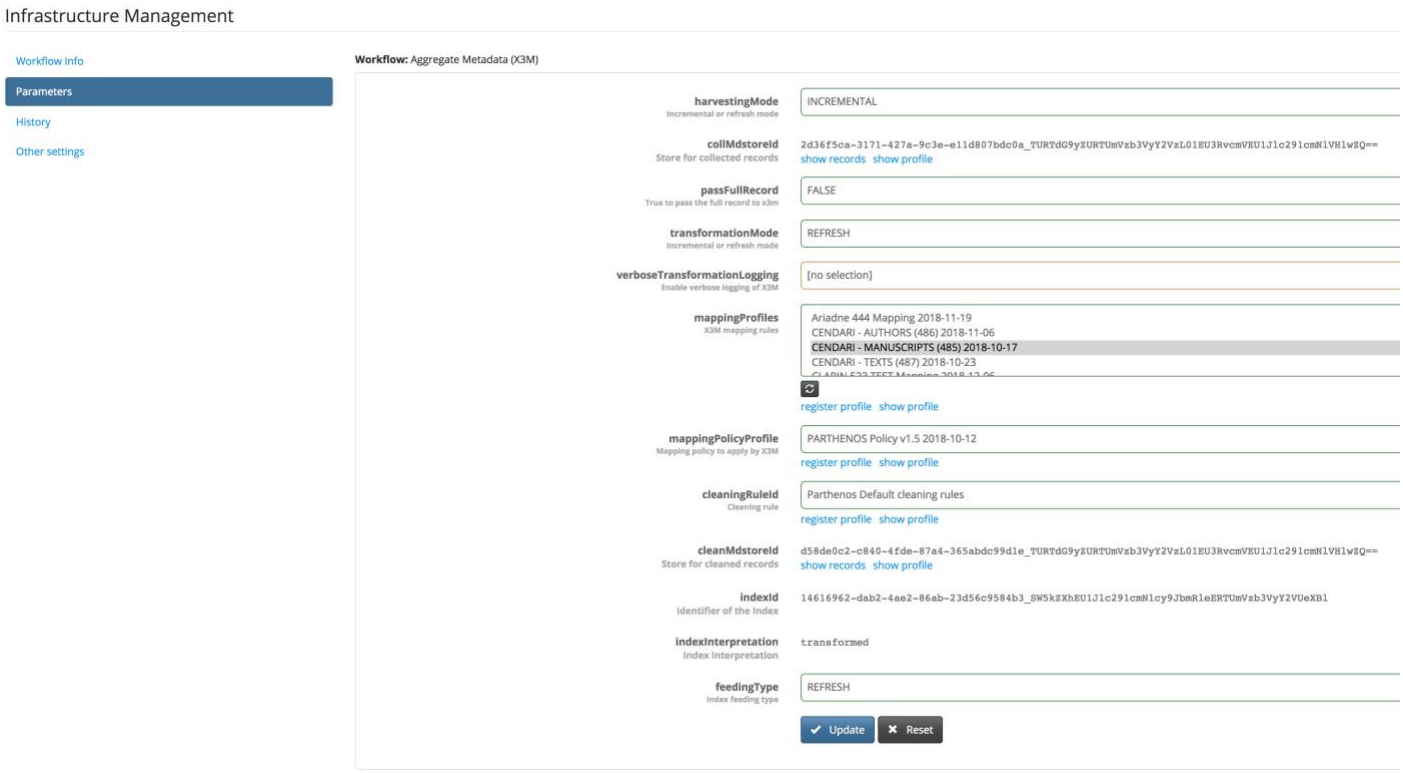


Figure 12. Setting parameters for a D-NET aggregation workflow: overview

Figure 13 shows details about the parameters to be set for the configuration of the transformation step:

- *passFullRecord*: Boolean. The PARTHENOS aggregator always adds, or updates if they exist, an OAI header and about sections to the collected records. All collected records have therefore the form of an OAI record[4]:

  ```
  <oai:record>
      <oai:header>…</oai:header>
      <oai:metadata>…</oai:metadata>
      <oai:about>…</oai:about>
  <oai:record>
  ```

  where the actual descriptive metadata of the resource is inside the `<oai:metadata>` element. Some mappings have been defined to work on the whole OAI record, others only on the descriptive metadata. When the parameter *passFullRecord* is set true, the aggregator is instructed to pass to the X3ML engine the whole OAI record. When set to false, the content of

[4] Namespace definition omitted for brevity and formatting.

<oai:metadata> is extracted and passed to the X3ML engine for transformation.

- *transformationMode*: REFRESH or INCREMENTAL. Aggregator managers can decide if all the collected records must be transformed (e.g. when the mapping is updated all records must be re-transformed) or if only the new/updated records must be transformed (applies only to RIs that are capable to expose their metadata in incremental mode).
- *verboseTransformationLogging*: Boolean. X3ML engine can be configured to verbosely log details about the ongoing transformation.
- *mappingProfiles*: list of mappings. Aggregator managers can select one or more X3ML mappings from the list. The list is populated based on the mappings currently registered in the PARTHENOS aggregator. In Figure 13, X3ML engine is configured to apply one mapping (the one in grey) to the input metadata records. Multiple mappings can be selected if needed. The links below the list box allows aggregator managers to register new mappings (“register profile”) or view the selected mappings (“show profile”).
- *mappingPolicyProfile*: mapping policy. Aggregator managers can select one URI generation policy file to pass to the X3ML Engine.
- *cleaningRuleId*: D-NET cleaning rule. X3ML mappings do not address semantic interoperability at the level of values for all fields of the PARTHENOS Entities model. With this parameter aggregator managers instruct the D-NET Metadata Cleaner service about the thesauri to apply to clean values of transformed metadata records.

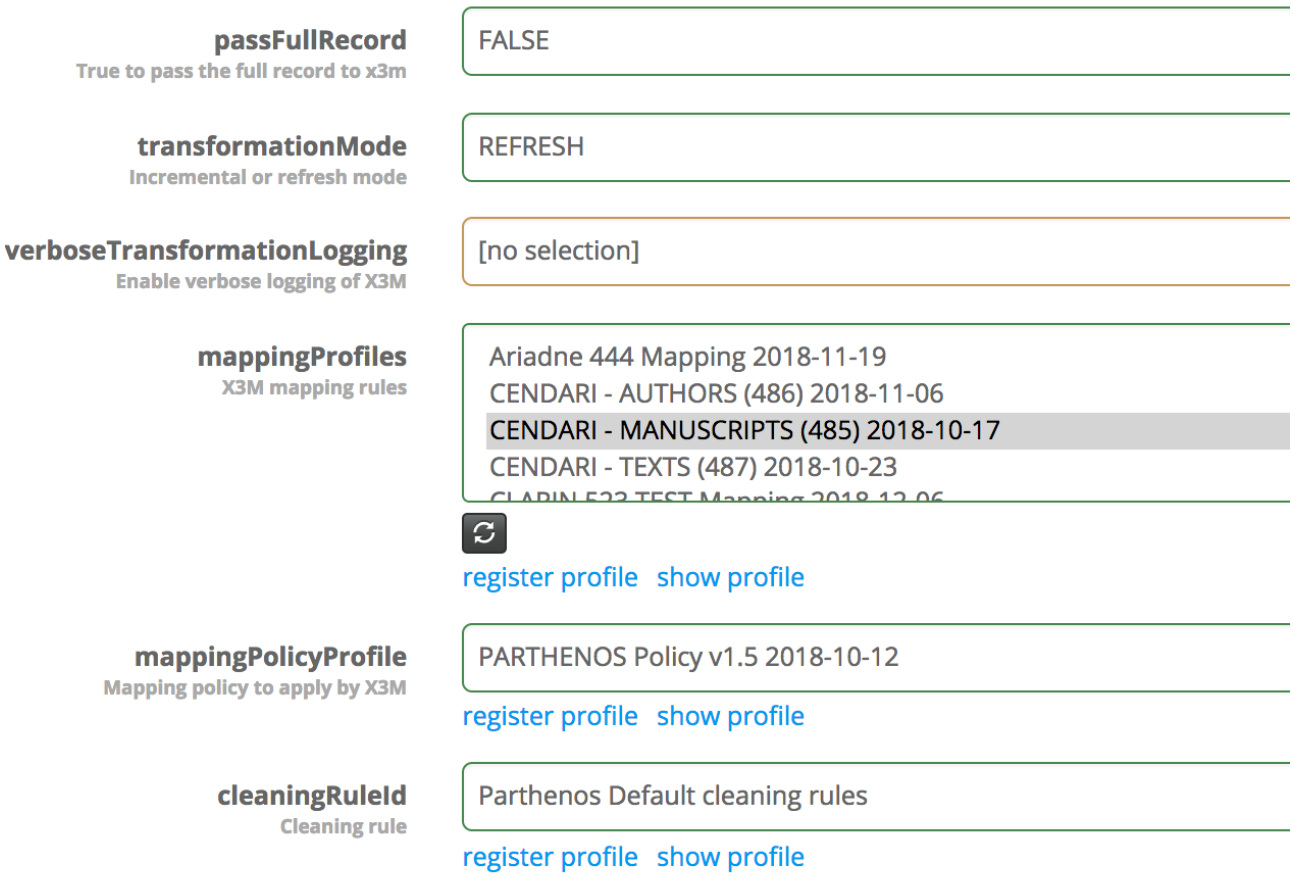


Figure 13. Transformation parameters

5. Mappings

5.1 Overview

The X3ML toolkit is deployed in the PARTHENOS infrastructure and integrated in the PARTHENOS Virtual Research Environment. As of December 2018, the PARTHENOS consortium created twenty-eight official mappings (out of a total of one hundred and fifty-two mappings, mostly created for testing, sampling and getting familiar with the 3M Editor). Twelve RI users have collaborated to their creation with the support of FORTH. Table 1 lists the twenty-eight official mappings. Details on each mapping is provided in the following sub-sections.

Table 1. PARTHENOS official mappings (December 2018)

Provider	Resource Mapped	Mapping Name	Mapping Responsible	Mapping ID
ARIADNE	ACDM:MASTER	ACDM -> PE Mapping File Official	Ilenia Gallucio; Achille Felicetti	444
CENDARI	CENDARI Manuscripts	CENDARI Manuscripts > PE Mapping File Official	Maurizio Sanesi	485
CENDARI	CENDARI Authors	CENDARI Authors > PE Mapping File Official	Maurizio Sanesi	486
CENDARI	CENDARI Texts	CENDARI Texts -> PE Mapping File Official	Maurizio Sanesi	487
CLARIN	CMDI: Datasets	CLARIN CMDI Dataset Mapping File Official	Matteo Lorenzini; Matej Durco; Davor Ostojic	548
CLARIN	CMDI: Services	CLARIN CMDI Service Mapping File Official	Matteo Lorenzini; Matej Durco; Davor Ostojic	373
Cultura Italia	Portal	CulturaItalia MUSEID-Italia -> PE Mapping File Official	Tiziana Scarselli; Sara De Giorgio	312
Cultura Italia	Actors	CulturaItalia People -> PE Mapping File Official	Tiziana Scarselli; Sara De Giorgio	416
DARIAH DE	DARIAH DE Portal	DARIAH DE -> PE Mapping File Official	Matteo Lorenzini	417
DARIAH GR	Dyas Register	DARIAH GR Dyas -> PE Mapping File Official	Athanasios Karasimos	515
DARIAH IT	Projects	DARIAH IT - Projects -> PE Mapping File Official	Maurizio Sanesi	453
DARIAH IT	People	DARIAH IT - People -> PE Mapping File Official	Maurizio Sanesi	452
DARIAH IT	Contributions	DARIAH IT - Contributions -> PE Mapping File Official	Maurizio Sanesi	451
DARIAH IT	Partners	DARIAH IT - Partners -> PE Mapping File Official	Maurizio Sanesi	450
EHRI	EHRI Model	EHRI -> PE Mapping File Official	Charles Riondet	328
Huma-Num	Nakala – Collection Level	Nakala -> PE Mapping File Official	Hélène Gautier; Nicolas Larrousse	516
Huma-Num	Nakala - Item Level	Nakala - Item Level Karnak	Hélène Gautier;	520

	Consortium 3D	-> PE Mapping File Official	Nicolas Larrousse	
Huma-Num	Nakala - Item Level KARNAK	Nakala - Item Level KARNAK -> PE Mapping File Official	Hélène Gautier; Nicolas Larrousse	517
Huma-Num	Nakala - Item Level MOM	Nakala - Item Level MOM -> PE Mapping File Official	Hélène Gautier; Nicolas Larrousse	521
Huma-Num	Nakala - Item Level AOROC	Nakala - Item Level AOROC -> PE Mapping File Official	Hélène Gautier; Nicolas Larrousse	522
Huma-Num	Isidore	Isidore -> PE Mapping File Official	Hélène Gautier; Nicolas Larrousse	432
Huma-Num	Isidore Item Level	Isidore Item Level -> PE Mapping File Official	Hélène Gautier; Nicolas Larrousse	398
LRE	LRE	LRE -> PE Mapping File Official	Fahad Khan	447
Metashare	Metashare	Metashare -> PE Mapping File Official	Fahad Khan	439
PARTHENOS	PARTHENOS Top Level	PARTHENOS Register Top Level -> PE Mapping File Official	George Bruseker	467
PARTHENOS WP3	Policy Wizard	WP3 Policy Wizard -> PE Mapping File Official	George Bruseker, Vyacheslav Tykohonov, Hella Hollander	335
PARTHENOS WP4	Standards DB	WP4 Standards List -> PE Mapping File Officia	Maurizio Sanesi	449
PARTHENOS WP5	WP5 PARTHENOS Standard Reference Resources	WP5 PARTHENOS Standard Reference Resources -> PE Mapping File Official	George Bruseker	547
PARTHENOS WP8	International Contacts	WP8 International Contacts -> PE Mapping File Official	George Bruseker, Sheena Basset	464

5.2 Mappings by content providers

In the following sections, mapping created for datasets of each RIs are described. For each mapping the following information are given:

Mapping information

- *ID*

  Identifier of the mapping assigned by 3M Editor.
- *URL to 3M Editor*

  URL to view the mapping in 3M Editor (requires authentication via the PARTHENOS Virtual Research Environment).
- *Status*

  Status of the mapping (e.g. completed, to be finalized, tested, under testing).

- *OAI header*

  Tells if the OAI header is available in the metadata records to map and if the mapping uses it to generate provenance information.
- *Issues*

  Problems encountered in creating or testing the mapping, if any.
- Main PARTHENOS Entities Covered

  This describes the main entities that were mapped to from PARTHENOS Entities or CIDOC-CRM more generally. The function of this is to enable the researcher to understand the main data represented in this dataset according to the PARTHENOS Entities Model
- Base URI

  This gives the base namespace which has been used to generate the URIs from the mapping.
- *Generated PARTHENOS entities*

  Which PARTHENOS entities are considered, and therefore generated, by the mapping. Please note that mappings typically generate also other entities that are defined in other namespaces (e.g. CRM[5], CRMdig[6], CRMsci[7], FRBR[8], SKOS[9]). Here we list only the entities that are defined in the PARTHENOS namespace (CRMpe[10]).

5.2.1 PARTHENOS Top Level Entities

Top-level entities are a special data source for the aggregator, as it is a single XML file containing metadata descriptions about entities that cannot be automatically retrieved from any endpoint of any RIs. The XML file has been prepared by FORTH with the collaboration of all RIs in the consortium. The file can be found at: http://data.d4science.org/em1EemhBdUZ0bjNGTWJNNjlxVDltcm9acDFmMHlBSVVHbWJQNStIS0N6Yz0.

