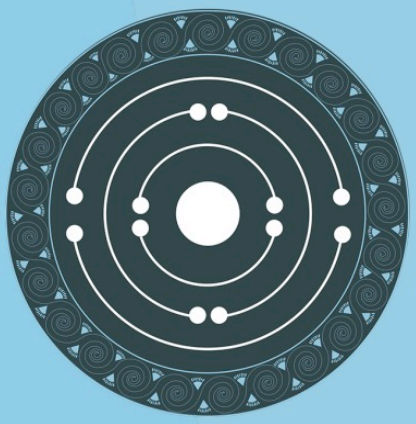

PARTHENOS

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

D5.4 Report on the Integration of Reference Resources

PARTNER(s) FORTH, CLARIN D, CNR

DATE 31/10/2017

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 the Integration of Reference Resources

Deliverable Number	D5.4
Dissemination Level	[PUBLIC with CC-BY distribution]
Delivery date	31 October 2017
Status	Final
Author(s)	George Bruseker Felix Helfer Martin Doerr Maria Daskalaki Carlo Meghini

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	5.4
Document title	Report on the Integration of Reference Resources
Author(s)	George Bruseker, Felix Helfer, Martin Doerr, Maria Daskalaki, Carlo Meghini
Dissemination level/distribution	PUBLIC with CC-BY distribution

Document History

Version/date	Changes/approval	Author/Approved by
V 0.1 09.01.17	TOC	George Bruseker
V 0.2 02.08.17	First draft	George Bruseker
V 0.3 04.09.17	Content enrichment	Felix Helfer, George Bruseker
V 0.4 02.10.17	Content revision	Martin Doerr, Maria Daskalaki
V 0.5 28.10.17	Content revision	Carlo Meghini
V 0.6 30.10.17	Formatting and Editing	Maria Theodoridou
V 0.7 30.10.17	Formatting and Editing	Sheena Bassett
Final 31.10.17	Final edit	George Bruseker

Table of Contents

Table of Figures

Table of Tables

1. Executive Summary

Taking up the challenge of creating a Research Infrastructure (RI) enabling integration of data across disciplines involves, at the level of conceptual modelling and mapping, two major intellectual and practical labours. On the one hand, a schema matching activity against a common expression must be achieved in order to render some subset of the available datasets interpretable in a common form. On the other hand, once such schema matching has been achieved, there remains a need for alignment on the level of actual data values. Because of different practice resulting from institutional policy, disciplinary approach and linguistic form, amongst others, data values contained in matched schemas will almost certainly differ, even though they refer to the same things. Before the desired interoperability of datasets can be achieved, a strategy for binding and connecting these various data forms together must be adopted and enforced. Desirable interoperability at the level of data values means that end users of the system will be able to use common vocabularies to query to and discover results from source systems implementing widely varying input systems or, inversely, start from variant forms of vocabulary and be delivered results from a normalized form. This work then has to do with vocabulary management and the ability to manage and connect a plethora of different but related vocabularies across disciplinary and linguistic boundaries. It also has to do with identifying best practice in the research infrastructure environment. Heterogeneity of data is a fact of the information space which should be approached as a situation to be managed (Plato, 1921), not eliminated. Nevertheless, there are identifiable information categories of common use where there are good reasons to seek common vocabularies which all participants in a RI can appeal to and use, rather than each making their own standard. In doing so we can reduce information fragmentation but also support and implement well structured vocabularies for categories of things of common interest and/or build such best practice standard vocabularies where there is a demonstrable lack in the field.

This document forms an interim report on the activities within PARTHENOS WP5 in collaboration with WP4 to adopt such a vocabulary management strategy and to identify high level standardized vocabularies for use in the data integration activities into the Joint Resource Registry carried out by WP6. This document first outlines the basic strategy adopted for vocabulary management in the PARTHENOS project and then provides an analytic presentation of the vocabularies deemed necessary for management of data at

the level of the RI. It then goes on to look at the specific research activity to find and identify the best available standards for vocabularies at the level defined by the PARTHENOS Entities, the management and tracking of information regarding datasets, software, services, projects and people, as the set of objects of interest for management at an infrastructural and cross-infrastructure level. The intent at this level is to enable an understanding of available resources and their interrelations in order to facilitate information management at a high level, making strategic decisions with regards to what information may be brought together in useful bundles in order to enable large scale research projects through Virtual Research Environments for example. In the final version of this report, we will look at vocabularies of interest for matching and integrating at the content level across Research Infrastructures representing the different constituent communities of the PARTHENOS project, e.g.: History, Linguistic Studies, Archaeology, Heritage and Applied Sciences and Social Sciences.

2. Vocabulary Management Strategy

2.1. The Problem

The activity of classifying and distinguishing groups of things within the world is a basic element of intellectual activity that leads, historically, to the elaboration of a plethora of terminological systems for describing the world around us. Both at a folk level and at the scientific level, human beings constantly partition the world intellectually into various classes of things by which to separate and distinguish collections of items of interest. Such classes are used, in turn, to build up a discourse over the groups of items so designated. This discourse, again, may have purely practical aims, e.g. separating the edible from the inedible, where the method is often tacit, or for scientific purposes, e.g. the taxonomic differentiation of biological species, where more or less explicit methods guide such processes. The plurality of classificatory systems and their recalcitrance to a reduction to a uniform and consistent classificatory *lingua universalis* is well known. Depending on the function that a classificatory system was devised for - the contextual goals that it was set out to achieve - its division of the world into this or that set of categorical units will reflect a particular intention and interest towards the world. This interest limits and focusses the different significant perceptible features of the world by which criteria for dividing up the the world into significant units of discourse is carried out. It is a consequence of this phenomenon that there is a general pattern of incommensurability amongst classificatory systems which makes the effort to unify the different visions of the world extremely difficult to achieve with rigour and fidelity to the original system. Such incommensurability at the level of detail is as typical for folk systems of classification (e.g. varying kinship systems) but also at scientific level (e.g. classificatory systems in biology and physics).

The problem of the method and very possibility of providing harmonized and correct classificatory systems which are able to mitigate if not solve this heterogeneity problem is one that has a deeply rooted and global philosophical history. In the Western tradition, we can refer to the efforts of Plato in *the Sophist (Plato, 1921)* to communicate a method of correct division of things which stands as an early effort to conceptualize and address this difficulty in the Western tradition. The dialogue outlines a method to effect division or *diairesis* over an area of concern, in order to find the correct and real categories of thing on the basis of which to have an epistemically valid discourse. Such early efforts at class

definitional rectitude encountered many philosophical challenges from competing schools. Perhaps no critique was as famous as the amusing episode in which Diogenes offered a ‘plucked chicken’ as an instance of man according to the classification arrived at by method of *diairesis* defining man as a ‘featherless biped’. Just as lively a debate occurred in other philosophical traditions with very different founding conditions. One may reference, notably, the work of Zhuang Zi (Zhuangzi, 2003) and his exploration of the epistemic problematics of discovering the correct division of the world - traditionally noted in defiance of the work of Kong Zi on executing a ‘rectification of names’ (Confucius, 2016) - where he famously describes the intuitive effort of the expert butcher to find the joints of the animal requiring a deprogramming of pre-existing rules and thoughts in order to follow the ‘joints of the world’ itself.