[5] CIDOC-CRM namespace: http://www.cidoc-crm.org/cidoc-crm/
[6] CRMdig namespace: http://www.ics.forth.gr/isl/CRMext/CRMdig.rdfs/
[7] CRMsci namespace: http://www.ics.forth.gr/isl/CRMext/CRMsci.rdfs/
[8] FRBR namespace: http://www.cidoc-crm.org/frbr/
[9] SKOS namespace: http://www.w3.org/2004/02/skos/core
[10] CRMpe namespace: http://parthenos.d4science.org/CRMext/CRMpe.rdfs/

Mapping information

- *ID*: 467
- *URL to 3M Editor*: https://mapping-d-PARTHENOS.d4science.org/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping419
- *Status*: completed and tested
- *OAI header*: OAI header not available
- *Issues*: the testing phase revealed some issues related to the UTF-8 encoding which had been fixed.
- Main PARTHENOS Entities Covered:
  - PE26_RI_Project
  - PE17_Curated_Data_E-Service
  - PE7_Data_Hosting_Service
  - PE15_Data_E-Service
  - PE13_Software_Computing_E-Service
  - PE24_Volatile_Dataset
  - PE25_RI_Consortium
  - E40_Legal_body
- Base URI: http://parthenos.d4science.org/handle/Parthenos/REG/
- Generated PARTHENOS entities:
  - PE1_Service
  - PE7_Data_Hosting_Service
  - PE13_Software_Computing_E-Service
  - PE15_Data_E-Service
  - PE17_Curated_Data_E-Service
  - PE25_RI_Consortium
  - PE26_RI_Project
  - PE28_Curation_Plan
  - PE29_AccessPoint
  - PE34_Team
  - PE36_Competency_Type
  - PE37_ProtocolType
  - PE_38_Schema

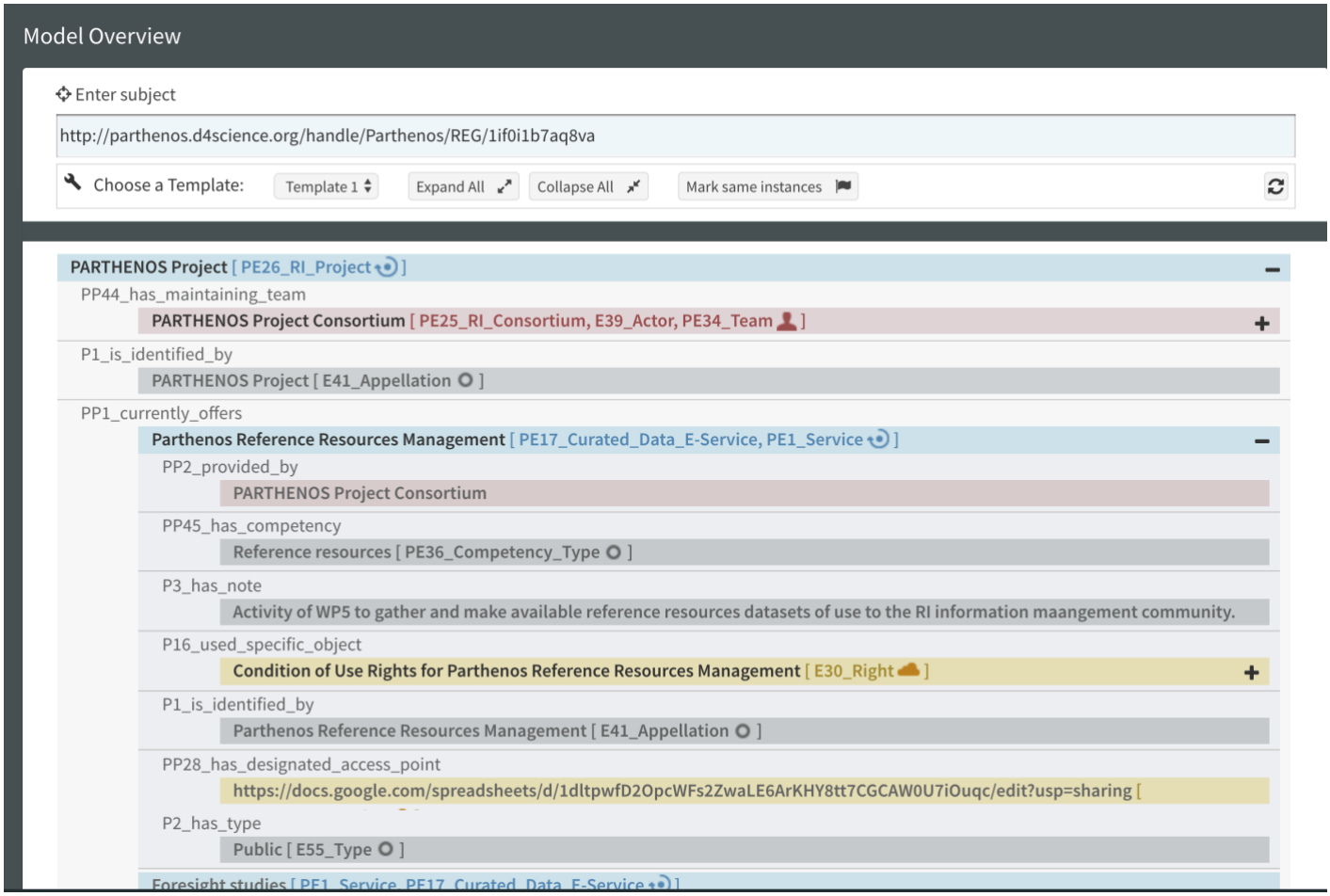


Figure 14 Example of mapped entity of the PARTHENOS Top Level

5.2.2 ARIADNE

The ARIADNE project aims to integrate the archaeological resources made available by the partners of the project for the purposes of discovery, access and integration on a research infrastructure. These resources include data, services and language resources, such as metadata formats, vocabularies and mappings. The registry is addressed to cultural institutions, private or public, which wish to describe their assets in order to make them known to e-infrastructures. The registry data model, called ACDM (ARIADNE Catalogue Data Model), extends the DCAT vocabulary[11] and describes the available resources among the various partners of the project. ACDM is a very rich and powerful model and it has been decided to map only a subpart of it during the first mapping phase. The result is one single mapping that covers ACDM datasets and basic descriptive information (contributors, rights and licenses, spatial and temporal coverage, and provenance).

[11] Data Catalogue Vocabulary (DCAT) https://www.w3.org/TR/vocab-dcat/

Mapping information

- *ID*: 444
- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping444
- *Status*: mapping completed, under testing.
  - *OAI header*: not available
  - *Issues*: The mapping cannot be thoroughly tested before metadata record aggregation because of data sparsity in individual records. Solution discussed will be to produce sample records for each of the twenty relevant parts of the mapping to make sure that the mapping works for that part. FORTH will then test to make sure all is fine and then final aggregation can be done by CNR.
  - Main PARTHENOS Entities Covered;
    - PE22_Persistent_Dataset
    - E39_Actor
    - E53_Place
  - Base URI: http://parthenos.d4science.org/handle/Ariadne/AriadnePortal/
  - Generated PARTHENOS entities:
    - PE15_Data_E-Service
    - PE22_Persistent_Dataset
    - PE29_Access_Point
    - PE38_Schema

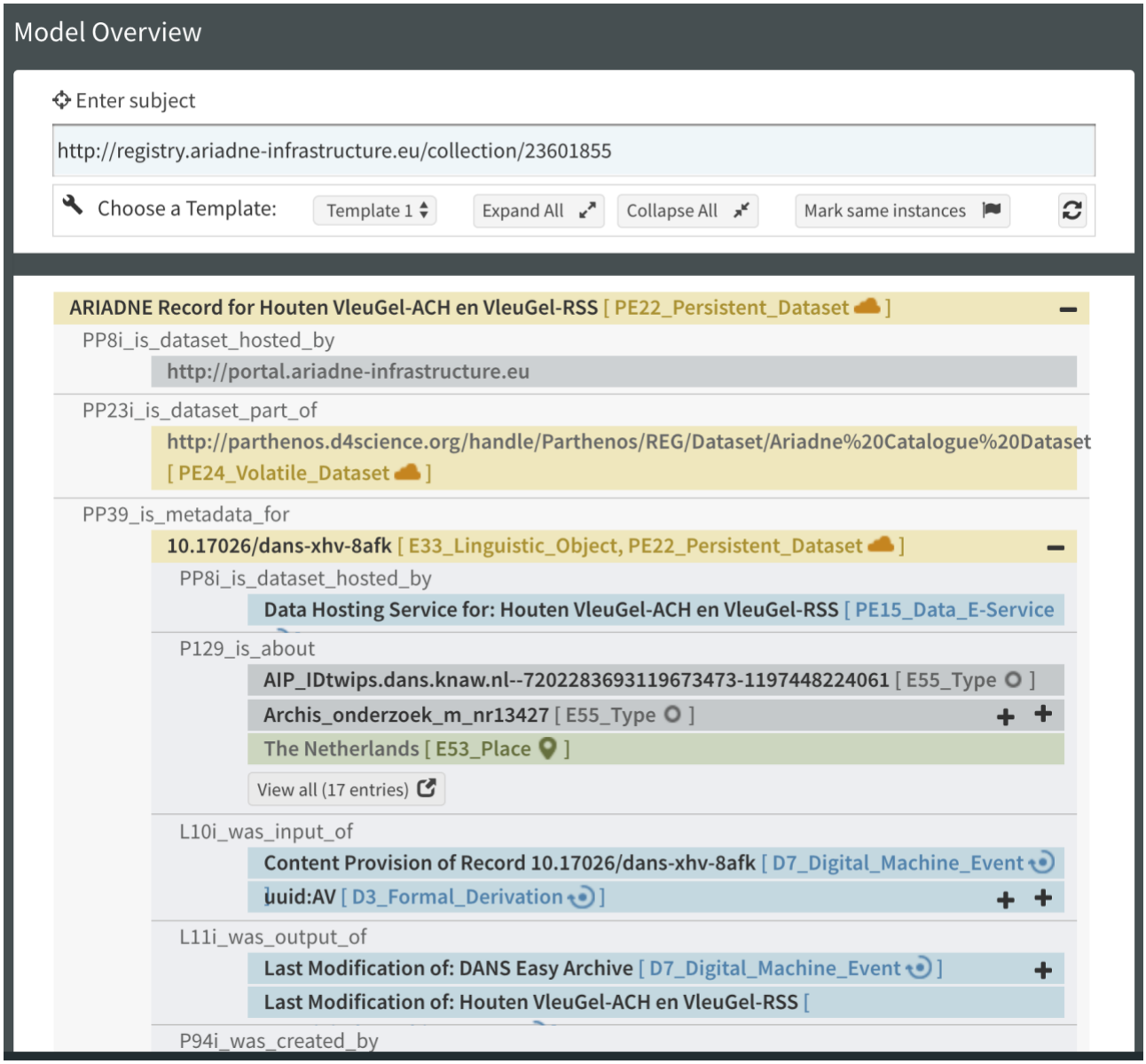


Figure 15 Example of mapped entity of Ariadne

5.2.3 CENDARI

The CENDARI dataset is covered by three mappings from research data created in the CENDARI project with regards to medieval manuscripts. These mappings cover authors, texts (in the abstract sense) and manuscripts (as physical objects. The CENDARI mappings offer an example of the use of the PARTHENOS infrastructure to create a rich mapping of primary research data from a particular dataset into the CIDOC CRM ontology.

Mapping information

- *ID*: 485, 486, 487

- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&id=Mapping485&lang=en
- Status: Complete
- OAI header: N/A
- Issues: Needs review from content experts in future.
- Base URI: http://parthenos.d4science.org/handle/Cendari/CendariDB/
- Main PARTHENOS Entities Covered:
    - E22_Man-Made_Object
    - E21_Person
    - E73_Information_Object

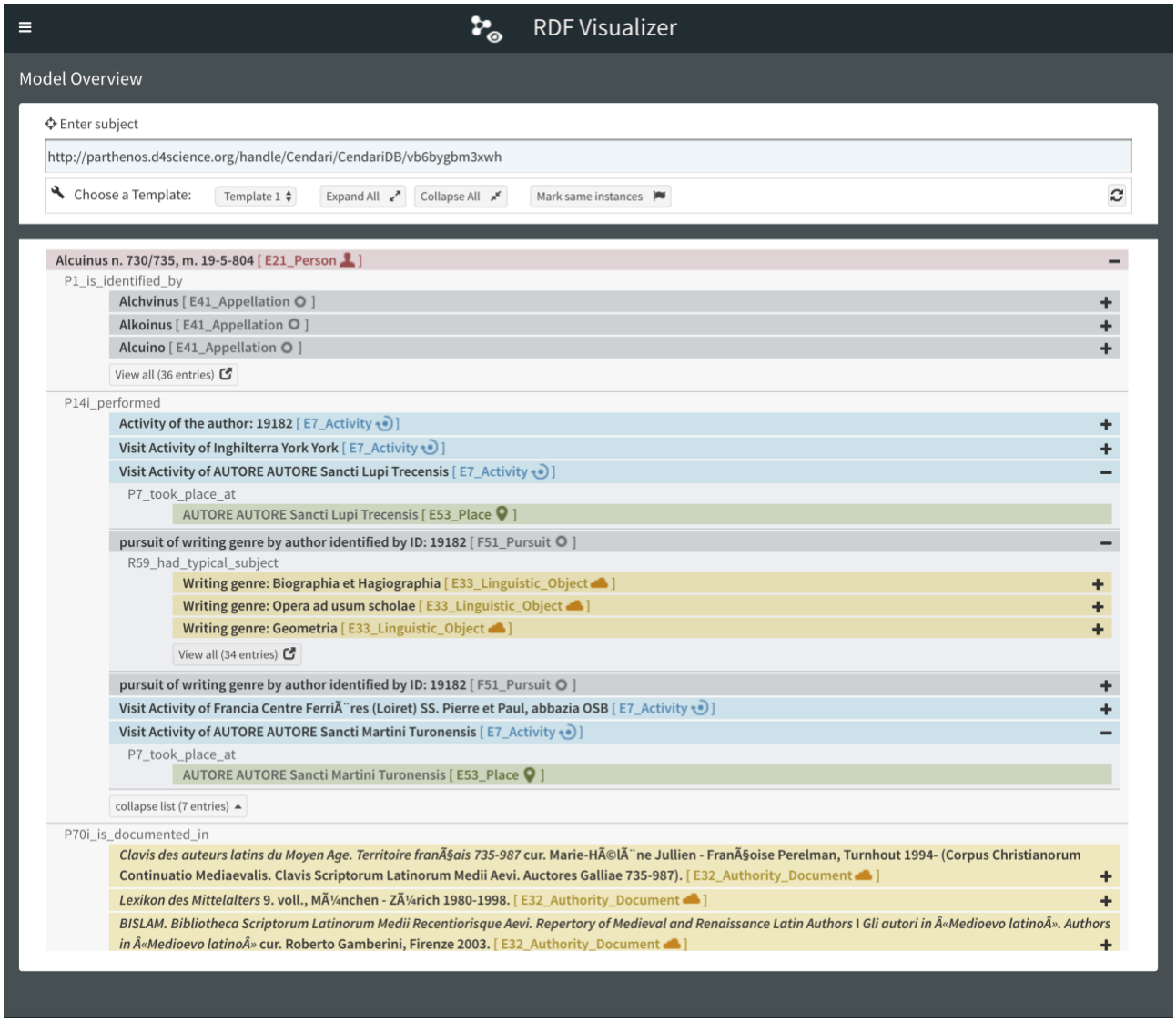


Figure 16 Example of mapped entity of CENDARI

5.2.4 CLARIN

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 [9][11]. It aggregates the metadata about the resources from over sixty data providers, containing more than 900,000 records. The backbone of the VLO is CMDI (Component Metadata Infrastructure) [8][9], which offers a flexible standardised framework to facilitate formalized descriptions for a wide range of resources, contributing to resource discovery within the linguistic domain. In order to deliver the information about CLARIN resources to PARTHENOS, it is required to map CMD metadata schemas to PARTHENOS Entities (PE). This reports about the approach adopted for the mapping between CMDI and PE model.

CMDI Model

The Component Metadata Infrastructure (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 new metadata creation. The CMDI Component Registry [7] was created as 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 VLO. Currently, it contains around one thousand components and around two hundred profiles. Fields in the components are linked to the concepts defined in the CLARIN Concept Registry (CCR) [10], successor of ISOcat data category registry which openly specifies stable definitions of semantic concepts, ensuring interoperability between the various profiles.

Mapping approach

The default approach to mapping is 1:1 cross-walks between the “local” source schema specific to individual research infrastructure and the target schema CIDOC-PE. However, CMDI is not just one schema but a framework for creating and reusing schemas. In fact, currently more than two hundred different schemas have been defined. It is, therefore, not feasible for the PARTHENOS project to define the mapping in a traditional way, i.e. as 1:1

cross-walks between source and target schema. 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 CMDI profiles to well-defined concepts. The developed mapping solution aims to identify PE properties which are (near) equivalent to the concepts of CCR (see Figure 14), to derive XPath patterns for any profile by matching concepts in the corresponding XML schema, and finally to use the XPaths to extract values from actual CMD instances (records) to generate a corresponding entity description adhering to PE model.

The generated mapping is converted to a format required by X3ML, pushing all processing logic to the PARTHENOS side. In order to automatize the whole process, it has been developed a simple java application 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 location to be resolved against a given individual CMD profile. The whole procedure is depicted in Figure 17.

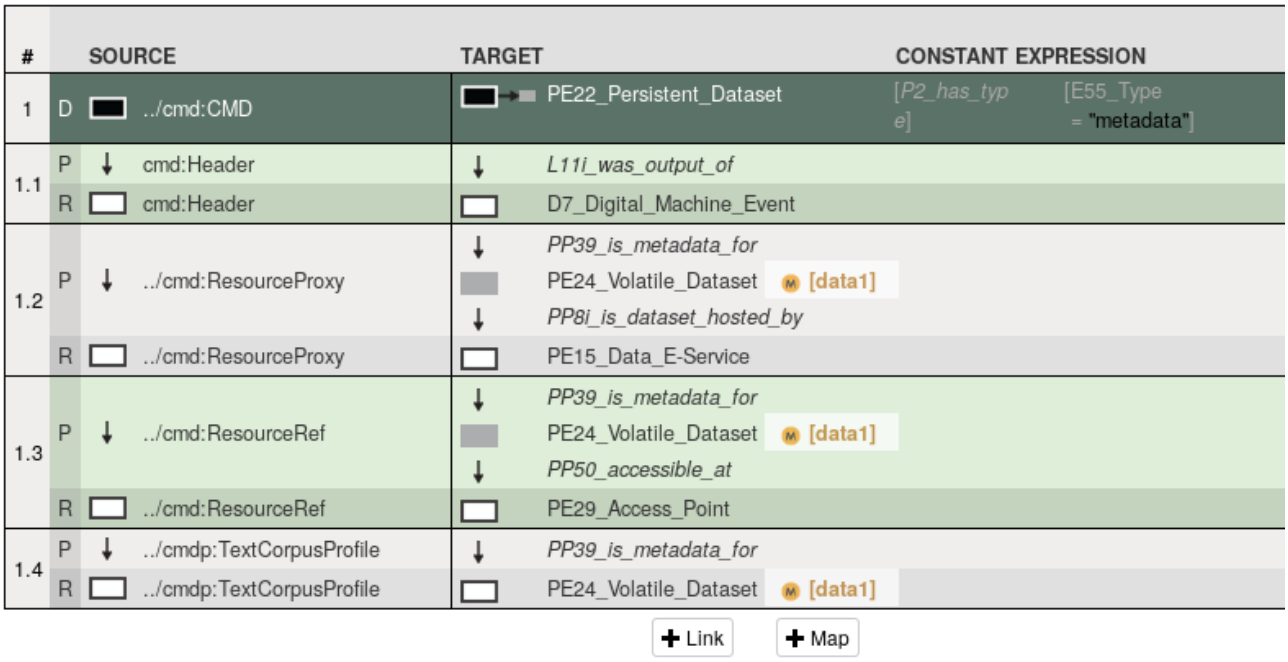


Figure 17. CLARIN Mapping Definition Phase

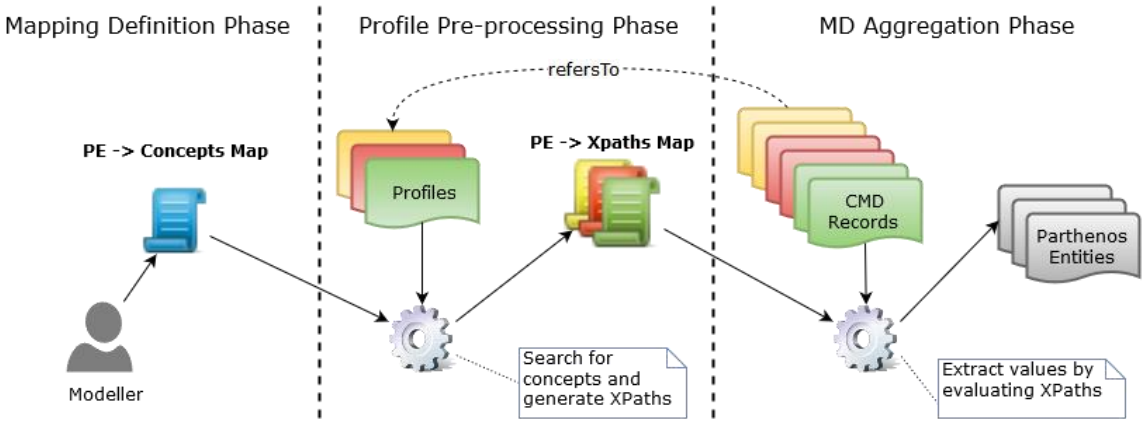


Figure 18. Automatic generation of X3ML mapping files

Mapping structure

Following the general model of CMD framework, we distinguish global mappings of the generic CMD envelope applicable to all CMD records (selected mapping examples in Table 2) and “local” mappings custom to individual CMD profiles (Table 3).