The problem of classificatory heterogeneity, however, cannot be relegated to the dustbin of history but represents an on-going and diachronic problem. This problem takes on a new urgency and interest in an information age, where the production of systematic information structures is no longer the realm of a fantastic technocratic dream of Socrates but a lived everyday reality and even environment for human beings. Information systems allow ever greater amounts of empirical data to be generated by scientists and scholars deploying an ever wider array of classificatory schemas in order to pursue their research. Historical, linguistic and methodological differences mean that there are ever larger amounts of datasets that refer to real world entities which may fall in the same general domain of interest but which cannot easily be accessed by potentially interested parties due to the fragmentation of classificatory systems. In facilitating an ever greater production of data, information technologies have not solved the problem of the babel of taxonomies but rather made it ever clearer by facilitating more production of expert data incorporating masses of heterogenous classificatory systems.

Within the context of a research infrastructure, and even more so within the context of a multi-disciplinary research infrastructure such as PARTHENOS, adopting a solution for the harmonization of such vocabularies is paramount. Without a long term strategy, even if temporary alignments of data can be undertaken, the continuous generation of new classifications in accordance with the consequence of new results and the opening of entirely new research fields will result in an obsolescence and ossification of information over time. Establishing common, acceptable standard vocabularies in any research

discipline is difficult and contentious. Such projects are long term investments which offer the benefit of compatibility and harmonization of results but at the risk, if carried out incorrectly, to stifle research by establishing inflexible canonical classifications unable to take into account new categorizations which may reveal new information about the world under study. The situation within the PARTHENOS project is further exacerbated by the fact that it aims not to serve an individual disciplinary community but rather to support research across disciplines and thus enable question posing and answering beyond traditional disciplinary boundaries. Such an ambition means that a resort to disciplinary best practices is not even an option. Rather, we are compelled to look for systematic methodological solutions that go beyond traditional disciplinary boundaries.

2.2. Previous Solutions

In line with the spirit and aim of PARTHENOS as a catalyzing action for finding common solutions and best practices from existing and well established Research Infrastructures, the effort to meet this problem begins from existing research available within the network. In particular, the DARIAH project[1] has had as a specific focus the creation of a solution to vocabulary heterogeneity within the humanities. This research focus has resulted in the creation of a Thesaurus Maintenance WG[2] that deals specifically with this topic on a continuous basis. The research of this WG stands as an important starting point for the PARTHENOS project which can take up its findings and principles and generalize them for the members of the entire PARTHENOS consortium.

Particularly in the work, “Thesaurus Maintenance Methodological Outline” (Thesaurus Maintenance Working Group, VCC3, DARIAH EU, 2015) a rigorous and practical methodological approach for addressing this problem as an informatics question is laid out.

The vocabulary management problem is not, as we have seen, new and has been addressed by a number of different generic information management strategy types historically. The effort to effectuate a practical *lingua universalis* of classificatory systems is in effect an agenda to build a vocabulary of vocabularies, a meta-vocabulary to bind them

[1] http://www.dariah.eu/
[2] http://www.dariah.eu/activities/working-groups/thesaurus-maintenance/

all. The authors of TMMO outline meta-vocabulary management as a specific problem of modern information management, and before proceeding to present their own solution, analyze previous efforts to meet the problem and their relative strengths and weaknesses, as a basis from which to learn and build. They analyze three major types of strategy that have been used to address this problem: the exhaustive subject classification system, taxonomic subject classification and the centralized controlled authority approach.

The exhaustive subject classification approach is evidenced in such standards as the Library of Congress Subject Heading[3] system. Able to draw on the collective cataloguing experience of thousands of libraries, LCSH creates an enormous vocabulary tree containing information from all different branches of science and scholarship. This provides a fantastic resource which has a clear empirical basis of enabling the discovery of many resources. Since its classification, however, draws from the disciplines themselves which in turn classify with regards to their own specific domain of interest, the LCSH, while providing a category for virtually anything, cannot provide a hierarchical synthetic view of overlapping areas of interest. That is to say, one has to already know where one should be searching and for what in order to be able to find it. Serendipitous discovery of related but disciplinarily distinct results is not facilitated. Another disadvantage to the LCSH type approach is that it necessarily treats classifications as static and relatively slow changing systems, whereas in a research environment classifications are fluid and changing dynamically, deployed as hypotheses and reformed according to empirical results. The ability to support such dynamic vocabularies while relating them to better known terms remains unaddressed by an LCSH type approach, perhaps largely because this functionality largely falls outside of the remit of libraries regardless.

The Dewey Decimal System[4], also devised within the library context, can be seen as a more promising tool for a meta-vocabulary since it takes a principled position on the hierarchical organization of information into a universal classificatory regime. That being said, it also proves inadequate to serve as a meta-vocabulary of the kind needed by a research environment. In part, this holds for the same reasons that LCSH is not appropriate. It is not designed to support rapidly changing hypothesis-style terminologies such as are deployed on a regular basis by scholars and scientists as they build to

[3] https://www.loc.gov/aba/publications/FreeLCSH/freelcsh.html
[4] https://www.oclc.org/en/dewey/features/summaries.html

conclusions. The methodological reason that it is unfit for purpose as a top level meta-vocabulary is that, while it adopts hierarchical semantic organization of data, it does not have an ontologically oriented methodology for creating these divisions, but rather builds levels of disjoint partitions from properties selected arbitrarily for the purpose of partitioning. This results in a system that is systematically incommensurable with any other sequence of partitioning, and may force arbitrary classification of things. This methodological shortcoming, with regards to the function of a top level meta-vocabulary, is significant because it means that it potentially fails in important integrations of relevant information that could be achieved through a systematic approach to developing the hierarchical semantics between classes.

Lastly, it is worthy to point out the work of the HEREIN project,[5] which aims to establish a central authority to gather multi-disciplinary vocabularies and organize them into a top level meta-vocabulary. While gathering inputs from an impressive range of partners with important geographic and linguistic distribution, the project is weighed down by its own successes. Centrally managing and deciding on the semantic clarification of such a plethora of vocabularies is a task that is unsustainable for a single central entity and especially for a project to undertake. The work of maintaining such a vocabulary is enormous. The ability to support a continuous updating and integration of data is required both at a technical but as much at a social scientific level, in order to maintain the relevance and use of the system. The constant production of new vocabularies by scientists and scholars requires a high degree of flexibility and a methodology that enables a decentralization of this task through the application of well known and public principles by which to effectuate the integration.

The above analysis of the existing successes and limitations of high level efforts to integrate systemic classificatory knowledge served as the ground from which the DARIAH research group elaborated a new strategy and methodology for devising such a system to allow practical data integration using a principle methodology for creating semantically coherent classificatory hierarchies in a distributed environment.