Table 2. CLARIN global mapping

CMDI XPath	CIDOC-PE	Note
/cmd:CMD	**crmpe:PE22_Persistent_Dataset**	Metadata record itself also is 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 → **rdf-schema#Literal**	
cmd:MdCreator	D7 → crmdig:L23_used_software_... → **crmpe:PE21_Persistent_Software**	
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 3. CLARIN: examples of local mappings

CMDI	CIDOC-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

Currently, three mappings have been implemented and tested, corresponding to *datasets* [12](ID:548) and *services* (ID:373). These stable mapping schemas represent the main structure of the template used by the semi-automatic mapping generator[13].

Mapping information (548 - datasets)

- *ID*: 548
- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping548
- *Status*: mapping completed and tested. Validated by FORTH. A valid RDF file is generated by 3M editor
- *OAI header*: OAI header not available
- *Base URI*: http://parthenos.d4science.org/handle/Clarin/VLO/
- *Issues*:
- *Main PARTHENOS Entities Covered*:
    - PE22 Persistent *Dataset*
    - D7 Digital Machine Event
    - PE15 Data E-Service
    - PE24 Volatile Dataset
    - PE35 Project
    - PE24 Team
    - E40 Legal Body
- *Generated PARTHENOS entities*:
    - PE15_Data_E-Service
    - PE21_Persistent_Software
    - PE24_Volatile_Dataset
    - PE29_Access_Point
    - PE34_Team
    - PE35_Project
    - PE38_Schema

[12] Corresponding to the majority of CLARIN records
[13] Cfr. Par. “Mapping Approach”

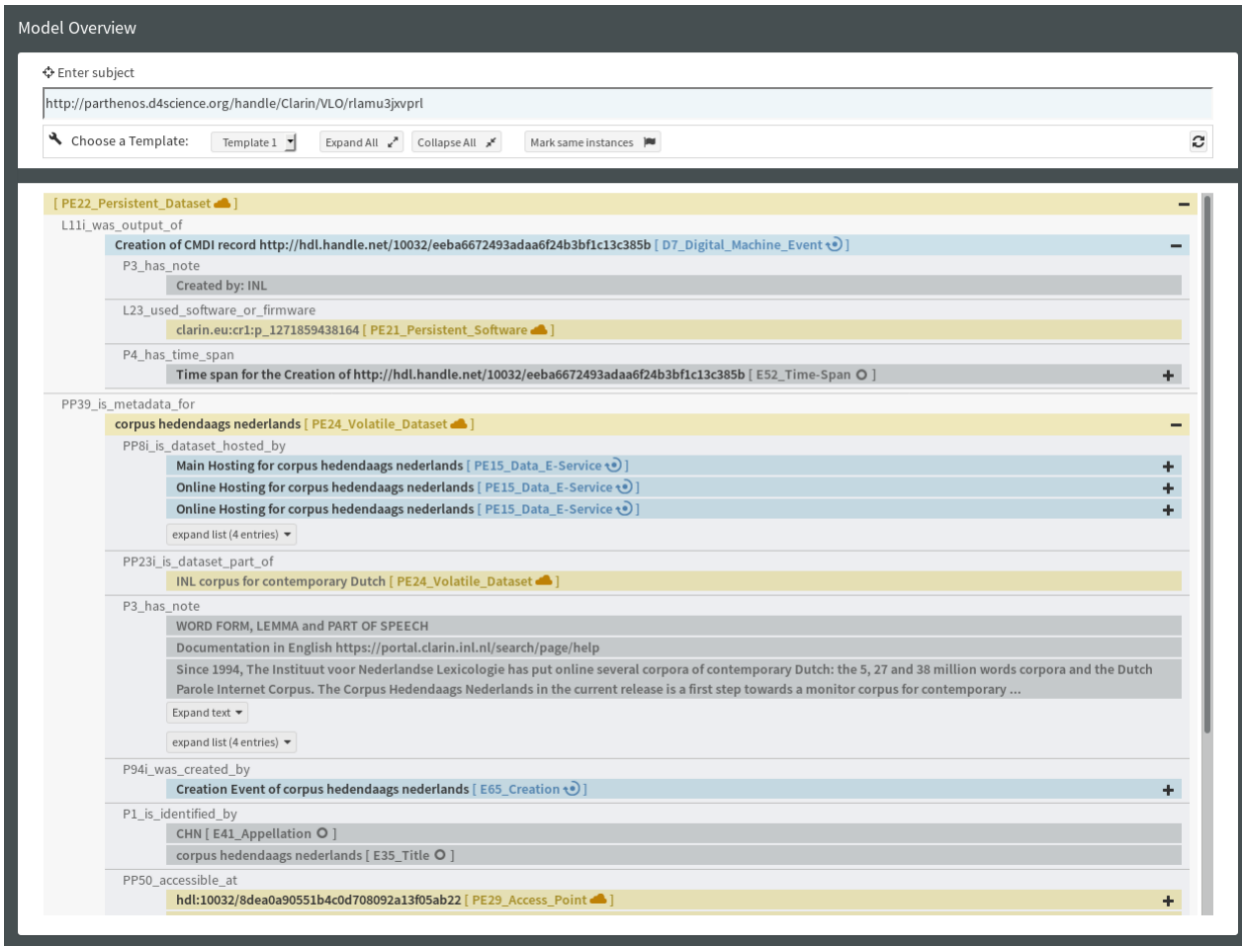


Figure 19 Example of mapped entity of CLARIN (dataset)

Mapping information (373 – services)

- *ID*: 373
  - *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping373
  - *Status*: mapping completed and tested. Validated by FORTH. A valid RDF file is generated by 3M editor
  - *OAI header*: OAI header not available
  - Issues:
    - Minor[14] issues about:
      - Standardisation of the encoding type.

[14] Solution already discussed

    - Major issues about:
        - OAI endpoints must be fixed with correct URIs.
- *Main PARTHENOS Entities Covered:*
    - PE22 Persistent Dataset
    - D7 Digital Machine Event
    - PE8 E Service
- *Base URI:* http://parthenos.d4science.org/handle/Clarin/VLO/
- Generated PARTHENOS entities:
    - PE1_Service
    - PE8_ E-Service
    - PE21_Persistent_Software
    - PE22_Persistent_Dataset
    - PE29_Access_Point
    - PE36_Competency_Type
    - PE38_Schema

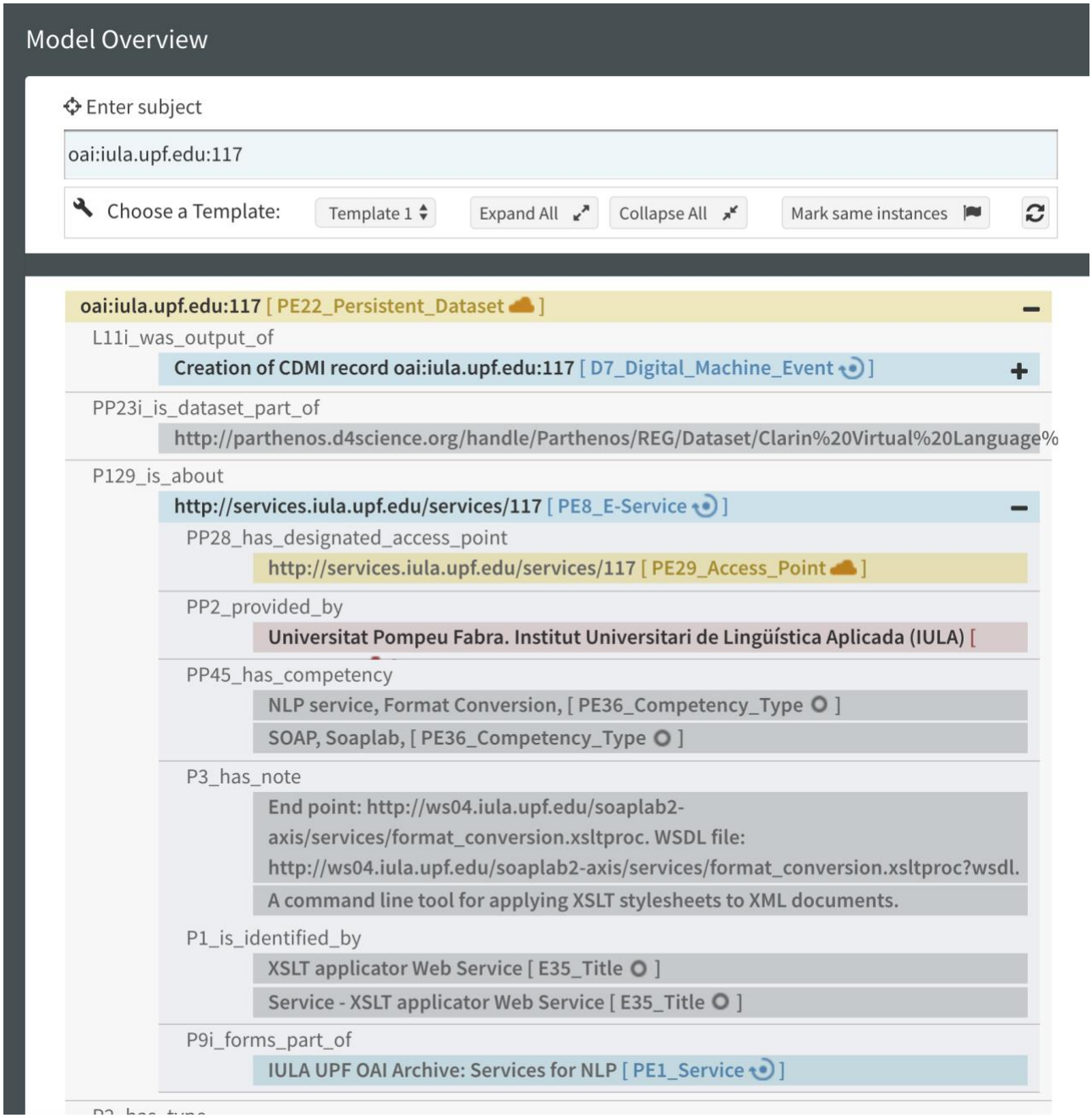


Figure 20 Example of mapped entity of CLARIN (service)

5.2.5 CulturaItalia

CulturaItalia is the Portal of Italian Culture, managed by the Central Institute for the Union Catalogue of Italian Libraries (ICCU) of Ministry of cultural heritage, activities and tourism (MiBACT). CulturaItalia, as national aggregator, plays an important role for the development of European RIs on Cultural Heritage such as ARIADNE, DARIAH and Europeana, making available cooperative networks and agreements and coordinating technical activities.

A specific Dublin Core Application Profile has been designed in order to cover the complex domain of the “Italian Culture” and to guarantee the interoperability of various kinds of cultural resources. This application profile is called PICO AP from the name of the Project in whose context the CulturaItalia portal was developed. The PICO AP combines in one metadata schema all DC Elements, all DC Element Refinements and Encoding Schemes from the Qualified DC and other refinements and encoding schemes specifically conceived to retrieve information pertaining to Italian culture.

PICO AP is mapped into PARTHENOS Entities Model via two X3ML mappings. Mapping 312 for the mapping of metadata about museum collections and mapping 416 for the mapping of metadata about Italian museums.

Mapping information (312 - collections)

- *ID*: 312
- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping312
- *Status*: completed and tested
- *OAI header*: OAI header available and mapped
- *Issues*: Each metadata describes a museum collection and lists the URI of the parts that form the collections, without specifying a name or title of the parts. The lack of titles does not affect the operation of mappings: with those titles, collections would be easier to discover in the PARTHENOS infrastructure.
- *Base URI*: http://parthenos.d4science.org/handle/Culturaitalia/ICCUMDI/
- Main PARTHENOS Entities Covered:
    - PE22_Persistent_Dataset
    - PE24_Volatile_Dataset
- Generated PARTHENOS entities:
    - PE22_Persistent_Dataset
    - PE24_Volatile_Dataset

Mapping information (416 - museums)

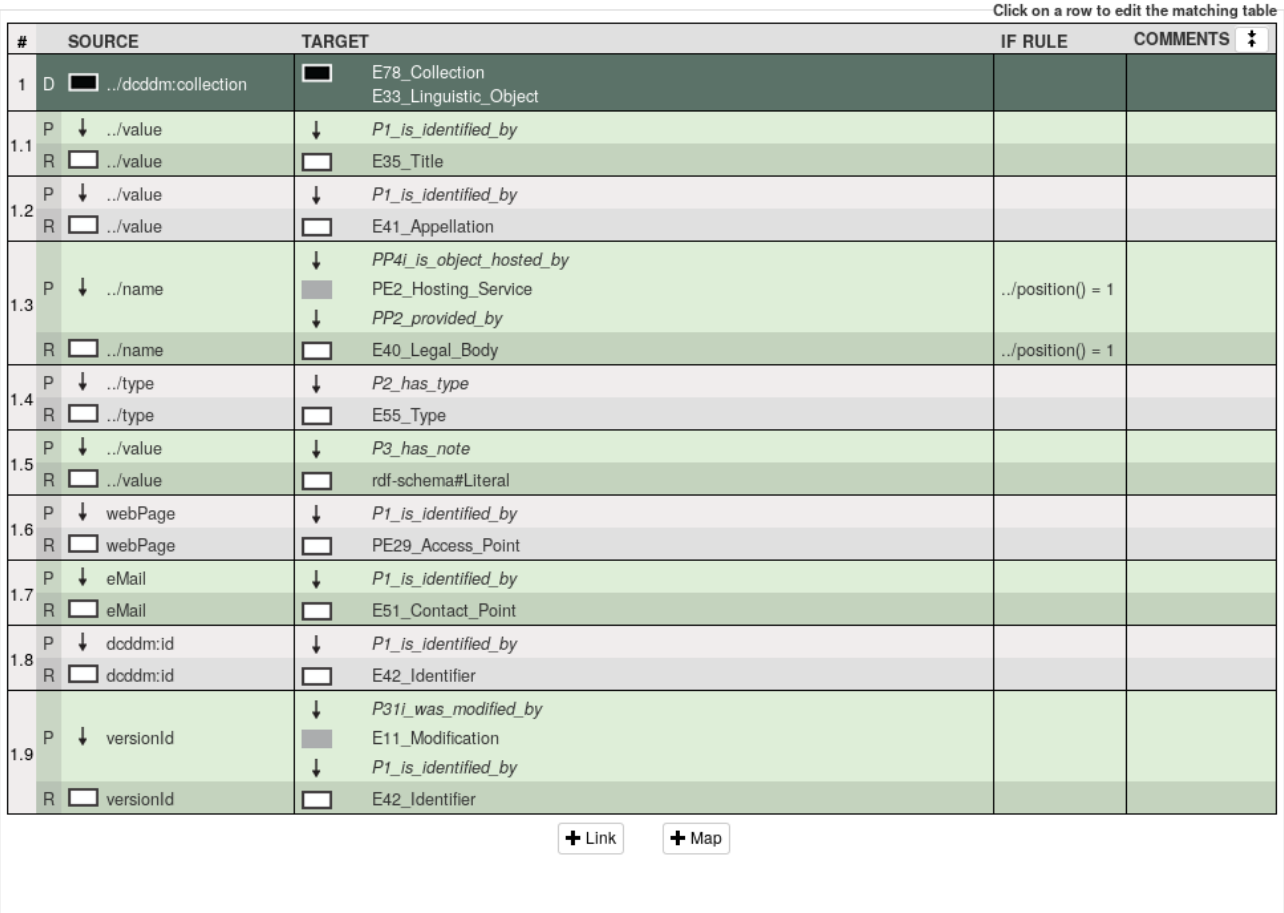
- *ID*: 416
- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping416
- *Status*: completed, under testing
- *OAI header*: OAI header available and mapped
- *Issues*: None.
- *Base URI*: http://parthenos.d4science.org/handle/Culturalitalia/ICCUCI/
- Main PARTHENOS Entities Covered:
  - PE22_Persistent_Dataset
  - E39_Actor
- Generated PARTHENOS entities:
  - PE22_Persistent_Dataset
  - PE1_Service

5.2.6 DARIAH DE

DARIAH-DE supports research in the humanities and cultural sciences with digital methods and procedures. The research infrastructure of DARIAH-DE consists of four pillars: teaching, research, research data and technical components. As a partner in DARIAH-EU, DARIAH-DE helps to bundle and network state-of-the-art activities of the digital humanities.