[5] http://www.herein-system.eu/

2.3. The Back Bone Thesaurus Solution

The Back Bone Thesaurus solution is documented most recently in a DARIAH report by the Thesaurus Maintenance Working Group's entitled, "A model for sustainable interoperable thesauri maintenance" (Thesaurus Maintenance Working Group, VCC3, DARIAH EU, 2016). This document outlines both the basic method adopted and the results heretofore of a top level meta-vocabulary. It is inspired by the UMLS Metathesaurus.[6]

The authors identify five basic requirements for the generation of a sustainable and effective meta-vocabulary: the adoption of a semantic approach, a clear method to semantic division, creation of top level terms based on bottom up analysis of existing classificatory systems, open ended development of complete vocabulary including top terms and the ability to carry out this work as a distributed collective project. In greater detail this entails the following. The semantic approach of building a hierarchy of terms that spans disciplines and is based on the real world referent of terminologies is necessary to meet the integrative functionality envisioned for a meta-thesaurus. An approach that cannot critically analyze and integrate classification systems into a general system will not deliver the data integration capacity that a meta-vocabulary promises. It is not enough, however, to engage in a semantic method for generating top level terms of the meta-vocabulary but there must be an explicit and communicable principle for generating top level classes and the distinctions that they entail and then impose back into the overall collection of classificatory systems. The methodology that the WG proposes to achieve this is a bottom up adduction of higher level meanings through the analysis of a broad body of classificatory systems as evidence. That is to say, an analysis must be done of classificatory systems that one wants to integrate and from these draw candidate top level terms for the meta-thesaurus. The top level terms should not be imposed by means of an *a priori* theory as is done in the Dewey Decimal System, but rather must be discovered through an analysis of existing sources and the development of a clear understanding of their common referents in order to be able to provide a functional and clear specification of top level terms. It is on the basis of this that high level classes with explicit scope notes that indicate the nature of the kind of classification system they entail can be developed. In

[6] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

order to meet the needs of research, however, this bottom up approach must be left fundamentally open. The derived top level classes come to serve as hooks upon which sufficiently described vocabularies can be hung in order to create a semantically consistent hierarchy. That being said, the possibility remains open that with the addition of new areas of knowledge either from new research or from the integration of domains not yet covered, the hierarchy would expand and be enriched either at the top level hierarchy or at any element below. This points to the final key element to the methodology propounded by the WG which is that the construction of the BBT should be carried out by a distributed group that is able to comment and organize the thesaurus without a central authority but with the clear methodological principles elaborated within the WG documents as authority in order to produce the top level categories.

Thus, in effect, what is proposed is a federation of vocabularies that are brought together through an open ended backbone and which are open to tighter integration on an as needed basis. Such need arises organically according to the mutual interests of groups of researchers to create integrated classifications of more specific resources/objects of research. The BBT strategy allows for this open ended extension by offering a declared method for building new branches in the tree allowing all *groups to follow the same method* even on lower levels of generalization and in very specific communities of practice.

The top level model proposed by DARIAH at this point consists of the following facets and hierarchies:

activities

- disciplines
- human interactions
- intentional destructions
- functions
- other activities

natural processes

- natural disasters
- geneses

materials

material things

- mobile objects
- built environment
- physical features
- structural parts of material things

types of epochs

conceptual objects

- symbolic objects
- propositional objects
- methods
- concepts

groups and collectivities

roles

- offices
- roles of interpersonal relations

geopolitical units

The basic idea of the use of the BBT from the user side is to find places within the top level hierarchy to which the top-terms or high-level terms of their classificatory system belong and properly hang them into the overall structure. It may be that a classificatory system is made up of terms in one hierarchy that pertain to multiple distinct generalizations in the BBT. Even then BBT is able to handle integration in a logically consistent way. Parts of a vocabulary can be split across multiple high level facets in the BBT. Where a candidate vocabulary is a flat list with no declared top term, it may be necessary to introduce auxiliary intermediate generalizations in the source classificatory system which would then, in turn, link into the BBT in a semantically consistent way. Following this linking process, terms from distinct classificatory systems referring to the same real world areas of interest can be searched together with other relevant classifications via the root in the class tree. This will enable benefits to the end user searching for information who will be able to use different classificatory systems for the same general class of things. Eventually, this can also enable the curation of such classification systems into common classificatory systems insofar as they show true compatibility and commonality of use. This functionality can be supported by a SKOS enabled vocabulary editor. Within the context of DARIAH and with continued support of PARTHENOS, the Themas tool[7] is one tool which

[7] http://www.ics.forth.gr/isl/index_main.php?l=e&c=243

could enable this functionality. Where the end user cannot find an appropriate high level facet or hierarchy under which to place terms of their classificatory system, a process of discussion of extension and expansion of the BBT itself should be launched. This functionality is presently enabled by the Submission and Connect Management Tool (Thesaurus Maintenance Working Group, VCC3, DARIAH EU, 2017) built and maintained by DARIAH.

2.4. The Back Bone Thesaurus and PARTHENOS

The information management strategy of PARTHENOS is based on the PARTHENOS Entities Model which is used as a common ontology, based on CIDOC CRM, in order to integrate data arising from Research Infrastructure registries regardless of disciplinary interest. It enables integration of data at the level of schema matching, bringing data encoded in miscellaneous schemas into a sufficiently general schema that they are globally queryable according to a common structure. This, however, achieves only part of the data integration picture since, for data to be tightly integrated, it must make use of the same or compatible structured vocabularies for expressing data values that are susceptible to standardization. Such data values are usually ‘type’ fields such as ‘subject’ or ‘material’ or ‘object kind’ etc. Additional data values that are susceptible to standardization include such data as is recorded in field types such as ‘period’ which relates a data item through some semantic relation to a, hopefully, well known periodization structure. Likewise, data values encoded in fields for expressing information such as ‘place’ which refer to well known geographic units can be standardized against well known gazetteers. This standardization or matching of vocabularies ensures against basic errors in data entry but also creates common terms of reference for classifying and referencing real world items. Such classification goes into detail that goes beyond the level of detail needed to generate a common semantic model such as the PARTHENOS Entities Model, but is a necessary correlate work that must be matched to the ontology in order to create the tight data integration that should be delivered to end users in order to facilitate their ability to find the resources they are looking for, be those datasets, software, services, actors or others.

The Back Bone Thesaurus Solution provides a high level means of carrying out the complicated task of aligning relevant vocabularies under common roots in order to allow integrated cross vocabulary search even when different local vocabularies are adopted by different scholars and groups for similar classes of referents. The solution proposes, however, that there is a well known set of existing standards relative to the domain of work in question.

The PARTHENOS project via its proposed semantic model represents, initially, a level of data management that is a step away from the content itself and has to do with the management thereof. Since the attention of researchers is primarily on their content, standardized vocabularies for such meta-metadata is not well known and well established. Harmonization at this level, however, directly affects the first stage of integration in the PARTHENOS strategy. This, therefore, establishes an initial need to identify the relevant potential vocabularies for use together with the model to provide a common representation of cross-disciplinary research infrastructure data. This represents an important first stage of research to establish a not-yet existing collection of reference standard vocabularies for this type of meta-metadata.

Before this initial setup is achieved, a further, more ambitious harmonization of vocabularies relevant to creating a cross-disciplinary harmonization of vocabularies relative to the content of study cannot yet be initiated. This next step will require a parallel research process to be derived from WP4 in order to identify the important vocabularies for PARTHENOS partners and then work on their integration into the BBT system.

Therefore, the strategy for vocabulary harmonization undertaken in the context of T5.3 will be be executed in two stages. The first stage of this activity will focus on the identification of the necessary vocabularies to support data values standardization relative to the PARTHENOS Entities Model. These will be matched to the high level categories in the BBT. The second stage of this task will gather some of the important related standard vocabularies of greatest relevance to the user communities of the member RIs of PARTHENOS and perform the same matching operation, aiming to make these vocabularies tractable to cross-disciplinary search across data integrated on the content level.

The adoption of the Themas and Submission Connection Tool or similar tools within PARTHENOS would enable a sustainable continuous development of this activity and provide a valuable cross-disciplinary research resource. In turn, the activity of PARTHENOS in adopting the BBT is consistent with the strategy and methodology proposed by the Thesaurus Management WG. It can provide additional new empirical material in the form of well known and identified classificatory systems from which to enrich and expand the BBT beyond the initial scope of humanities research to a cross-disciplinary resource capable of facilitating research beyond disciplinary boundaries.

3. Structured Vocabularies for PARTHENOS Entities

This section describes the general research process engaged for the identification of relevant well defined vocabularies to be used in relation to the entities described by the PARTHENOS Entities Model.

3.1. Joint Research Registry, PE and Vocabulary needs

The PARTHENOS Entities Model (PEM) itself represents a product of research over the data organization practices of Research Infrastructures based on the work of T5.3 of the PARTHENOS Project. It provides a semantic model of the world of data management for scientific and scholarly research with a focus on connecting researchers to the producers and maintainers of data in order to be able to identify mutually relevant resources for exploitation within collaborative Virtual Research Environments by the integration of data into common formats and their investigation through traditional and digital methods of research. The process and outcome of developing this model is described in D5.1 of the PARTHENOS Project. The semantic model itself, however, is used particularly in PARTHENOS in order to build a Joint Research Registry which adapts the model in order to build a common, cross RI registry of resources at a high level. The process and initial outcome of the development of this registry is described in D5.2 of the PARTHENOS Project. The Joint Resource Registry is initially populated by a rich description of the top level Actors, Datasets, Software, Services and Projects which make up the PARTHENOS community. It is then enriched through the integration of data on the resources availed in

each RI which is mapped to PEM using the X3ML Toolkit Suite.[8] It is at this point that the need for a set of standardized vocabularies shows itself. While integration is achieved at the schema level, there are a number of distinct classificatory schemes deployed by each RI for the same objects either implicitly or explicitly that must be harmonized in order to provide a usable query environment within the JRR.

Since, as mentioned above, the types of entities being classified by such vocabularies belong not to the subject of research of scholars themselves but apply to the processes of maintaining and preserving such resources, there is a lack of well known and identified standard vocabularies to which to harmonize. Therefore, the first research with regards to building integrated reference resources, is to find appropriate reference resources for integration. In what follows, we will describes the PARTHENOS Entities Model as implemented as an application profile within the Joint Resource Registry, what standard vocabularies it entails and the standards that were identified to meet these needs. Finally, we will look at an initial linking of these standard vocabularies into the BBT meta-vocabulary.

3.2. PE Minimal Metadata Information Types and their Standardized Vocabulary

The PARTHENOS Entities are structured in order to be able to build - or create data translations from/to - information systems that aim to document information resources and the activities of holding, curating and managing these resources as well as the contexts of these activities, e.g. projects. There is a special focus on enabling the connection of resources to the actors responsible for and interested in them. Translated from a conceptual model into an information architecture, we can speak of the elaboration of an application profile that suggests a minimal level of data management necessary in order to support such a data management goal. The elaboration of such an application profile has been executed in PARTHENOS as the 'minimal metadata' set (defined in D5.1). In this section, we will highlight chief elements of this application profile and where they create a demand for standardized vocabularies in order to move beyond schema matching to integrated ways of classifying and identify individual resources that will enable tightly integrated and highly queryable data.

[8] http://www.ics.forth.gr/isl/index_main.php?l=e&c=721

Each part of the information profile intends to help ask and answer certain basic questions that one would like to be able to ask of a dataset on this information space and receive robust answers. We will present the data model suggested for significant high level entities in the model and then indicate the data elements which are candidates for the application of a standardized vocabulary. We will then elaborate on the vocabularies selected for use in PARTHENOS and evaluate their relative merits.

We will look at profiles for: Projects, Services, Datasets, Software and Actors and the vocabularies they require. For each entity type we will look at their general intended use and in particular what questions they aim to help a researcher answer. Then we will look at their instantiation as an application profile in an implemented model adopting the PARTHENOS Minimal Metadata recommendations. For each application profile, we will look at the metadata it requires, represent this in a semantic schema and indicate where a control vocabulary is needed and which vocabulary was selected (where such a selection was possible). Where no appropriate vocabulary could be found, we aim to carry the research on in the second phase of T5.3 activity to fill the gaps identified where possible by working with the relevant RIs.

Please note that in the semantic diagrams that follow a colour coding is used to make the reading of the diagrams easier. This coding is as follows:

Colour	General Entity Type
Blue	Temporal Entity
Yellow	Conceptual Entity
Brown	Physical Entity
Pink	Agency Entity
Green	Geometric Entity

Table 1: Color coding of semantic diagrams

3.2.1. Projects

A project in the PARTHENOS Entities model is a long term encompassing activity that gains its existence by the formation of a team that has the will and the capacity to carry it out and retains this existence so long as this team continues to exist with the same aim regardless of its internal composition. It is distinguished as a type of activity by the will to a long term goal into which many activities and provisions of service may belong. A research infrastructure project and a research consortium form specializations of the general notion of project and team respectively. The documentation of a project provides a general context for understanding under what conditions services were enacted, datasets and software produced and who was involved.

With the project classes we wish to support answering the following types of questions to the information model:

- What is it? (Identity)
- What activities does it support? (Part/Whole)
- When was it available? [Access]
- Who carried it out? (Agency)

3.2.1.1. Project

The minimal metadata set profile proposed for Project is as follows:

Label	Mandatory(?)	Field Type	Description
ID	Y	String	The identifier used to indicate the project.
Type	Y	Controlled Vocabulary [1]	A typology for classifying the kind of project.
Title	Y	String	The name by which the project is known or referred to.
Description	N	Long Text	A textual description of the service
Supports	N	Link	Link to activities and services supported by the project.
Project Duration	N	Date	The duration of the project.