Mapping information (417 - collection)

- *ID*: 417
- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&id=Mapping417&lang=en
- *Status*: completed, under testing
- *OAI header*: OAI header available and mapped
- *Issues*: None.
- *Base URI*: http://pathenos.d4science.org/handle/DariahDE/
- Main PARTHENOS Entities Covered:
  - E78_Collection
  - E40_Legal_Body



Click on a row to edit the matching table

#		SOURCE		TARGET	IF RULE	COMMENTS
1	D	../dcddm:collection		E78_Collection E33_Linguistic_Object		
1.1	P	../value		*P1_is_identified_by*		
	R	../value		E35_Title		
1.2	P	../value		*P1_is_identified_by*		
	R	../value		E41_Appellation		
1.3	P	../name		*PP4i_is_object_hosted_by* PE2_Hosting_Service *PP2_provided_by*	../position() = 1	
	R	../name		E40_Legal_Body	../position() = 1	
1.4	P	../type		*P2_has_type*		
	R	../type		E55_Type		
1.5	P	../value		*P3_has_note*		
	R	../value		rdf-schema#Literal		
1.6	P	webPage		*P1_is_identified_by*		
	R	webPage		PE29_Access_Point		
1.7	P	eMail		*P1_is_identified_by*		
	R	eMail		E51_Contact_Point		
1.8	P	dcddm:id		*P1_is_identified_by*		
	R	dcddm:id		E42_Identifier		
1.9	P	versionId		*P31i_was_modified_by* E11_Modification *P1_is_identified_by*		
	R	versionId		E42_Identifier		

Figure 21 DARIAH-DE mapping definition on the 3M Editor

DARIAH members involved in task 5.2 of PARTHENOS initially focused on the mapping for metadata from the Greek registry (see Section 5.2.7). Mapping 515 generated for DARIAH GR/DYAS will be revised, if needed, to be applied also to the metadata available from the German registry of DARIAH.

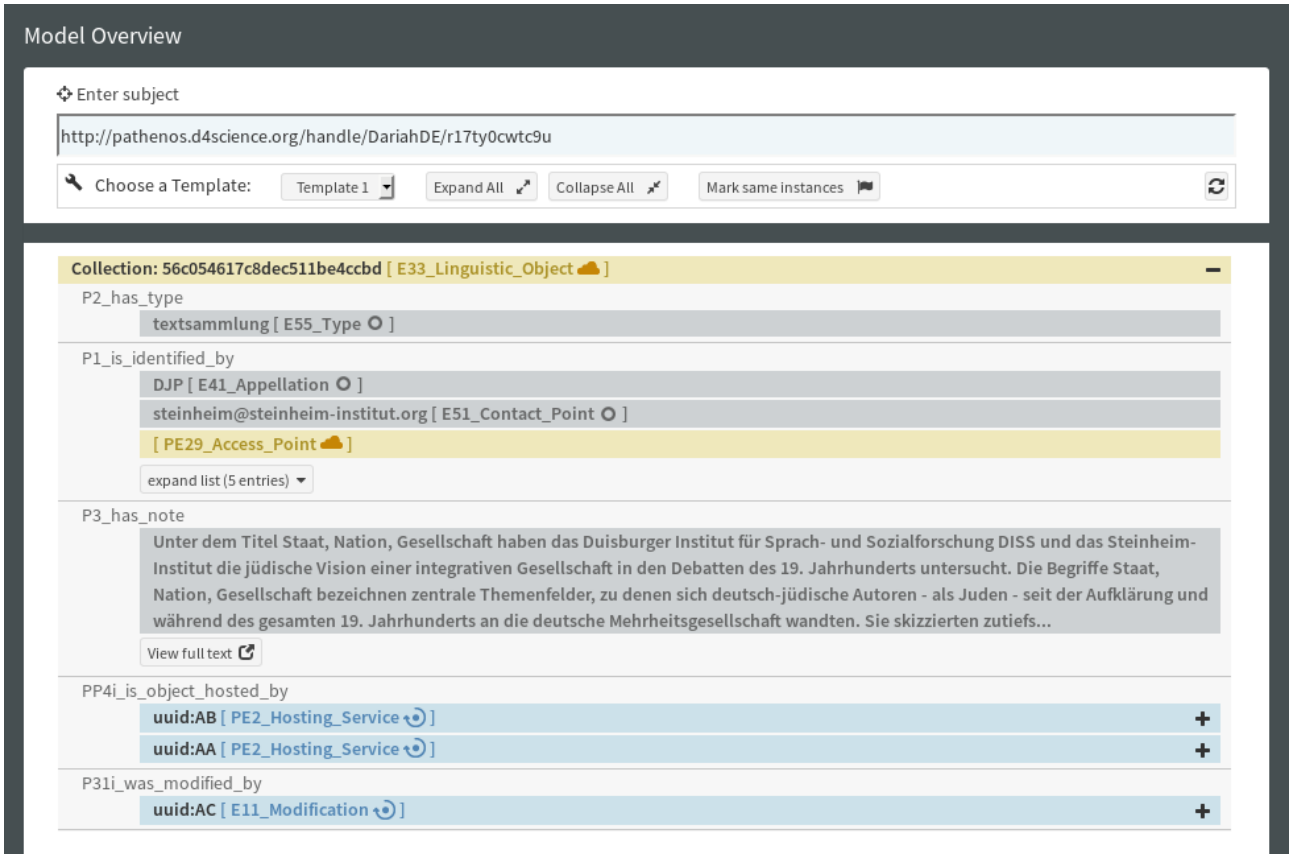


Figure 22 Example of mapped entity of DARIAH-DE

5.2.7 DARIAH GR/DYAS

The mission of DYAS, the Greek Research Infrastructure Network for the Humanities, is:

- to support the Greek communities of humanities researchers in advancing their work using ICT and in exchanging knowledge and working practices
- to broaden the scope of and opportunities for research through the interconnection of various distributed digital resources
- to promote the access, use, creation and long-term preservation of research data, both primary and secondary, in digital form.

DYAS is also in charge of operating the Greek component of the European Infrastructure for Arts and Humanities, DARIAH. DYAS is designed as a distributed infrastructure with members at distinct levels of involvement:

- management nodes, providing the services of the infrastructure and setting the specifications for digital resources,
- curators, responsible for specific collections and added-value repositories,
- affiliates, providing selected metadata for ingestion by the management nodes.

The DYAS Organizations and Collections Registries provide access information on Greek institutions, individuals and analogue and digital collections. It covers seventeen disciplines of Humanities and Arts and all the fields fall within the Greek History, Culture, Heritage and Language categories. The registry data model is based on FOAF for Persons and Organisations and Dublin Core Collections Application Profile (DCCAP) for Collections; it presents, with detailed information the available collections and organisations of Humanities and Arts.

Mapping information

- *ID*: 515
- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping515
- *Status*: completed, tested, possibly to be revised for additional harmonization.
- *OAI header*: OAI header available and mapped
- Issues:
    - Timespans in source records use the same field (dc:created) for both start and end date, making it programmatically difficult to distinguish what is what
    - Places, right types, and curation technique fields contain numeric values. Data expert from DARIAH GR must provide FORTH and CNR with an explanation of those values in order to set-up the proper harmonization process in the aggregation workflow.
- *Base URI*: http://parthenos.d4science.org/handle/DariahGR/Dyas/
- Main PARTHENOS Entities Covered:
    - PE22_Persistent_Dataset
    - E78_Collection
- Generated PARTHENOS entities:
    - PE2_Hosting_Service
    - PE22_Persistent_Dataset
    - PE32_Curated_Thing

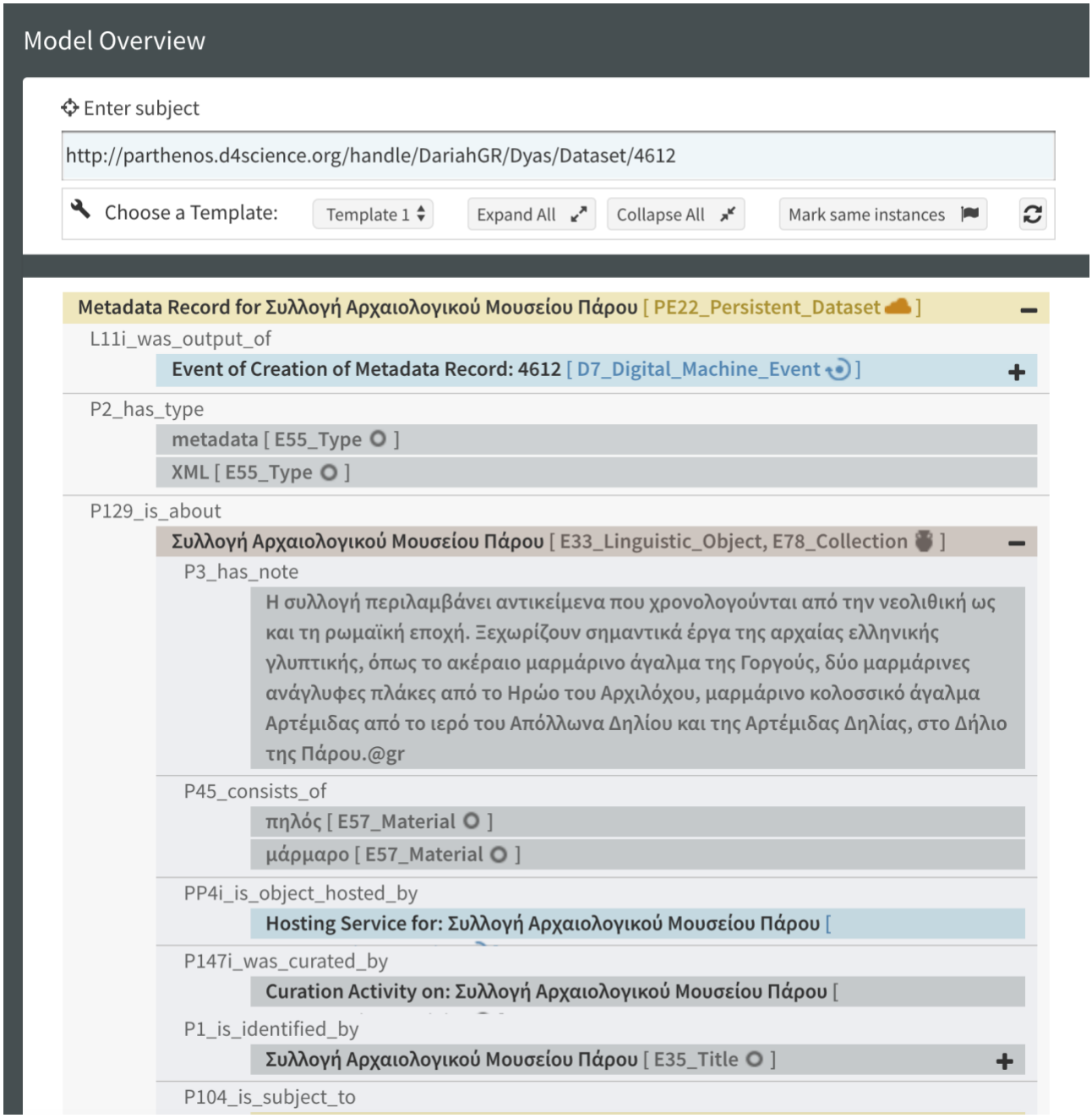


Figure 23 Example of mapped entity of DARIAH GR/DYAS

5.2.8 EHRI

EHRI offers an online environment to freely access to rich information about Holocaust-related archival institutions and their collections across Europe and beyond. It is trans-national in scope, containing information about Holocaust-related archival institutions in more than fifty countries. Metadata records from EHRI conform to the conceptual standards proposed by the International Council on Archives:

- ISAD(G): General International Standard Archival Descriptions
- ISDIAH: International Standard for Describing Institutions with Archival Holdings
- ISAAR(CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families.

And the related following encoding standards:

- EAD: Encoded Archival Description
- EAG: Encoded Archival Guide
- EAC(CPF): Encoded Archival Context (Corporate Bodies, Persons, Families).

Additionally, thesaurus data is encoded in a manner aligned with the Simple Knowledge Organisation System (SKOS). The EHRI catalogue currently contains more than 150,000 descriptions of archival materials, 474 descriptions of archival institutions that hold archival materials and authority files on 3,231 Corporate Bodies and 620 Personalities related to the history of the Holocaust.

At this stage of the project, EHRI members created an X3ML mapping from EAD to PE Model. Additional mappings from EAG and EAC could be provided in the next months to further enrich the PARTHENOS Joint Resource Registry.

Mapping information

- *ID*: 328
- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping328
- *Status*: completed and tested.
- *OAI header*: OAI header not available
- Issues: Source records contain a declaration of namespace (*<ead xmlns="urn:isbn:1-931666-22-9"*) that is not processable by X3ML Engine. CNR therefore implemented a work-around on the aggregation side, removing the namespace declaration at metadata record harvesting time.
- Base URI: http://parthenos.d4science.org/handle/EHRI/PORTAL/

- Main PARTHENOS Entities Covered:
  - PE22_Persistent_Dataset
  - E78_Collection
  - E39_Actor
- Generated PARTHENOS entities:
  - PE2_Hosting_Service
  - PE22_Persistent_Dataset

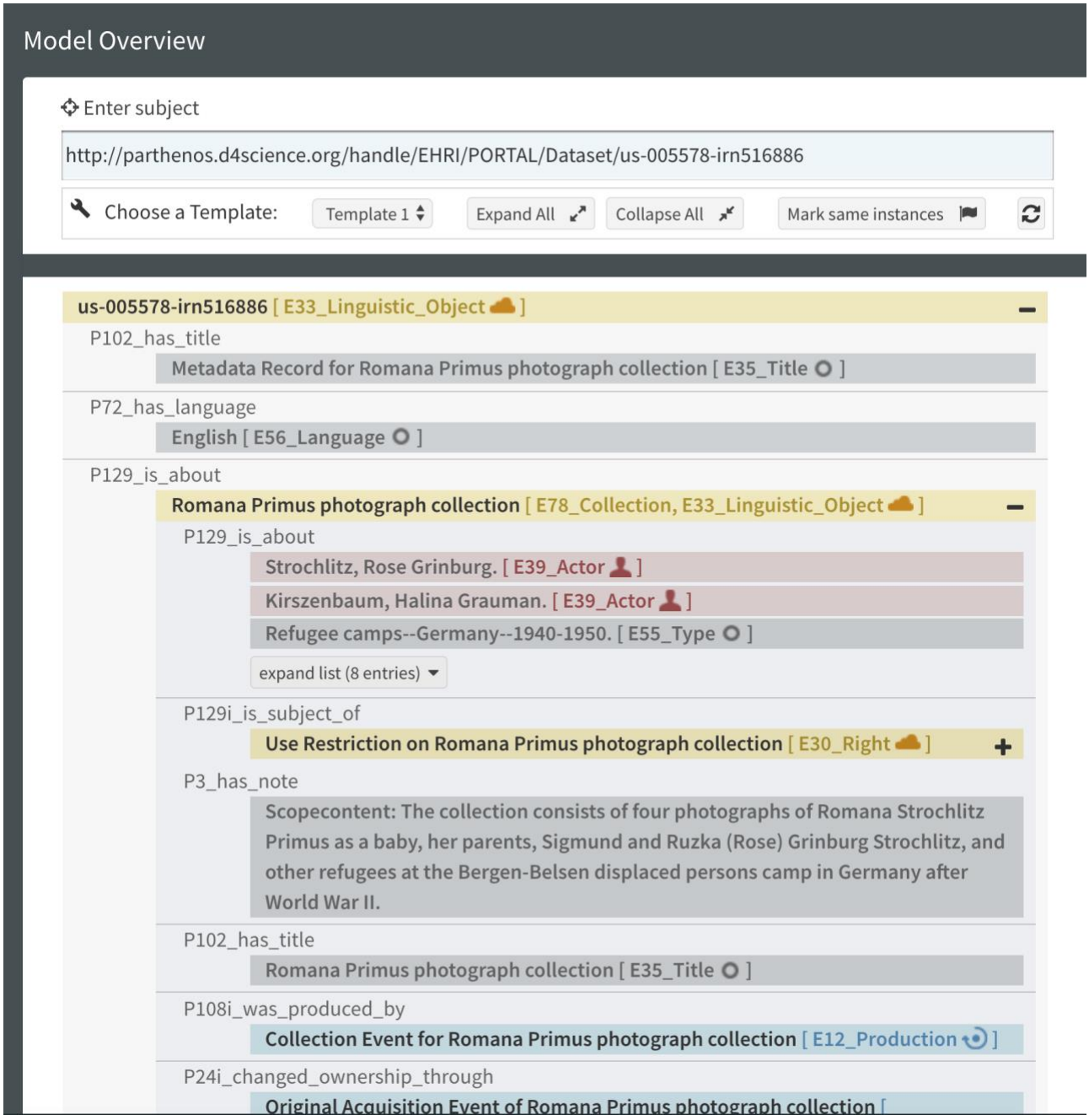


Figure 24 Example of mapped entity of EHRI

5.2.9 Huma-Num

Huma-Num is a major French research infrastructure aimed at facilitating the turning of digital research in the humanities and social sciences. Huma-Num offers services dedicated to the production and reuse of scientific data. To do this, Huma-Num supports research teams throughout their digital projects to allow the sharing, reuse and preservation of data thanks to a chain of devices focused on interoperability. The aim is to foster the exchange and dissemination of metadata, but also of data itself via standardized tools and lasting, open formats.

For the PARTHENOS project, Huma-Num provides metadata records describing the content hosted in two of its main services: Isidore and Nakala. ISIDORE is a platform allowing access to digital data in the Humanities and Social Sciences (e.g. archival and multi-media materials, articles, manuscripts, art exhibitions, survey data).

Noting that many teams and research projects do not have the necessary digital infrastructure that will provide a persistent and interoperable access to their digital data, Huma-Num has implemented a service called NAKALA exposure. NAKALA offers three types of services: one to give access to the data, another one to expose metadata and one to give PID to both access data and metadata.

For both Isidore and Nakala, Huma-Num prepared a dedicated endpoint from which the PARTHENOS aggregator can aggregate one XML file that describes some of the available data collections. Each XML file can be mapped into the PE Model via a dedicated mapping (mapping 432 for Isidore, mapping 433 for Nakala).