Maintaining Team	Y	Link	Link to the team maintaining the project.

Table 2: PE35 Application Profile Minimal Metadata Configuration

The semantically encoded expression of the minimal metadataset for PE35 Project is as follows:

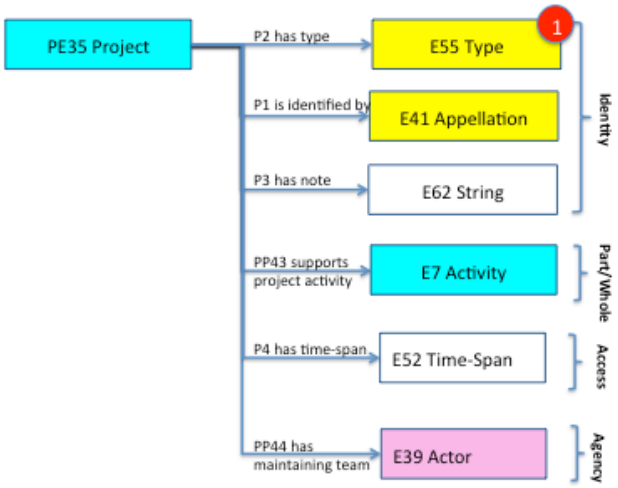

Figure 1: PE35 Project Minimal Metadata Application Profile Schema

The PE35 Project minimal metadata application profile makes reference to one field which require standardization according to common vocabularies. The following table summarizes the final results of chosen standards relative to these fields.

	Min Metadata Field Name	Path	Role	Recommended Standard
1	Type	PE35→P2→E55	Identity	None

Table 3: Recommended standards for PE35 Application Profile

3.2.2. Services

Services are a central notion within research infrastructures, since the goal of such consortia is not limited to the amassing of a collection of data but rather to the provision of a series of long standing activities which form a physical and social infrastructure wherein a community of researchers can dynamically engage and build on each other's research, experience and outcomes. Services are defined in the PARTHENOS Conceptual Model as the willingness and ability to do something for someone else. They are a kind of long standing activity that can be activated by users/customers of RIs. Services as activities

gain identity through the actors who offer them and the kind of service offered as well as the services actual and potential outputs. The notion of service is what binds products such as datasets of software to actual institutions and practices, allowing one to understand their provenance and communicate with the people behind such products. Therefore, it is fundamentally necessary to capture information about the service within the context of Research Infrastructure management.

In the PARTHENOS Entities model a general class is declared for services to capture any instance of service in general. The model then makes three high level divisions between Hosting Services, Curating Services and E-Services. These are particularly of relevance within Research Infrastructures. Hosting Services, on the one hand, have to do with the offer and ability to hold and give access to an object, without doing anything to it. Curating Services are an entirely different activity. They have to do with the willingness and ability to manage an aggregate of things according to a plan. E-Services have to do with the offer of an electronic service that allows an automated access through a network to a computing environment capable of delivering services automatically. These three service classes are deployed through multi-inheritance in the conceptual model to build the possible derivations of general kinds of services. This allows both a granular depiction of complex services that involve both hosting and e-services (e.g. a web based hosting service) but also general hosting services (e.g. the temporary storage of art by a museum for some group).

Knowledge of services and their capacities are crucial to members of Research Environments in order to have an understanding of the resources available to them.

With the service classes we wish to support answering the following types of questions to the information model:

- What is it? (Identity)
- What can it do? (Identity)
- What is it part of? [Service/Project] (Part/Whole)
- When is it available? [Access]
- What conditions are there to use? (Access)
- What technical conditions are there to use? (E-Access)
- What does it manage? (Stewardship/Curation)

- How does it manage what is manages? (Stewardship/Curation)
- What does it hold? (Hosting Info)

Translated into application profiles for execution in an information system we can look at three basic profiles: Service, Curated Data E-Service and Curated Software E-Service. The former provides a profile for the description of any service in general. The latter two provide a minimal dataset for monitoring in the case of services that combine the offers of hosting, curating and offering an e-service for access, in the one case for datasets and, in the other, for software.

3.2.2.1. Service

The minimal metadata set profile proposed for Services is as follows:

Label	Mandatory (?)	Field Type	Description
ID	Y	String	The identifier used to indicate the service.
Type	Y	Controlled Vocabulary [1]	A typology for classifying the kind of service.
Title	Y	String	The name by which the service is known or referred to.
Description	N	Long Text	A textual description of the service
Competency	Y	Controlled Vocabulary [2]	The function of a service.
Is/Was Part of	N	Link	The service of which this service forms a part.
Supported by	N	Link	The project which supports this service.
Declared Begin/End	N	Date	The date that the service providers indicates as the beginning and/or ending of the offer of the service
Conditions of Use / Rights Type	N	Controlled Vocabulary [3]	Indicate the type of conditions that the use of this service are subject to (Open Access, Open Access - required registration, license-based, on request, embargo)
Conditions of	N	Link	Link to the actual text outlining conditions

Use / Rights Text			of use
Provided by	Y	Link	The actor that provides the service.
Contact Person	N	Link	The contact person for this particular service.
Communication Address	Y	String	The contact address for this contact person, any type.
Communication Address Type	N	Controlled Vocabulary [4]	The type of the contact address provided.

Table 4: PE1 Application Profile Minimal Metadata Configuration

The semantically encoded expression of the minimal metadataset for PE1 Service is as follows:

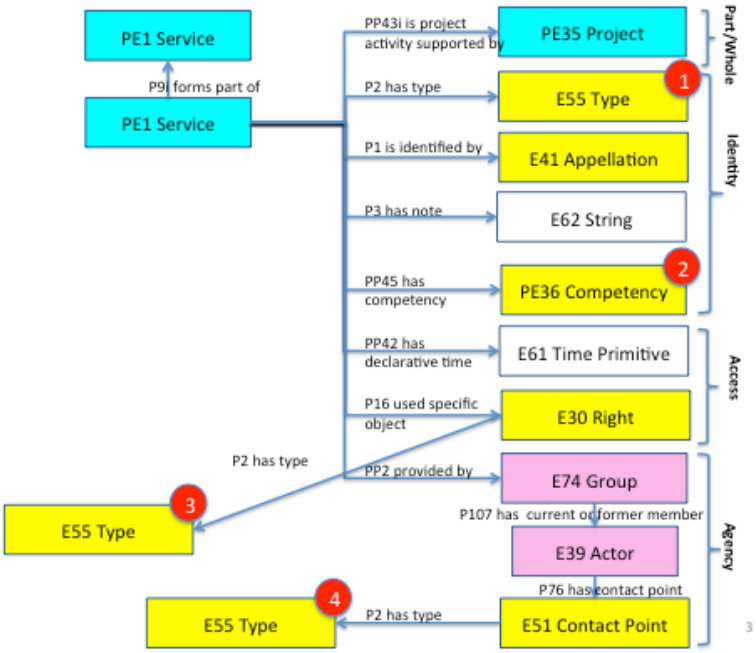

Figure 2: PE1 Service Minimal Metadata Application Profile Schema

The PE1 Service minimal metadata application profile makes reference to four fields which require standardization according to common vocabularies. The following table summarizes the final results of chosen standards relative to these fields.