Mapping information (432 & 398 - Isidore)

- *ID*: 432 & 398
- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3meditor/index?type=mapping&action=view&lang=en&id=mapping432
- *Status*: under refinement
- *OAI header*: OAI header not available
- Issues:
- Base URI: http://parthenos.d4science.org/handle/Humanum/ Isidore

- Main PARTHENOS Entities Covered:
    - E24_Volatile_Dataset
    - PE22_Persistent_Dataset
    - E39_Actor
- PE15_Data_E-Service
- Generated PARTHENOS entities:
    - PE17_Curated_Data_E-Service
    - PE24_Volatile_Dataset
    - PE29_Access_Point

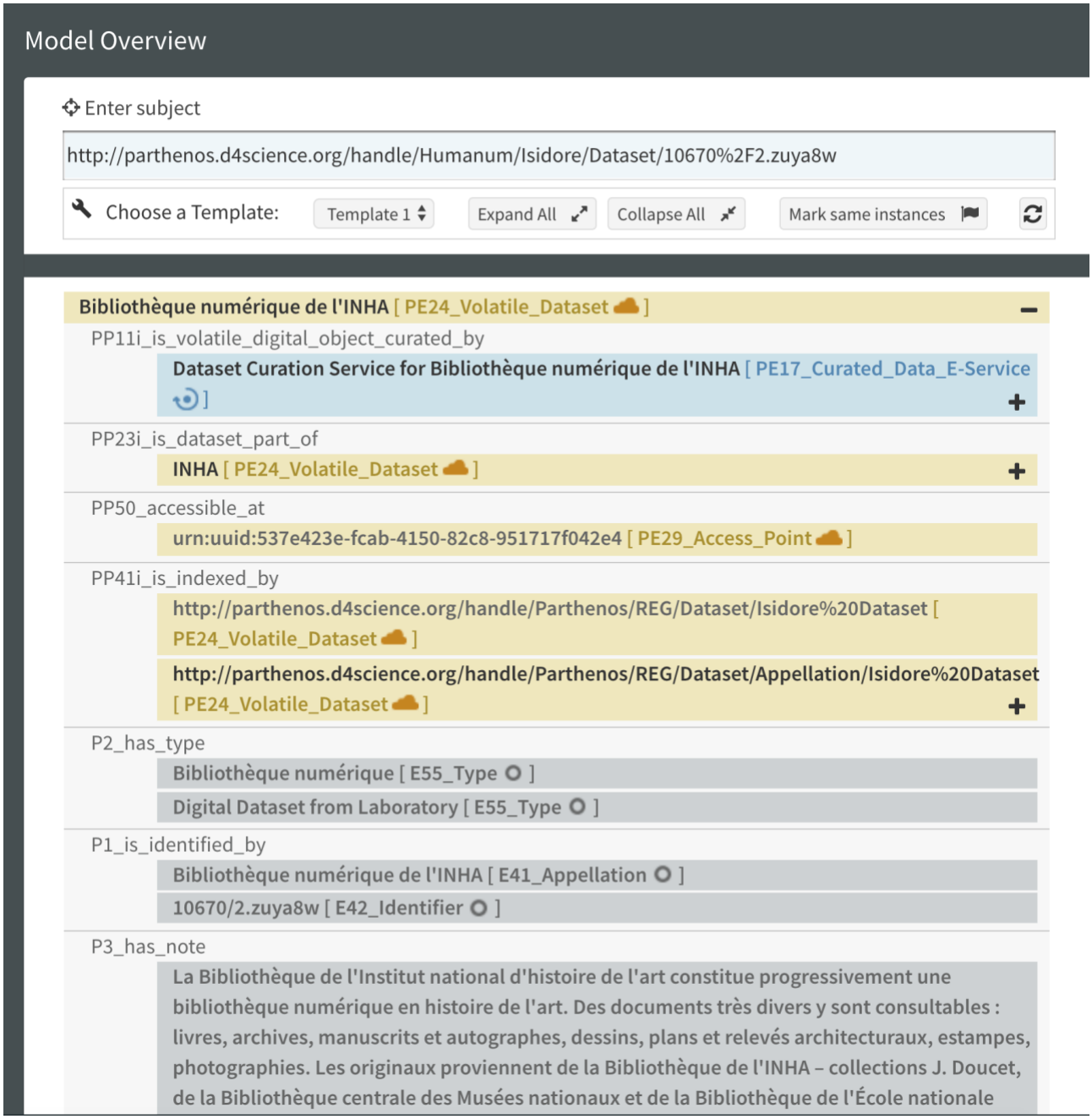


Figure 25 Example of mapped entity of Huma-Num Isidore

Mapping information (516 & 517,520,521,522 - Nakala)

- *ID*: 516 & 517,520,521,522
- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping516
- *Status*: complete
- *OAI header*: OAI header not available

- Issues:
- Base URI: http://parthenos.d4science.org/handle/Humanum/Nakala/
- Main PARTHENOS Entities Covered:
    - PE24_Volatile_Dataset
    - PE22_Persistent_Dataset
- Generated PARTHENOS entities:
    - PE12_Data_Curating_Service
    - PE15_Data_E-Service
    - PE24_Volatile_Dataset
    - PE29_Access_Point
    - PE34_Team

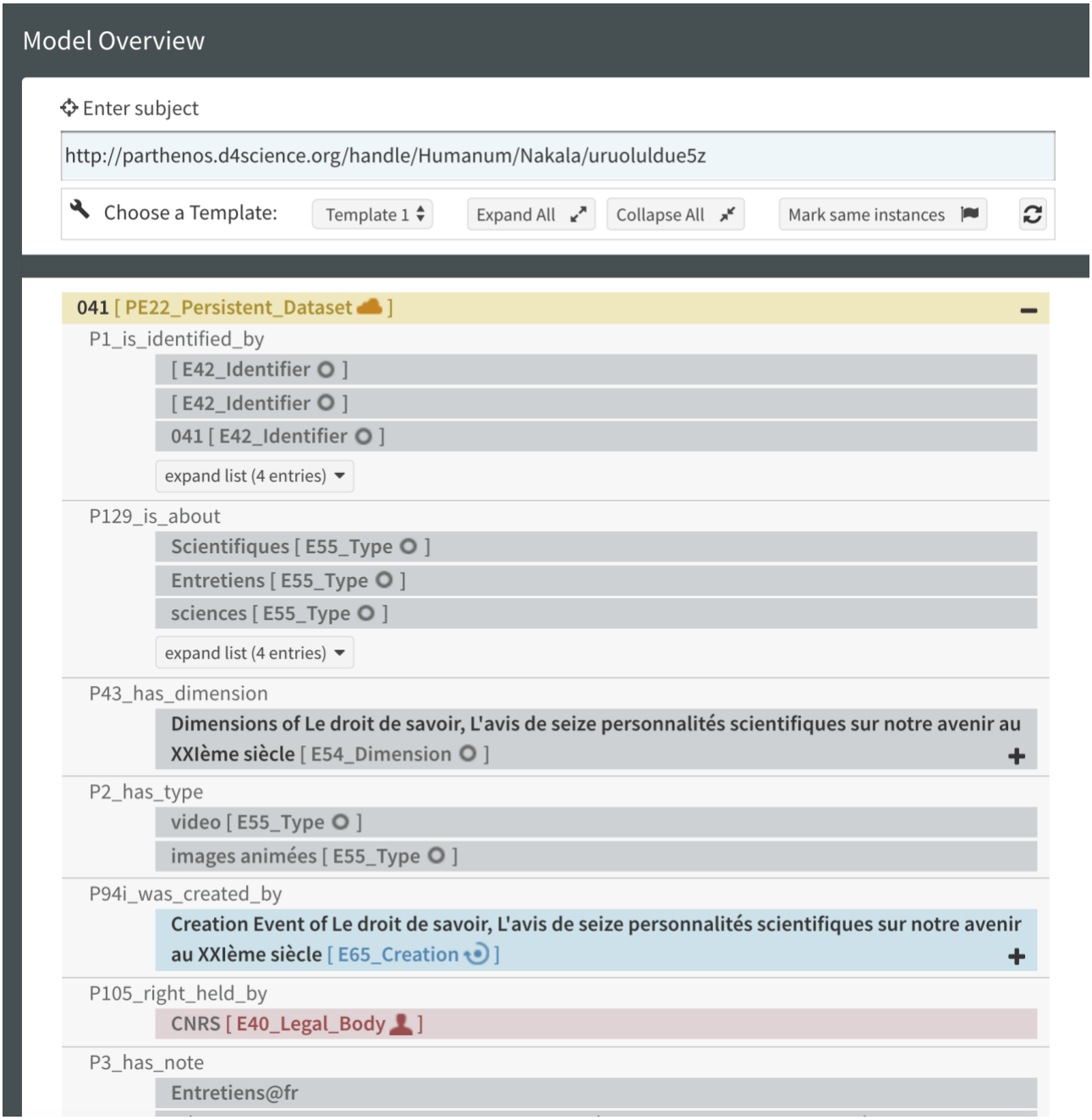


Figure 26 Example of mapped entity of Huma-Num Nakala

5.2.10 LRE Map

The Language Resource and Evaluation Map initiative issued out of the FLaReNet[15] project, whose mission was to develop a common vision of the area of LRs and to foster a European strategy for consolidating the sector, thus enhancing competitiveness at EU level and worldwide. The FLaReNet project produced a set of recommendations for the sector of digital Language Resources (LRs), encompassing creation, standardization,

[15] FLaReNet: http://www.flarenet.eu/

curation and long-term preservation. The correct documentation of LRs was indicated as crucial, and an initiative at the Language Resources and Evaluation Conference (LREC2010) was launched in collaboration with ELRA, to crowd-source the usage of LRs in papers submitted to the conference. The initiative continued within the following LREC conferences and extended to other events; today the LRE MAP[16] is a large repository of data, documenting language resources (well-known ones, but also minor ones and resources under development) using a lightweight metadata scheme. Authors are asked to enter a description for each language resource (whether their own or those of others) that they have used to carry out the research described in their paper. The LRE MAP is not actually a catalogue of language resources, but a collection of instances of uses of resources, so, for instance, well known and used LRs (e.g. Princeton WordNet, or the British National Corpus) have several entries in the LRE map

Mapping information

- ID: 447
- URL to 3m Editor: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&id=Mapping447&lang=en
- Status:
- OAI Header:
- Issues:
- Base URI: http://parthenos.d4science.org/handle/LRE/LRECatalog/
- Main PARTHENOS Entities Generated:
    - PE20_Volatile_Digital_Object
    - E7_Activity

[16] LRE Map: http://www.resourcebook.eu/searchll.php

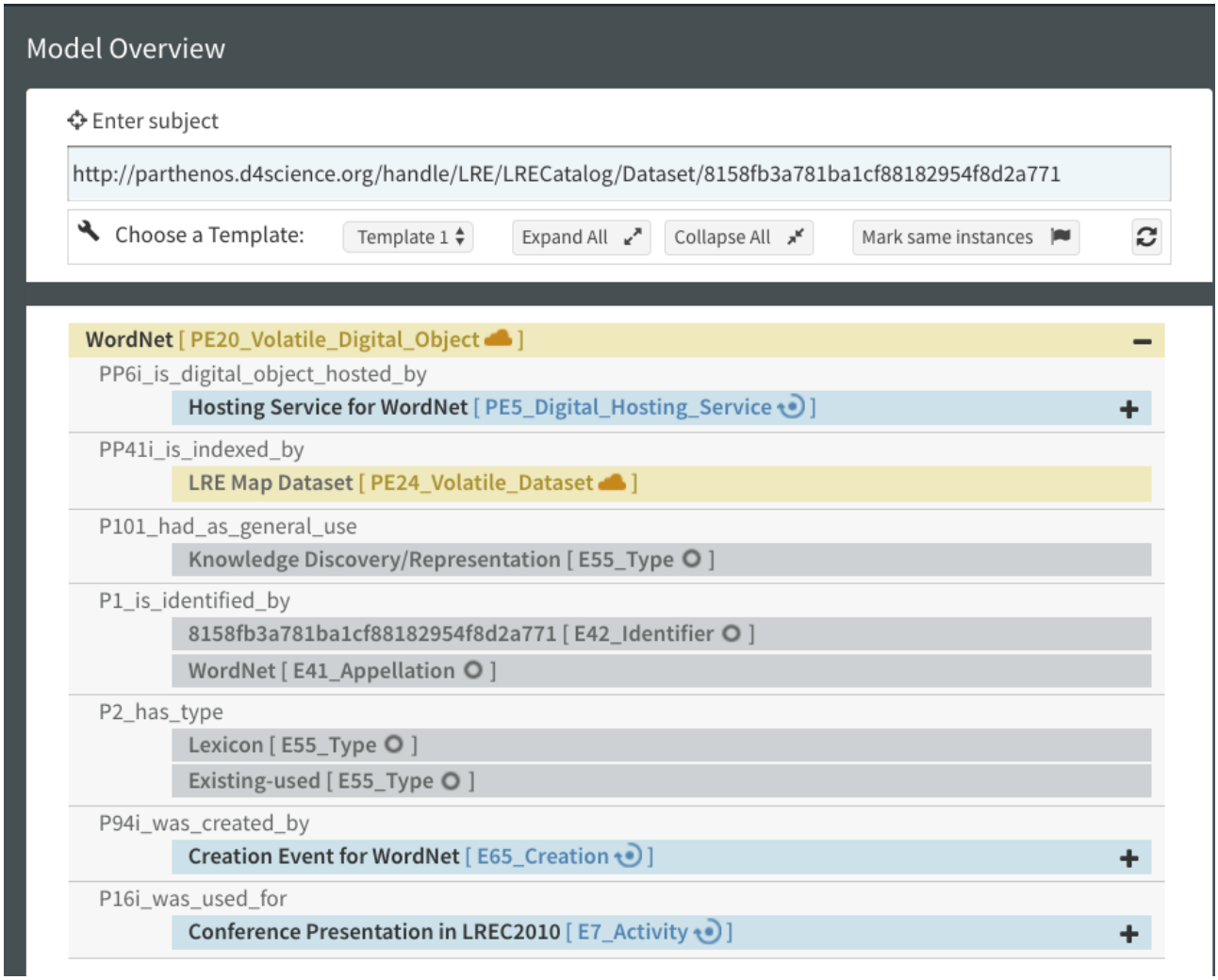


Figure 27 Example of mapped entity of LRE

5.2.11 METASHARE

The META-SHARE registry federation was implemented in the framework of the METANET Network of Excellence[17]. The META-SHARE registry contains information about Language Resources, Licenses, Projects, Actors and Documents. As to Language Resources, five different profiles are available, for Corpora, Lexical and Conceptual Resources (Lexicons, Ontologies…), Tools and Services (such as NLP software and online applications) and Language descriptions (e.g. language models or grammars).

[17] METANET: http://www.meta-net.eu/

Mapping information

- *ID*: 439
- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&action=view&lang=en&id=Mapping439
- *Status*: completed, under testing
- *OAI header*: OAI header not available
- Issues: None.
- Base URI: http://parthenos.d4science.org/handle/MetaShare/MetashareCatalog/
- Main PARTHENOS Entities Covered:
    - PE24_Volatile_Dataset
    - PE22_Persistent_Dataset
    - E21_Person
    - E31_Document
- Generated PARTHENOS entities:
    - PE3_Curating_Service
    - PE15_Data_E-Service
    - PE21_Persistent_Software
    - PE22_Persistent_Dataset
    - PE24_Volatile_Dataset
    - PE29_Access_Point
    - PE36_Competency_Type

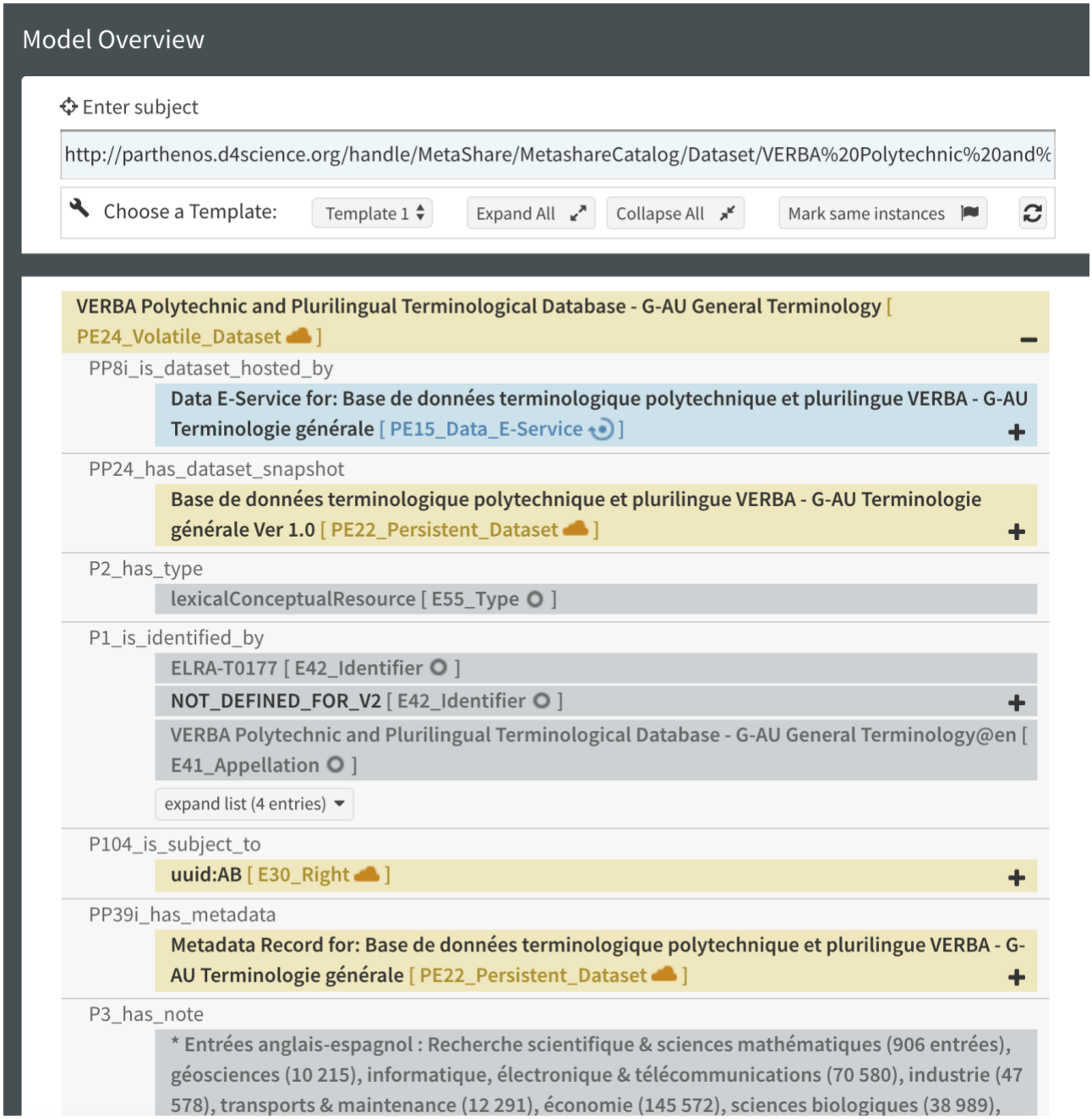


Figure 28 Example of mapped entity of Metashare

5.2.12 WP3 Policy Wizard

"Besides a theoretical deliverable, Data Archiving and Networked Services (DANS) has developed an easy to use tool called the PARTHENOS policy wizard; an interactive guide helping different users, like researchers, data managers or policy makers, to find and access policies and common guidelines tailored for the different humanities disciplines and various research activities." – PARTHENOS Policy Wizard Poster DHBenelux 2018.

Mapping information

- *ID*: 335
- *URL to 3M Editor*: https://mapping-d-PARTHENOS.d4science.org/3MEditor/Index?type=Mapping&id=Mapping335&lang=en
- *Status:*
- *OAI header:*
- *Issues:*
- *Base URI:* http://parthenos.d4science.org/handle/Parthenos/REG/
- *M*ain PARTHENOS Entities Covered:
    - PE22_Persistent_Dataset
    - E29_Design_or_Procedure

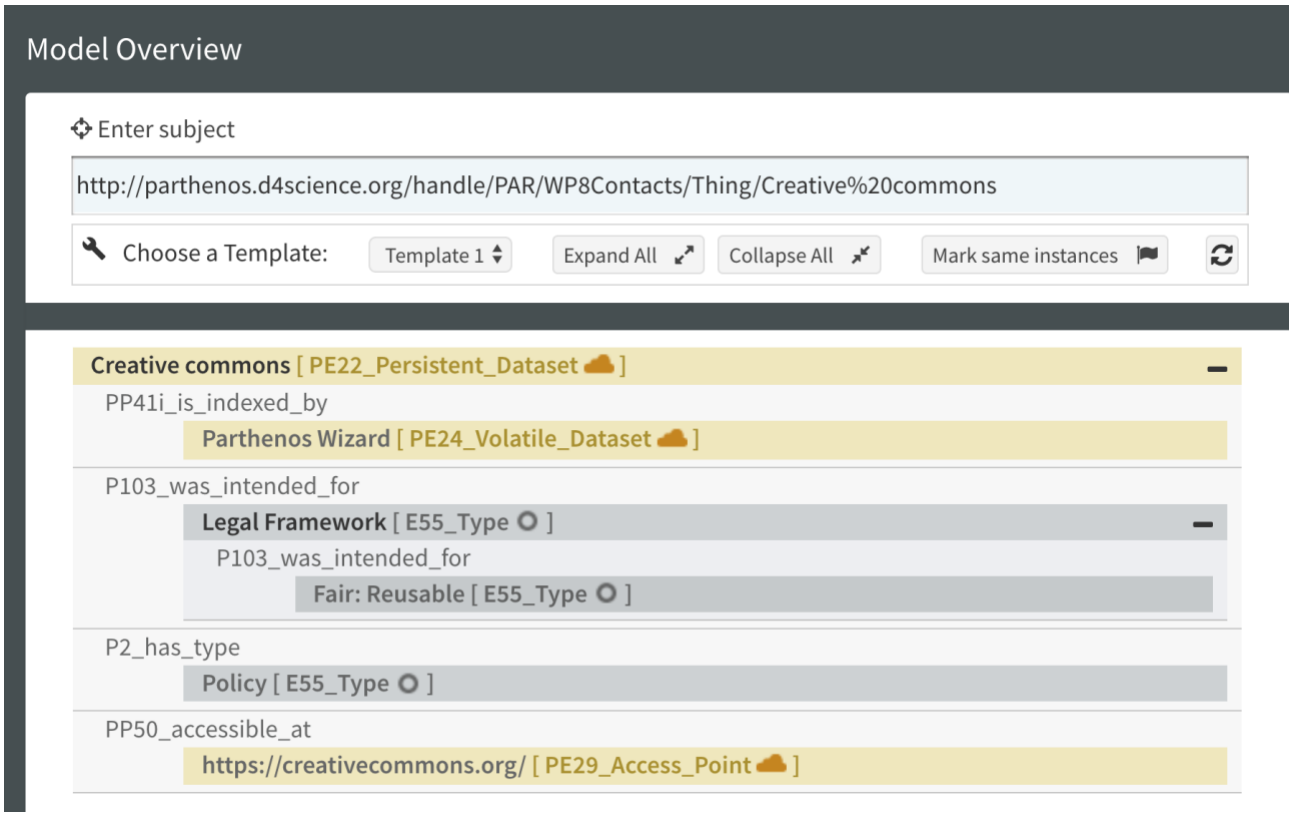


Figure 29 Example of mapped entity of WP8

5.2.13 WP4 Standards List

As part of WP4, the Standards Survival Toolkit (SSK), a collection of research use case scenarios, illustrating best practices in Digital Humanities and Heritage research, was implemented. As part of that implementation data gathering was done with regards to relevant standards to be referenced by the SSK. This dataset contains that list.

Mapping information

- *ID*: 449
- *URL to 3M Editor*: https://mapping-d-PARTHENOS.d4science.org/3MEditor/Index?type=Mapping&id=Mapping449&lang=en
- Status:
- OAI header:
- Issues:
- Base URI: http://PARTHENOS.d4science.org/handle/PARTHENOS/REG/
- Main PARTHENOS Entities Generated:
    - PE22_Persistent_Dataset
    - E29_Design_or_Procedure

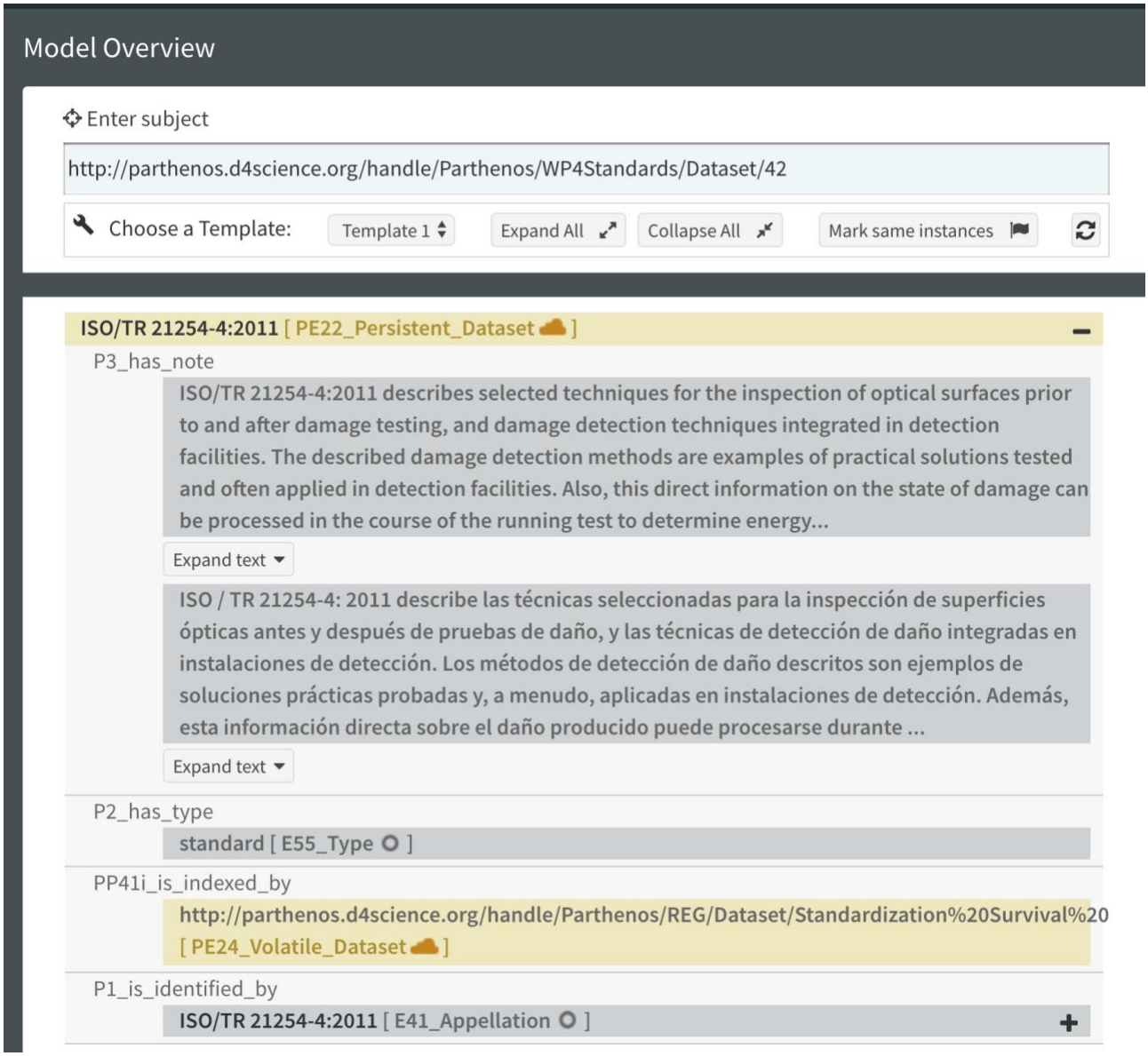


Figure 30 Example of mapped entity of WP4

5.2.14 WP5 Standard Reference Resources

The PARTHENOS Standard Reference Resources dataset was produced in the context of WP5 T5.3 which aimed to collate useful reference resources for standardizing data in the context of a Research Infrastructure information management scenario. The resources listed here document the sources used to generate standardized thesauri for normalizing data in the PARTHENOS Joint Resource Registry. This is the metadata for the thesauri resources employed.

Mapping information

- *ID*: 547
- *URL to 3M Editor*: https://mapping-d-parthenos.d4science.org/3MEditor/Index?type=Mapping&id=Mapping547&lang=en
- Status:
- OAI header:
- Issues:
- Base URI: http://parthenos.d4science.org/handle/Parthenos/REG/
- Main PARTHENOS Entities *Generated*:
    - PE24_Volatile_Dataset
- *Generated PARTHENOS entities*:

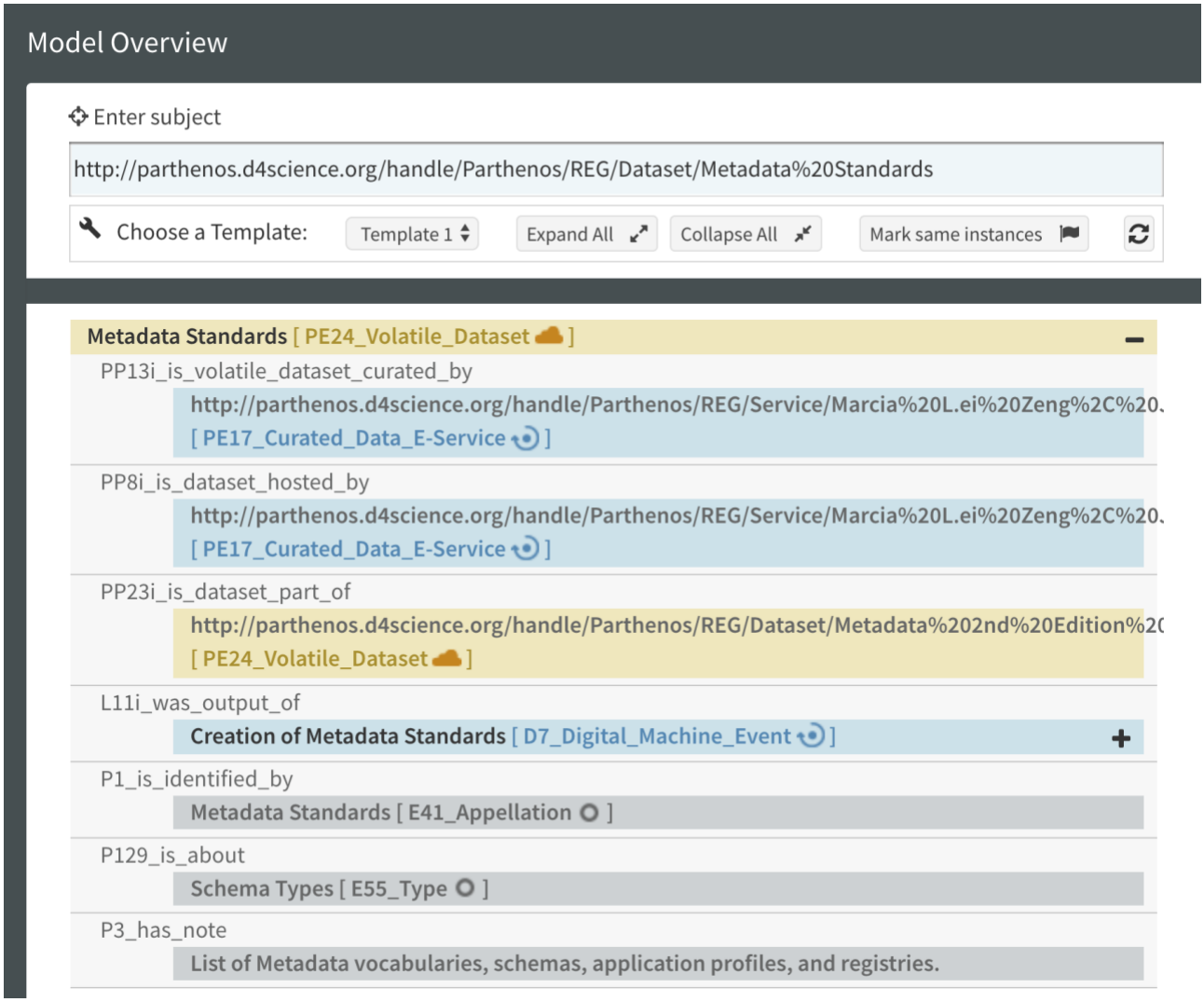


Figure 31 Example of mapped entity of WP5

5.2.15 WP8 International Contacts

Part of the activities of WP8 included the generation and curation of an-up-to-date list of international contact in the digital humanities, their areas of activity and the means to contact them. This list is represented here.

Mapping information

- *ID*: 464
- *URL to 3M Editor*: https://mapping-d-PARTHENOS.d4science.org/3MEditor/Index?type=Mapping&id=Mapping464&lang=en
- Status:
- OAI header:
- Issues:
- Base URI: http://parthenos.d4science.org/handle/Parthenos/REG/
- Main PARTHENOS *Entities* Generated:
    - E74_Group

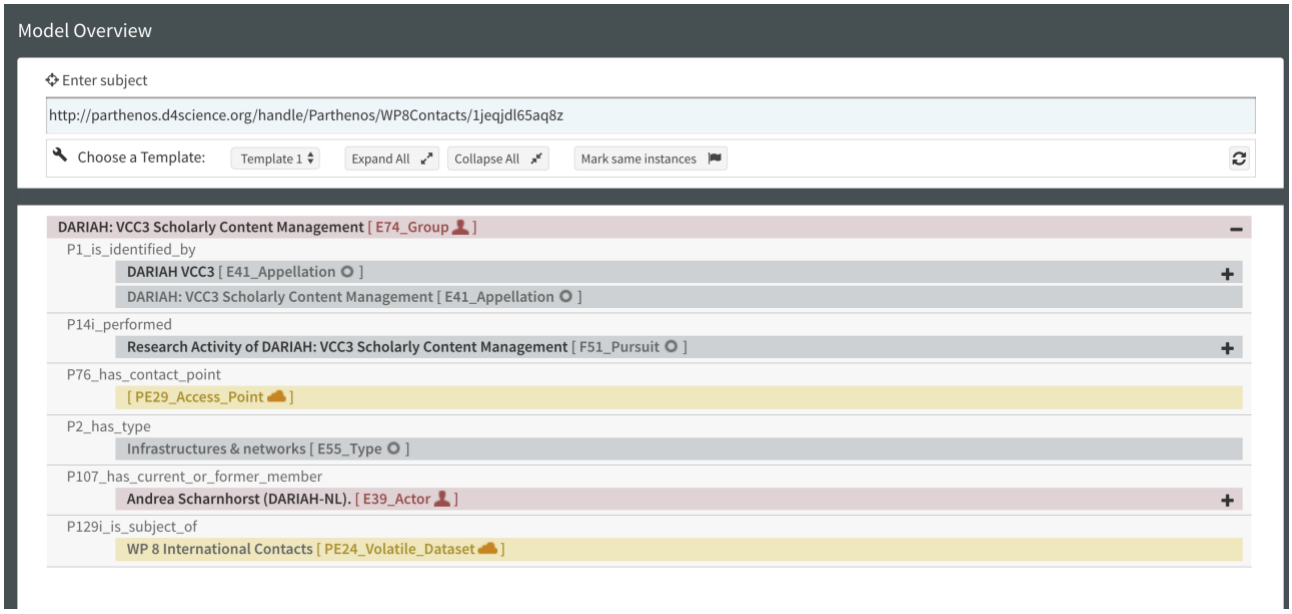


Figure 32 Example of mapped entity of WP8

6. Metadata Quality Checking

A critical issue in a large-scale, heterogeneous aggregation endeavour as pursued in PARTHENOS is the quality of the (meta)data as this has a decisive impact on the usability of the data for resource discovery. Not surprisingly, the usual classes of problems with data quality in aggregation scenarios were encountered:

- **Missing data** - there are often large lacunae in the aggregated data space. This can be due to erroneous or incomplete mapping, but mostly it is in the source instance data that certain characteristics of a resource are not made explicit, and/or are left out.
- **Literals referring to entities** - ideally a reference to an entity should be done with an unambiguous identifier, a URI. However, often due to limitations of the source metadata schema and/or the metadata authoring tools, simple literals are used to denote entities such as persons or institutions. This approach is inherently prone to spelling variations and ambiguous references. The PEM offers a clean way to model these entities, however, the problems in the source data counteract this potential.