	Min Metadata Field Name	Path	Role	Recommended Standard
1	Type	PE1→P2→E55	Identity	None

2	Competency	PE1→PP45→PE36	Identity	PARTHENOS Service Competency List
3	Conditions of Use / Rights Type	PE1→P16→E30	Access	PARTHENOS Rights List
4	Communicatoin Address Type	PE1→PP2→E74→P107 →E39→P76→E51→P2 →E55	Agency	CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details

Table 5: Recommended standards for PE1 Application Profile

3.2.2.2. Curated Data E-Service

The minimal metadata set profile proposed for Curated Data E-Services is as follows:

Label	Mandatory (?)	Field Type	Description
ID	Y	String	The identifier used to indicate the service.
Type	Y	Controlled Vocabulary [1]	A typology for classifying the kind of service.
Title	Y	String	The name by which the service is known or referred to.
Description	N	Long Text	A textual description of the service
Competency	Y	Controlled Vocabulary [2]	The function of a service.
Is/Was Part of	N	Link	The service of which this service forms a part.
Supported by	N	Link	The project which supports this service.
Declared Begin/End	N	Date	The date that the service providers indicates as the beginning and/or ending of the offer of the service
Conditions of Use / Rights Type	N	Controlled Vocabulary [3]	Indicate the type of conditions that the use of this service are subject to (Open Access, Open Access - required registration, license-based, on request, embargo)
Conditions of Use / Rights Text	N	Link	Link to the actual text outlining conditions of use

Provided by	Y	Link	The actor that provides the service.
Contact Person	N	Link	The contact person for this particular service.
Communication Address	Y	String	The contact address for this contact person, any type.
Communication Address Type	N	Controlled Vocabulary [4]	The type of the contact address provided.
Online Access Point	Y	String	URL where the service can be accessed by a client application
Online Access Point Type	N	Controlled Vocabulary [5]	Type of access point provided
Protocol	Y	Link	The access protocol, considered as a form of software, which the E-Service invokes
Protocol Type	N	Controlled Vocabulary [6]	Documentation of access protocol type when particular version of software not referenced
Protocol Parameters	N	Link	Link to the schema of parameters to use in the protocol invoked
Curates Volatile Dataset	N	Link	Reverse link from the dataset that is curated by this service.
Curation Plan	N	Link	Link to the curation plan guiding the dataset curation provide by this service.
Curation Plan Type	N	Controlled Vocabulary [7]	Link to the controlled vocabulary of curation plan types for e-curation of datasets.
Hosts Dataset	N	Link	Reverse link from the dataset that is hosted by this service.

Table 6: PE17 Application Profile Minimal Metadata Configuration

The semantically encoded expression of the minimal metadataset for PE17 Curated Data E-Service is as follows:

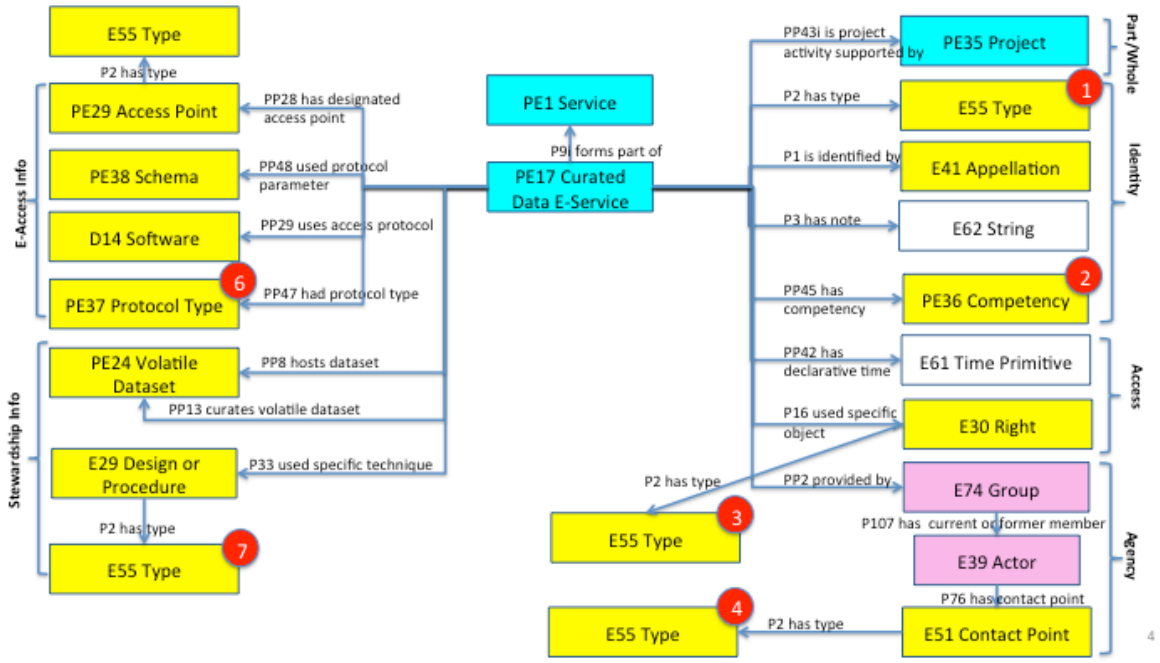

Figure 3: PE17 Curated Data E-Service Minimal Metadata Application Profile Schema

The PE17 Curated Data E-Service minimal metadata application profile makes reference to seven fields which require standardization according to common vocabularies. The following table summarizes the final results of chosen standards relative to these fields.