  In many cases, entities are degraded to just literal properties of a given resource, e.g. a given publishing house as “publisher” property, or an actor in the role of a creator or contributor for a given resource. A major challenge in the mapping efforts was to generate PEM entities out of these underspecified references, especially to generate a sensible, stable URI denoting a given entity.
- **Variation of values / Variability of descriptors** - related to the previous problem, values in fields often come in various spellings or language variants, with strong adverse effect on the recall and discoverability of the resources.
- **Underspecified semantics in the original metadata schema** - when mapping the source schemas to the PEM, we encountered multiple situations where the meaning of certain elements in the source schema are not well defined. A typical example is mingling/convolution of instances of Service and Software. While these are semantically clearly separated in the PEM, in the source formats, as in colloquial use, these are often used interchangeably.

All listed issues have a strong influence on the quality of the resulting harmonized metadata and dramatically hamper the recall. Literal reference to entities 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.

Descriptive statistics, i.e. information about the number of occurrences of various phenomena in a given dataset, is crucial for getting an overview and for understanding the data. It is also an indispensable input for any quality assessment or curation work. What follows is a selection of numerical summaries, as well as an overview of the minimal metadata coverage for five main entities: Thing, Dataset, Service, Actor, Software. At the end of the chapter a few data quality issues are singled out and briefly described.

6.1 Statistics

This selection of some preliminary (as of 2019-02-25) statistics give a rough indication of the content of the aggregated data space available via the PARTHENOS Virtuoso server [18] . These statistics were generated mainly as SPARQL queries via the PARTHENOS Discovery tools (cf. D6.4). The following dimensions are covered:

- **Data Providers** - how much data is coming from each provider?
- **Classes** - how many instances of which class are present?
- **Properties** - what properties are used?
- **Types** – the PEM makes extensive use of the CIDOC-CRM typing construct *p02_hasType -> E55 Type*
  A dedicated page shows which types are most frequent and also distinguished use of types per Class, e.g. the CIDOC-CRM class *crm:E51_Contact_Point* may be further categorized with E55 Type as “email”, “phone” or “fax”
- **Aboutness** - Another crucial property in CIDOC-CRM is *crm:P129_is_about* which links a Dataset to a Concept (or in fact a crm:E89_Propositional_Object to any crm:E1_CRM_Entity

[18] PARTHENOS Virtuoso SPARQL endpoint: https://virtuoso.parthenos.d4science.org/sparql

6.1.1 By data source / provider

Each piece of information (triple) in the triple store holding the aggregated and transformed data is contextualized with respect to provenance. Through special annotations on the level of graphs, one can retrieve the information about provider and the data was collected.

The PARTHENOS Virtuoso server contains one special namespace for provenance information coming from the PARTHENOS Aggregator: *dnet:http://www.d-net.research-infrastructures.eu/provenance*. The aggregator creates and manages one provenance graph (with URI *http://www.d-net.research-infrastructures.eu/provenance/graph*) that holds provenance information about all the RDF/XML records of the aggregation:

- The date when the original was collected. This information is available in triples in the form *<RDF file ID>* *http://www.d-net.research-infrastructures.eu/provenance/collectedInDate* *<the date>*
- The date when the original was transformed into an RDF/XML compliant to the PE Model. This information is available in triples in the form *<RDF file ID>* *http://www.d-net.research-infrastructures.eu/provenance/transformedInDate* *<the date>*
- The endpoint (API) from which the aggregator collected the original XML file. This information is available in triples in the form *<RDF file ID>* *http://www.d-net.research-infrastructures.eu/provenance/collectedFrom* *<API ID>*
- The name of the data source/provider the endpoint belongs to. This information is available in triples in the form <API ID> *http://www.d-net.research-infrastructures.eu/provenance/isApiOf* *"Data source/provider name".*

All the triples coming from an RDF/XML file are stored in a dedicated graph whose URI is the RDF file ID.

To retrieve the number of graphs per provider (i.e. the number of RDF/XML aggregated per provider), the following SPARQL query making use of the graph annotations described above has been used:

```
SELECT ?source (count (?record) as ?cnt) WHERE {
      GRAPH dnet:graph {
            ?record dnet:collectedFrom ?api .
            ?api dnet:isApiOf    ?source.
            }
      }
```

GROUP BY ?source

Table 4 gives a purely quantitative view on the size of the datasets by individual providers obtained running the SPARQL query above.

Table 4 Number of RDF files per provider (February 2019)

Provider	Count of triples
CLARIN	9,516,093
CENDARI	4,579,112
Huma-Num - Nakala	3,714,159
ARIADNE	3,494,326
European Holocaust Research Infrastructure	1,620,113
CulturaItalia	445,416
LRE Map	442,946
METASHARE	244,565
DARIAH-GR DYAS	43,320
PARTHENOS WP8	3,389
PARTHENOS	2,126
PARTHENOS WP3	1,624
PARTHENOS WP4	704

6.1.2 By class

Overall instances for sixty classes from the PEM are present in the dataset. Table 5 lists the top thirty classes with the number of respective instances.

Table 5 Top thirty classes with higher number of instances

Class	# instances	Class	# instances
Time-Span	460,487	Place Appellation	61,587
Data E-Service	373,416	Right	57,660
Identifier	372,983	Pursuit	53,532
Access Point	359,476	Place	47,591
Persistent Dataset	358,991	Activity	41,280
Type	326,603	Dimension	33,468
D7_Digital_Machine_Event	258,488	Address	32,675
Creation	177,629	Collection	20,395
Actor	167,905	Attribute Assignment	20,178
Title	166,211	Acquisition	18,588
Publication Event	153,228	Language	18,508
Volatile Dataset	138,195	Death	15,199
Person	134,312	Birth	15,187
Man-Made Object	132,321	Production	12,904
Linguistic Object	128,925	Transfer of Custody	12,460
Information Object	108,448	Spatial Coordinates	12,395
Appellation	108,271	Digital Hosting Service	7,453

6.1.3 Properties

Currently one hundred and seventeen distinct properties are in use. Table 6 lists the fifty most used properties.

Table 6 The fifty most used properties and the number of their occurrences

Property name	# usage	Property name	# usage
type	8,162,677	was present at	135,053
label	6,060,296	PP32i_was_curated_by	135,008
has type	1,713,460	carried out by	121,296
is identified by	1,141,153	carries	100,344
has note	1,014,577	ongoing throughout	83,164
is dataset hosted by	398,940	performed	79,501
provides access point	378,178	is composed of	77,246
is about	371,322	was brought into existence by	76,085
transformedInDate	369,368	has language	71,846
collectedInDate	369,164	has dataset part	69,774
collectedFrom	368,170	is subject to	65,921
accessible at	357,351	has title	60,678
has time-span	340,469	had typical subject	53,534
at some time within	335,072	is digital object part of	41,288
L23_used_software_or_firmware	294,804	consists of	40,671
is listed in	218,194	took place at	39,246
L11i_was_output_of	211,895	has dimension	33,469
P14.1_in_the_role_of	207,172	has current or former residence	32,530
is domain of	206,028	is component of	30,586
has range	190,181	has former or current owner	29,671
is dataset part of	189,369	L10i_was_input_of	27,488
is metadata for	158,970	had participant	27,451
was created by	156,546	was attributed by	21,430
P4_has_time_span	140,020	assigned	20,192
PP2_provided_by	135,168	provided by	20,171

6.1.4 Types

E55_Type together with the generic property P02_hasType is an important mechanism in CIDOC-CRM to further classify/categorize entities of all kinds. Currently, 260,338 distinct types are in use.

Table 7, lists the number of distinct types/categories per class and reveals that, while for some classes a clean consistent categorisation has been achieved (small number of distinct types , e.g. for Access Point or Volatile Dataset), for others there still seems to be an issue with the mapping, generating too many types (Man-Made Object, Identifier). The assumption is that the number of types serving to categorize some class should be limited, definitely by orders of magnitude smaller than the number of categorized instances.

Table 7 Types by class and occurrences

Class	Count distinct types	Count instances
Access Point	5	349,946
Persistent Dataset	136	261,543
Volatile Dataset	4	133,354
Man-Made Object	132,382	132,321
Linguistic Object	554	112,484
Appellation	4	89,397
D7_Digital_Machine_Event	5	72,910
Identifier	92,008	47,191
Right	994	29,105
Creation	1	16,731
Information Object	15,915	15,915
Activity	4	15,323

6.1.5 Aboutness

Another important dimension for discovery and exploration of the dataset is the aboutness relation, i.e. all kinds of information that give a clue about the actual content of the resources, their spatial or temporal coverage, or which concepts and entities are mentioned, much like the subject headings in library catalogues.

Currently, 55,852 distinct concepts (linked to an entity with `crm:P129_is_about` property) are in the data. Table 8 indicates which entities (of class Dataset, Collection etc.) are tagged with which types of concepts (Type, Place, Period, Actor). How many distinct concepts for each combination exist and how many instances are actually described in this manner are provided in the last two columns, respectively.

Table 8 Aboutness: usage of concept classes and instances

Class	Concept Type	Count distinct concepts	Count instances
Linguistic Object	Type	14,966	66,489
Persistent Dataset	Type	6,787	64,383
Linguistic Object	Place	6,709	45,576
Persistent Dataset	Place	3,293	41,296
Volatile Dataset	Type	306	34,062
Persistent Dataset	Linguistic Object	18,518	20,133
Linguistic Object	Linguistic Object	18,518	20,133
Persistent Dataset	Collection	18,518	20,133
Linguistic Object	Collection	18,518	20,133
Persistent Dataset	Period	289	13,757
Linguistic Object	Period	289	13,757
Collection	Type	8,341	9,476
Persistent Dataset	Actor	6,333	6,333
Collection	Place	3,416	4,280
Linguistic Object	Actor	8,319	3,802
Collection	Actor	8,319	3,802
Volatile Dataset	Time-Span	249	242
Volatile Dataset	Period	10	14
Volatile Dataset	Place	2	1

6.2 Minimal metadata coverage of main types

Critical issue for quality of (meta)data is the coverage, i.e. how many of the expected fields are filled. In the course of the project a set of “minimal metadata” for the main entity types has been defined. The coverage of this minimal metadata has been evaluated against the dataset from 2019-02 using the query-automatization tool SparqLaborer (cf. D6.4).
Note: the indicated ratio is computed relative to the overall number of instances for a given main class, not considering that some of the relations or properties are only applicable to certain subclasses. Also, there may be multiple relations between entities of two classes, e.g. a Thing can have multiple Appellations, potentially yielding a coverage ratio greater than 1.

6.2.1 Metadata coverage of E70 Thing

Number of instances: **2,396,196**

Table 9 Coverage of minimal metadata of E70_Thing

Relation between instances of classes	CIDOC CRM path	Number of relations	Ratio of relations to instances
thing -label-> literal	E70-label->literal	1,605,130	0.67
thing -hasType-> e55type	E70-P2->E55	1,084,715	0.45
thing -isIdentifiedBy-> appellation	E70-P1->E41	1,011,776	0.42
thing -hasNote-> literal	E70-P3->E62	819,935	0.34
thing -isIdentifiedBy-> identifier	E70-P1->E42	465,356	0.19
informationObject -isAbout-> entity	E73-P129->E1	357,403	0.15
digitalMachineEvent -hadOutput-> digitalObject	D7-L11->D1	207,893	0.09
physicalThing -isComposedOf-> physicalThing	E18-P46->E18	70,562	0.03
thing -isSubjectTo-> right	E70-P104->E30	65,921	0.03
volatileDigitalObject -hasDigitalObjectPart-> digitalObject	PE20-PP18->D1	39,420	0.02
hostingService -hostsObject-> thing	PE2-PP4->E70	20,132	0.01

iInformationObject -refersTo-> entity	E73-P67->E1	17,324	0.01
volatileDataset -isIndexOf-> digitalObject	PE24-PP41->D1	7,608	<0.01
digitalHostingService-hostsDigitalObject-> digitalObject	PE5-PP6->D1	7,453	<0.01
curationActivity -curates-> collection	E87-P147->E78	2,914	<0.01
thing -rightHeldBy-> actor	E70-P105->E39	2,763	<0.01
curatingService -curates-> curatedThing	PE3-PP32->PE32	1,571	<0.01
digitalObject-isComposedOf-> informationObject	E73-P106->E73	0	0
volatileDigitalObject-hasSnapshot-> persistentDigitalObject_instance	PE20-PP17->PE19	0	0
digitalCuratingService-curates-> volatileDigitalObject	PE10-PP11->PE20	0	0
digitalMachineEvent-hadOutput-> digital object, digitalMachineEvent-carriedOutBy-> actor	D7-L11->D1 D7-P14->E39	0	0
digitalObject-hasCurrentKeeper-> accessPoint	D1-P50->PE29	0	0
digitalCuratingService - curatesVolatileDigitalObject-> volatileDigitalObject	PE10-PP11->PE20	0	0
digitalObject-hasDigitalObjectPart-> volatileDigitalObject	D1-PP18->PE20	0	0
persistentDigitalObject - hasPersistentDigitalObjectPart-> persistentDigitalObject	PE19-PP16->PE19	0	0

6.2.2 Metadata coverage of PE18 Dataset

Number of instances: 497,182

Table 10 Coverage of minimal metadata of PE18_Dataset

Relation between instances of types	CIDOC CRM path	number of relations	Ratio of relations to instances
dataset -hasNote-> literal	PE18-P3->Literal	742,975	1.49
dataset -isIdentifiedBy-> appellation	PE18-P1->E41	671,349	1.35
dataset -hasType-> e55Type	PE18-P2->E55	411,519	0.83
dataHostingService -hostsDataset-> dataset	PE7-PP8->PE18	372,757	0.75
dataset -label-> literal	PE18-label->literal	298,453	0.6
dataset -isAbout-> e55type	PE18-P129->E55	187,889	0.38
Dataset -hasDatasetPart-> volatileDataset	PE18-PP23->PE24	135,494	0.27
dataset -isAbout-> place	PE18-P129->E53	49,559	0.1
dataset -isAbout-> temporalEntity	PE18-P129->E2	38,085	0.08
volatileDataset -isIndexOf-> digitalObject	PE24-PP41->D1	7,608	0.02
volatileDataset -hasDatasetSnapshot-> persistentDataset	PE24-PP24->PE22	1,214	<0.01
dataCuratingService -curatesVolatileDataset-> volatileDataset	PE12-PP13->PE24	23	<0.01
persistentDataset -hasPersistentDatasetPart-> persistentDataset	PE22-PP20->PE22	0	0
volatileSoftware -hasRelease-> persistentSoftware	PE23-PP22->PE21	0	0
persistentDataset -isMetadataFor-> digitalObject	PE22-P39->D1	0	0
persistentDataset -isMetadataFor-> digitalObject	PE22-P39->D1	0	0

6.2.3 Metadata coverage of PE1 Service

Number of instances: 389,541

Table 11 Coverage of minimal metadata of PE1_Service

Relation between instances of types	CIDOC CRM path	Number of relations	Ratio of relations to instances
dataHostingService -hostsDataset-> dataset	PE7-PP8->PE18	372,757	0.96
eService -providesAccessPoint-> accessPoint	PE8-PP49->PE29	372,322	0.96
service -providedBy-> actor	PE12-PP2->E39	132,820	0.34
service -label-> literal	PE1-label->literal	37,053	0.1
hostingService -hostsObject-> thing	PE2-PP4->E70	20,132	0.05
service -hasE55type-> e55type	PE1-P2->E55	15,432	0.04
digitalHostingService -hostsDigitalObject-> digitalObject	PE5-PP6->D1	7,453	0.02
curatingService -curates-> curatedThing	PE3-PP32->PE32	1,571	<0.01
digitalHostingService-usedGeneralTechnique-> e55type	PE5-P32->E55	308	<0.01
eService -usedGeneralTechnique-> e55type	PE8-P32->E55	308	<0.01
Service -providedBy-> group	PE1-PP2->E74	38	<0.01
Service -hasCompetency-> competencyType	PE1-PP45->PE36	37	<0.01
Service -isIdentifiedBy-> appellation	PE1-P1->E41	26	<0.01
Service -hasNote-> actor	PE1-P3->E62	25	<0.01
eService -hasDesignatedAccessPoint->	PE8-PP28-	25	<0.01

accessPoint	>PE29		
dataCuratingService -curatesVolatileDataset-> volatileDataset	PE12-PP13->PE24	23	<0.01
eService -hasProtocolType-> protocolType	PE8-PP47->PE37	17	<0.01
service -hasDeclarativeTime-> literal	PE1-PP42->literal	15	<0.01
service -usedSpecificObject-> right, right -hasNote-> literal	PE1-P16->E30, E30-P3->E62	9	<0.01
service -usedSpecificObject-> right, right -hasType-> e55type	PE1-P16->E30, E30-p2->E55	5	<0.01
service -isIdentifiedBy-> identifier	PE1-P1->E42	0	0
service -hasType-> availabilityType	PE1-P2->PE39	0	0
service -providedBy-> group, group -hasMember-> person	PE1-PP2->E74, E74-P107->E21	0	0
digitalCuratingService -curatesVolatileDigitalObject-> volatileDigitalObject	PE10-PP11->PE20	0	0
softwareCuratingService -curatesVolatileSoftware-> software	PE11-PP12->D14	0	0
softwareComputingEService -runsOnRequest-> software	PE13-PP14->D14	0	0
softwareDeliveryEService -deliversOnRequest-> software	PE14-PP15->D14	0	0
eService -usesAccessProtocol-> software	PE8-PP29->D14	0	0
eService -usesProtocolParameter-> schema	PE8-PP48->PE38	0	0

6.2.4 Metadata coverage of E39 Actor

Number of instances: 172,956

Table 12 Coverage of minimal metadata of E39_Actor

Relation between instances of types	CIDOC CRM path	Number of relations	Ratio of relations to instances
activity -carriedOutBy-> actor	E7-P14->E39	200,791	1.16
service -providedBy-> actor	PE12-PP2->E39	132,820	0.77
actor -isIdentifiedBy-> appellation	E39-P1->E41	117,469	0.68
actor -label-> literal	E39-label->literal	57,684	0.33
actor -isIdentifiedBy-> identifier	E39-P1->E42	17,254	0.1
actor -hasNote-> literal	E39-P3->literal	16,533	0.1
actor -hasResidence-> place	E39-P74->E53	13,045	0.08
actor -hasType-> e55type	E39-P2->E55	9,347	0.05
actor -hasContactPoint-> contactPoint	E39-P76->E51	3,699	0.02
Thing -rightHeldBy-> actor	E70-P105->E39	2,763	0.02
actor .... contactPoint -label-> literal, literal = "email"	E39-P76->E51, E51-P2->E55, E55-label->literal	1,467	0.01
group -hasMember-> actor	E74-P107->E39	801	<0.01
actor -hasContactPoint-> accessPoint	E39-P76->PE29	674	<0.01
actor -hasContactPoint-> contactPoint	E39-P76->E45	578	<0.01
actor .... contactPoint -label-> literal, literal = "phone"	E39-P76->E51, E51-P2->E55, E55-label->literal	566	<0.01
digitalMachineEvent -hadOutput-> digital object, digitalMachineEvent -carriedOutBy-> actor	D7-L11->D1, D7-P14->E39	0	0
digitalMachineEvent -hadOutput-> digital object, digitalMachineEvent -carriedOutBy-> actor	D7-L11->D1, D7-P14->E39	0	0

6.2.5 Metadata coverage of E53 Place

Number of instances: 47,591

Table 13 Coverage of minimal metadata of E53_Place

Relation between instances of types	CIDOC CRM path	Number of relations	Ratio of relations to instances
Dataset -isAbout-> place	PE18-P129->E53	49,559	1.04
period -tookPlaceAt-> place	E4-P7->E53	39,246	0.82
place -isIdentifiedBy-> identifier	E53-P1->E42	32,560	0.68
place -label-> literal	E53-label->literal	28,920	0.61
attributeAssignment -assignedAttributeToAttributeTo-> place	E13-P140->E53	21,430	0.45
place -fallsWithin-> place	E53-P89->E53	14,276	0.3
actor -hasResidence-> place	E39-P74->E53	13,045	0.27
place -isIdentifiedBy-> identifier	E53-P1->E42	4,552	0.1
physicalThing -hasLocation-> place	E18-P53->E53	798	0.02
propositionalObject -refersTo-> place	E89-P67->E53	737	0.02
physicalObject -hasPermanentLocation-> place	E19-P54->E53	0	0

6.2.6 Metadata coverage of D14 Software

Number of instances: 26

Table 14 Coverage of minimal metadata of D14_Software

Relation between instances of types	CIDOC CRM path	Number of relations	Ratio of relations to instances
digitalMachineEvent -usedSoftware-> software	D7-L23->D14	290046	11155.62
software -label-> literal	D14-label->literal	19	0.73
software -wasIntendedFor-> e55type	D14-P103->E55	7	0.27
persistentSoftware -hasPersistentSoftwarePart-> persistentSoftware	PE21-PP19->PE21	0	0
volatileSoftware -hasSoftwarePart-> software	PE23-PP21->D14	0	0
software -isIdentifiedBy-> identifier	D14-P1->E42	0	0
software -isIdentifiedBy-> appellation	D14-P1->E41	0	0
softwareCuratingService -curates-> volatileSoftware	PE11-PP12->PE23	0	0
softwareDeliveryService -delivers-> software	PE14-PP15->D14	0	0
softwareComputingService -runs-> software	PE13-PP14->D14	0	0
softwareHostingService -hosts-> software	PE6-PP7->D14	0	0
digitalMachineEvent -hadOutput-> software, digitalMachineEvent -used-> programmingLanguage	D7-L11->D14, D7-P33->PE40	0	0

softwareHostingService -hostsSoftwareObject-> software	PE6-PP7->D14	0	0
softwareCuratingService -curatesVolatileSoftware-> software	PE11-PP12->D14	0	0
softwareComputingEService -runsOnRequest-> software	PE13-PP14->D14	0	0
softwareDeliveryEService -deliversOnRequest-> software	PE14-PP15->D14	0	0
eService -usesAccessProtocol-> software	PE8-PP29->D14	0	0
eService -usesProtocolParameter-> schema	PE8-PP48->PE38	0	0

6.3 Selected data quality issues

- **Missing Labels**
  While the mapping tool allows and encourages the definition of labels for all generated entities, there is still a substantial portion of entities without a label (2,752,985 out of 4,944,829). To some extent this can be attributed to auxiliary entities dictated by the data model, not directly relevant for the user, and thus not needing a descriptive label. However, even for the main entity types, the coverage is suboptimal: round 60% for Thing, Dataset or Place, 33% for Actor.
- **Random identifiers/UUIDs**
  In the mapping process, unique identifiers, ideally URLs adhering to a given nomenclature, need to be generated to identify the created instances. A fall-back mechanism is to assign randomly generated UUIDs (universal unique identifier), especially for “artificial” instances that are dictated by CIDOC-CRM, but are of no direct use to the user. However, there are cases where relevant entities, like Types or Places have been assigned a UUID. This potentially leads to major proliferation of instances: While the generated URI for the same concept referred to by the same literal string in two records would be the same, if only a UUID is assigned, this will be a new one for every record.

- **Inverse properties**

  The PEM defines the properties as pairs of inverse properties, i.e. *hasMember* vs. *memberOf*. In the individual mappings, presumably dictated primarily by the available information in the source data, there is no consistent use of such property pairs. This is semantically valid and triple stores are potentially capable of inferencing new facts from data in combination with the ontology axioms. (I.e. for each A -hasMember-> B, it would generate the corresponding triple B -isMemberOf-> A) However, it complicates the customization of the interface, as it is required to always check for both directions.
- **Places & Spatial Coordinates**

  Many instances of E53_Place are missing geocoding information. And if the geocoding is present it is mostly formatted in a non-standard way (distinction between latitudes and longitudes). Out of over 47.000 Places, only around 1.100 come with geocoding information, all of which from one provider (CulturaItalia). Places would lend themselves ideally for post-aggregation enrichment, where Place instances in the data could be automatically matched against large reference resources like Wikidata or GeoNames.
- **Image representation**

  Also in case of images (e.g. as photo of the described artefact) the data space is quite scarce, all available data (round 5,000 images) coming again just from one provider (CulturaItalia)
- **Same information in many graphs**

  Presumably due to modelling error in the mapping process, certain triples are repeated many times. E.g. the type of a collection consisting of many individual datasets may be indicated repeatedly in the context of every dataset. Albeit this is technically not incorrect, it clutters the data space with duplicate information leading to confusion of the user and potential problems (unexpected expansion of the result) in complex queries.
- **Disambiguation of entities**

  Due to the principal uncertainty when trying to disambiguate a literal reference to an entity, a policy has been adopted, that if the same literal is encountered in the context of one provider it is considered the same identical entity. However, when the identical literals are in records coming from different sources, the probability that they don’t refer to the same entity is deemed higher, therefore, in such

cases, two distinct entities are created, even though with an identical label or appellation.

For example, the Person “Andreas Witt”, which is introduced both via the provider PARTHENOS WP8 and LREMap, this leads to two distinct entities:

1) http://parthenos.d4science.org/handle/Parthenos/WP8Contacts/soccvikfef6p

2) http://parthenos.d4science.org/handle/LRE/LRECatalog/soccvikfef6p

7. Conclusions

In order to create a Joint Resource Registry and support resource discovery and data integration, a homogenous information space of metadata must be constructed. To this aim, a process for homogenization is required in order to tackle interoperability issues at different levels. The PARTHENOS infrastructure implements the homogenization process with the 3M Editor and the PARTHENOS aggregator. Using the 3M Editor, data experts of the Research Infrastructures (RIs) in the consortium can create mappings from the local format of their metadata into the PARTHENOS Entities Model. Those mappings and additional value cleaning functions are then applied by the aggregator in order to generate homogenous metadata records that can be fed into the Joint Resource Registry and exposed via well-known protocols like SPARQL and Solr API.

As of December 2018, the PARTHENOS consortium created twenty-eight official mappings (out of a total of one hundred and fifty-two mappings, mostly created for testing, sampling and getting familiar with the 3M Editor). Sixteen RI users have collaborated in their creation with the support of FORTH.

Figure 30 shows the coverage of the mappings with respect to the PARTHENOS Entities Model. Orange frames indicate that the entity is mapped from the PARTHENOS top-level entities source. Blue frames indicate that the entity is considered in at least one of the mappings of the RIs. The figure shows that mappings cover a good part of the PARTHENOS Entities Model and reflect the actual situation of the RIs and the services they offer. The availability of data curating services highlights the importance of research data in the different disciplines covered by the consortium and the interest of RIs in curation and preservation patterns for research data. The same attention is not reserved to research software: the lack of mappings for services related to software confirm the trend of RIs to not offer dedicated software hosting, curation and delivery services and rely on well-known services managed by third-parties (e.g. GitHub, Amazon and Microsoft). Metadata records about software are not widely used as well, as reported in Section 6.2.6 These facts highlight the importance of initiatives like Software Heritage[19] [17] that are

[19] Software Heritage, https://www.softwareheritage.org

committed to the preservation and sharing of publicly available software source code hosted in a plethora of source code repositories like GitHub, Google Code, and GitLab.

A critical issue in a large-scale heterogeneous aggregation endeavour, as pursued in PARTHENOS, is the quality of the (meta)data. In order to support machine-assisted (meta)data quality checks, a number of tools have been developed and configured during the PARTHENOS project. The D-NET Metadata Inspector proved to be beneficial to infrastructure managers to verify the outcome of the transformations, but more specific tools capable of exploiting semantic information in the mapped metadata records were required. Thanks to the PARTHENOS Discovery Tool and SparqLaborer (see D6.4), we have automatized the execution of SPARQL queries to analyse the aggregated PARTHENOS data space. Interesting quantitative counters have been selected and an analysis on the quality of the PARTHENOS metadata has been presented in Section 6. Such analysis may be useful as starting point for future work on data curation on the PARTHENOS aggregated data or can be used as a reference by projects/initiatives willing to realize an aggregative metadata infrastructure using the PE model (or any RDF/XML model) as common data model.

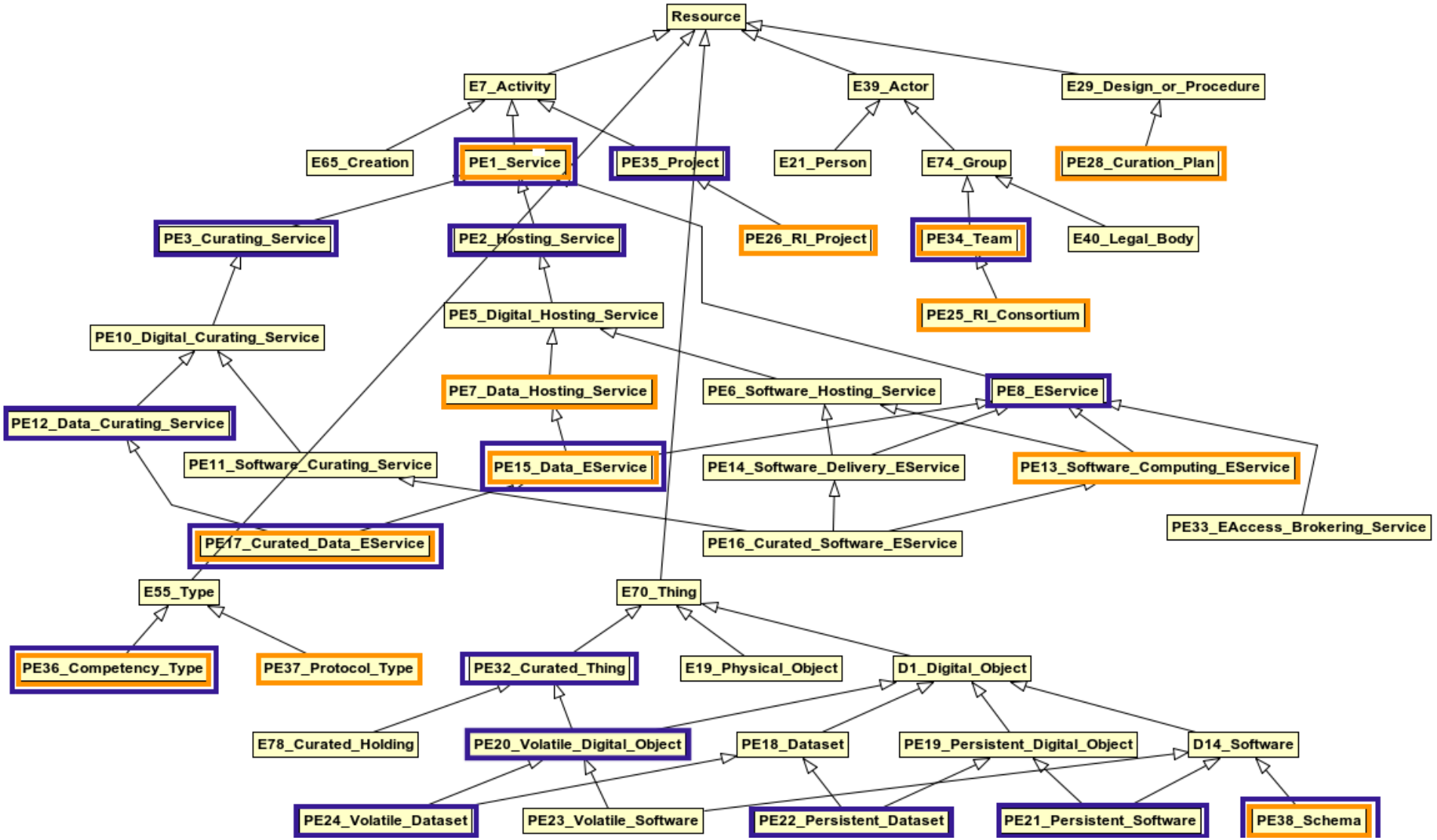


Figure 33 Coverage of the mappings with respect to the PARTHENOS Entities Model

8. References

[1] Manghi, P., Candela, L. and Pagano, P. (2010) 'Interoperability patterns in digital library systems federations', Proceedings of the 2nd DL.org Workshop on Making Digital Libraries Interoperable: Challenges and Approaches, in conjunction with ECDL 2010, ISTI-CNR, Glasgow, Scotland, UK.

[2] Bruseker, G., Doerr, M., Theodoridou, M. (2017). "D5.1 Report on the common semantic framework - Draft". Available at: https://goo.gl/dqLWXM.

[3] Bardi, A., Manghi, P., & Zoppi, F. (2014). Coping with interoperability and sustainability in cultural heritage aggregative data infrastructures. International Journal of Metadata, Semantics and Ontologies, 9(2), 138-154.

[4] Stasinopoulou, T., Bountouri, L., Kakali, C., Lourdi, I., Papatheodorou, C., Doerr, M. and Gergatsoulis, M. (2007) 'Ontology-based metadata integration in the cultural heritage domain', in Goh, D-H.L., Cao, T-H., Sølvberg, I. and Rasmussen, E. (Eds): Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, Springer, Berlin, Heidelberg, pp.165–175.

[5] Marketakis, Y., Minadakis, N., Kondylakis, H., Konsolaki, K., Samaritakis, G., Theodoridou, M., ... & Doerr, M. (2016). X3ML mapping framework for information integration in cultural heritage and beyond. International Journal on Digital Libraries, 1-19.

[6] Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., ... & Pagano, P. (2014). The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures. *Program*, *48*(4), 322-354.

[7] 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 *Proceedings of the Seventh Conference on International Language Resources and Evaluation* [LREC2010]. Pp. 43-47.

[8] 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.

[9] 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

[10] 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 [CAC2014]. In Selected Papers from the CLARIN 2014 Conference [CAC2014]. Pp. 36-53.

[11] 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.

[12] 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* [LREC2012]. Pp. 1029-1034.

[13] Gavrilidou, Maria, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Haris Papageorgiou, Monica Monachini, Francesca Frontini, et al. 2012. “The META-SHARE Metadata Schema for the Description of Language Resources.” In, 1090–97. http://www.lrec-conf.org/proceedings/lrec2012/index.html.

[14] Soria, Claudia, Núria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, Nicoletta Calzolari, and others. 2012. “The FLaReNet Strategic Language Resource Agenda.” In LREC, 1379–86. http://lrec.elra.info/proceedings/lrec2012/pdf/777_Paper.pdf.

[15] Bruseker, G., Doerr, M., Theodoridou, M. (2018). “D5.5 Report on the common semantic framework”. Available at: https://goo.gl/mQV3Mf

[16] A. Bardi, G. Bruseker, M. Durco, M. Kemps-Snijders, M. Lorenzini, M. Theodoridou (2017). “D6.2 report on services and tools”. Available at: http://www.parthenos-project.eu/Download/Deliverables/D6.2_Report_on_services_and_tools.pdf

[17] Jean-François Abramatic, Roberto Di Cosmo, and Stefano Zacchiroli. 2018. Building the universal archive of source code. Commun. ACM 61, 10 (September 2018), 29-31. DOI: https://doi.org/10.1145/3183558