	Min Metadata Field Name	Path	Role	Recommended Standard
1	Type	PE17→P2→E55	Identity	None
2	Competency	PE17→PP45→PE36	Identity	PARTHENOS Service Competency List
3	Conditions of Use / Rights Type	PE17→P16→E30	Access	PARTHENOS Rights List
4	Communicatoin Address Type	PE17→PP2→E74→ P107→E39→P76→ E51→P2→E55	Agency	CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details
5	Access Point Type	PE17→PP28→PE29 →P2→E55	E-Access	CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details
6	Protocol Type	PE17→PP47→PE37	E-Access	None
7	Curation Plan Type	PE17→P33→E29→ P2→E55	Stewardship	None

Table 7: Recommended standards for PE17 Application Profile

3.2.2.3. Curated Software E-Service

The minimal metadata set profile proposed for Curated Software E-Services is as follows:

Label	Mandatory (?)	Field Type	Description
ID	Y	String	The identifier used to indicate the service.
Type	Y	Controlled Vocabulary [1]	A typology for classifying the kind of service.
Title	Y	String	The name by which the service is known or referred to.
Description	N	Long Text	A textual description of the service
Competency	Y	Controlled Vocabulary [2]	The function of a service.
Is/Was Part of	N	Link	The service of which this service forms a part.
Supported by	N	Link	The project which supports this service.
Declared Begin/End	N	Date	The date that the service providers indicates as the beginning and/or ending of the offer of the service
Conditions of Use / Rights Type	N	Controlled Vocabulary [3]	Indicate the type of conditions that the use of this service are subject to (Open Access, Open Access - required registration, license-based, on request, embargo)
Conditions of Use / Rights Text	N	Link	Link to the actual text outlining conditions of use
Provided by	Y	Link	The actor that provides the service.
Contact Person	N	Link	The contact person for this particular service.
Communication Address	Y	String	The contact address for this contact person, any type.
Communication Address Type	N	Controlled Vocabulary [4]	The type of the contact address provided.
Online Access Point	Y	String	URL where the service can be accessed by a client application
Online Access Point Type	N	Controlled Vocabulary [5]	Type of access point provided

Protocol	Y	Link	The access protocol, considered as a form of software, which the E-Service invokes
Protocol Type	N	Controlled Vocabulary [6]	Documentation of access protocol type when particular version of software not referenced
Protocol Parameters	N	Link	Link to the schema of parameters to use in the protocol invoked
Curates Volatile Software	N	Link	Reverse link from the dataset that is curated by this service.
Curation Plan	N	Link	Link to the curation plan guiding the dataset curation provide by this service.
Curation Plan Type	N	Controlled Vocabulary [7]	Link to the controlled vocabulary of curation plan types for e-curation of datasets.
Hosts Software	N	Link	Reverse link from the dataset that is hosted by this service.
Delivers Software On Request	N	Link	Reverse link from Software that the service offers for download deliver.

Table 8: PE16 Application Profile Minimal Metadata Configuration

The semantically encoded expression of the minimal metadataset for PE16 Curated Software E-Service is as follows:

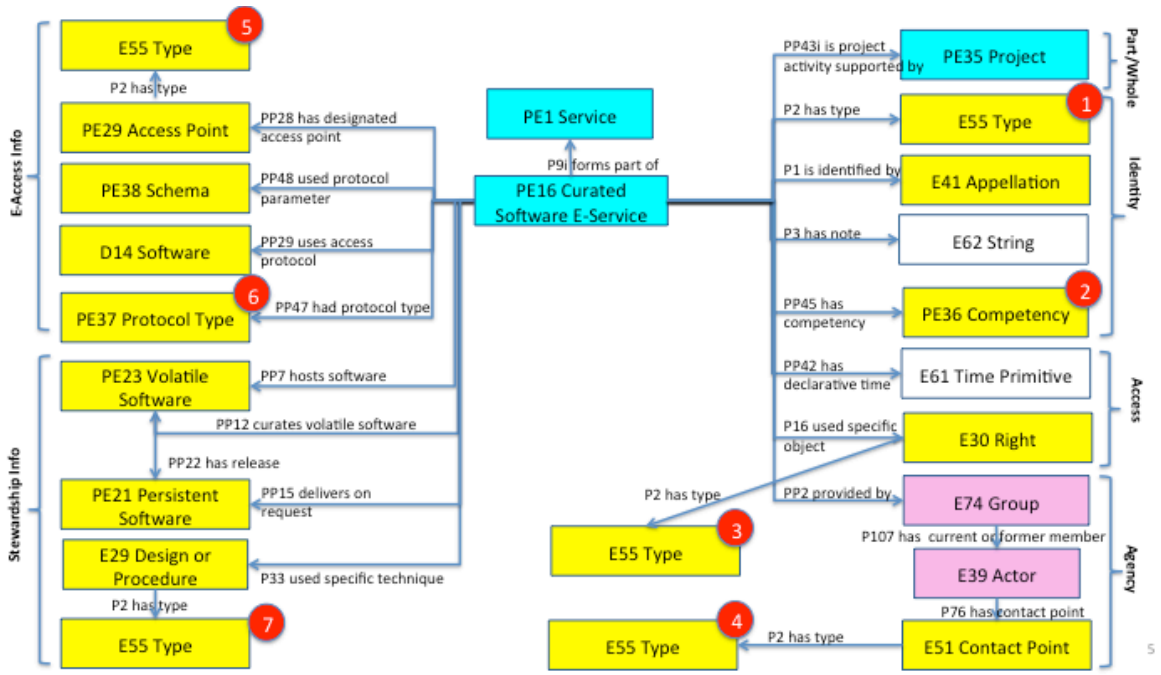

Figure 4: PE16 Curated Software E-Service Minimal Metadata Application Profile Schema

The PE16 Curated Software E-Service minimal metadata application profile makes reference to seven fields which require standardization according to common vocabularies. The following table summarizes the final results of chosen standards relative to these fields.

	Min Metadata Field Name	Path	Role	Recommended Standard
1	Type	PE16→P2→E55	Identity	None
2	Competency	PE16→PP45→PE36	Identity	PARTHENOS Service Competency List
3	Conditions of Use / Rights Type	PE16→P16→E30	Access	PARTHENOS Rights List
4	Communicatoin Address Type	PE16→PP2→E74→ P107→E39→P76→E51 →P2→E55	Agency	CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details
5	Access Point Type	PE16→PP28→PE29→ P2→E55	E-Access	CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details
6	Protocol Type	PE16→PP47→PE37	E-Access	None
7	Curation Plan Type	PE16→p33→E29→P2 →E55	Stewardship	None

Table 9: Recommended standards for PE16 Application Profile

3.2.3. Datasets

With the documentation of datasets, we implement the ontological distinction provided by the PE model between volatile and persistent digital objects. This corresponds roughly to what are loosely called ‘collections’ and ‘files’ or ‘resources’ which consist of encoded propositions about the world. There are different means of identifying these classes of datasets and different questions we would like to pose with regards to them in order to make them operational. A volatile dataset does not have a bit-wise identity from over time, but rather gains an identity by a continuity of activity over a collection of data, a curation

process that in turn adopts a plan which gives sense to the aggregate of data. It can also be known by its backups as offering a snapshot of the datastream at a certain moment. On the other hand, a persistent dataset accords more directly with naive notions of 'files' etc. These are bitwise identical overtime and of particular use in its identification and disambiguation is its participation in larger datasets and the manner in which it was produced.

More analytically a list of questions that we wish to be able to support the user to ask and answer with regards to datasets includes:

- What is it? (Identity)
- What is it part of? [Dataset] (Part/Whole)
- What is it about? (Relevance/Coverage/Content)
- Who has it? (Holding Info)
- How do I access it? (Holding Info/Use)
- How was it made? (Provenance)
- How is it structured? (Provenance/Use)
- Who manages the data? (Curation Info)

This motivates the articulation of the following two basic profiles which in turn motivate a series of required vocabularies.

3.2.3.1. Persistent Dataset

The minimal metadata set profile proposed for Persistent Datasets is as follows:

Label	Mandatory (?)	Field Type	Description
ID	Y	String	The identifier used to indicate the object.
Other IDs	N	String (Multi)	Additional identifiers given to the object.
Type	Y	Controlled Vocabulary [1]	A typology for classifying the kind of dataset contained in this information object.
Title	Y	String	The name by which the object is known or referred to.
Description	N	Long Text	A textual description of the object

Is/Was Part of	Y	Link	The digital object of which this digital object forms part.
Hosted by	Y	Link	The digital hosting service responsible for the hosting of this digital object.
Available at	Y	String	The electronic address at which the object is made available.
Available at Type	N	Controlled Vocabulary [5]	The type of access point at which the object has been made available.
Encoding Type	Y	Controlled Vocabulary [6]	The encoding(s) of the dataset in question.
Schema/ Format	N	Controlled Vocabulary [7]	The schema used to structure the dataset.
Subject	N	Controlled Vocabulary [2]	The role that the dataset can play in research
Spatial Coverage	N	Controlled Vocabulary [4]	The geographic scope for which the dataset has relevance.
Temporal Coverage	N	Controlled Vocabulary [3]	The temporal scope for which the dataset has relevance.
Created by	Y	Link	The link of the dataset to its creator

Table 10: PE22 Application Profile Minimal Metadata Configuration

The semantically encoded expression of the minimal metadataset for PE22 Persistent Dataset is as follows:

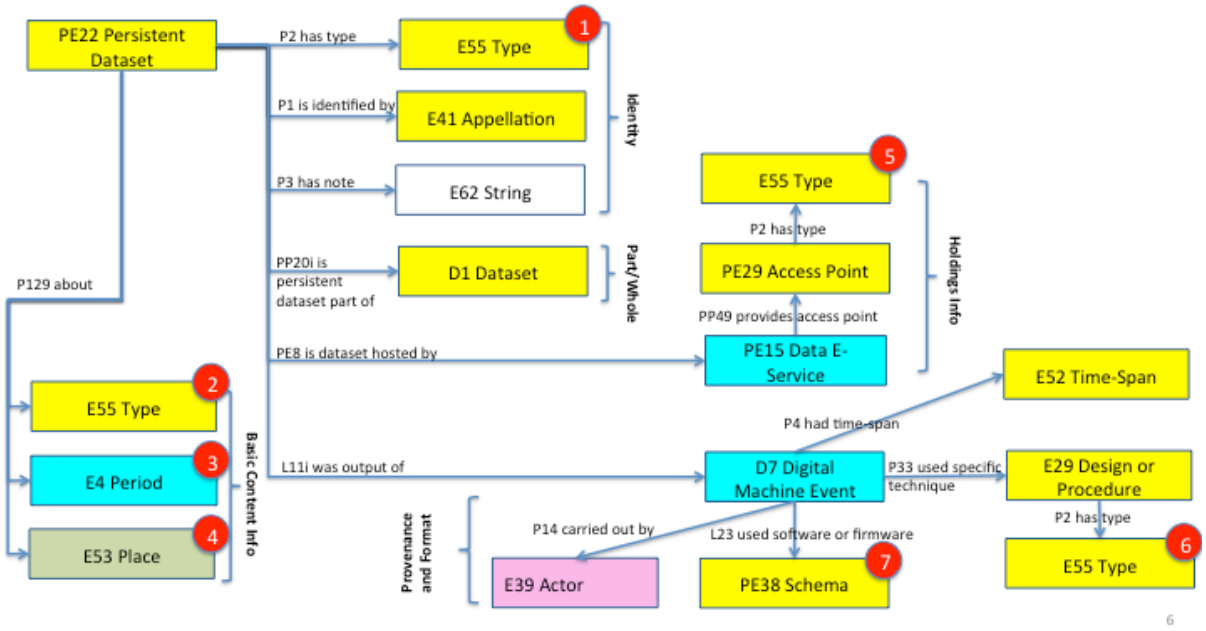

Figure 5: PE22 Persistent Dataset Minimal Metadata Application Profile Schema

The PE22 Persistent Dataset minimal metadata application profile makes reference to seven fields which require standardization according to common vocabularies. The following table summarizes the final results of chosen standards relative to these fields.

	Min Metadata Field Name	Path	Role	Recommended Standard
1	Type	PE22→P2→E55	Identity	CERIF - Output Types
2	Subject	PE22→P129→E55	Coverage	None
3	Temporal Coverage	PE22→P129→E4	Coverage	PeriodO
4	Spatial Coverage	PE22→P129→E53	Coverage	TGN
5	[E-Service] Access Point Type	PE22→PE8i→PE15→ PP49→PE29→P2→ E55	Holdings	CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details
6	Encoding Type	PE22→L11i→D7→P33 →E29→P2→E55	Provenance	File Format Overview and Information
7	Schema/Format	PE22→L11i→D7→L23 →PE38	Provenance	Metadata Standards

Table 11: Recommended standards for PE22 Application Profile

3.2.3.2. Volatile Dataset

The minimal metadata set profile proposed for Volatile Datasets is as follows:

Label	Mandatory (?)	Field Type	Description
ID	Y	String	The identifier used to indicate the object.
Other IDs	N	String (Multi)	Additional identifiers given to the object.
Type	Y	Controlled Vocabulary [1]	A typology for classifying the kind of dataset contained in this information object.
Title	Y	String	The name by which the object is known or referred to.
Description	N	Long Text	A textual description of the object
Is/Was Part of	Y	Link	The digital object of which this digital object forms part.
Hosted by	Y	Link	The digital hosting service responsible for the hosting of this digital object.
Available at	Y	String	The electronic address at which the object is made available.
Available at Type	N	Controlled Vocabulary [5]	The type of access point at which the object has been made available.
Curated by	Y	Link	The digital curating service responsible for the curation of this digital object.
Has Curation Plan	N	Link	The curation plan associated to this curated holding.
Has Curation Plan Type	N	Controlled Vocabulary [8]	The kind of curation plan adopted in the curation of the digital object.
Has Dataset Snapshot	Y	Link	The latest backup of the volatile dataset.
Encoding Type	Y	Controlled Vocabulary [6]	The encoding(s) of the dataset in question.
Schema/Format	N	Controlled Vocabulary [7]	The schema used to structure the dataset.
Subject	N	Controlled Vocabulary	The role that the dataset can play in research

		[2]	
Spatial Coverage	N	Controlled Vocabulary [4]	The geographic scope for which the dataset has relevance.
Temporal Coverage	N	Controlled Vocabulary [3]	The temporal scope for which the dataset has relevance.
Created by	Y	Link	The link of the dataset to its creator

Table 12: PE24 Application Profile Minimal Metadata Configuration

The semantically encoded expression of the minimal metadataset for PE24 Volatile Dataset is as follows:

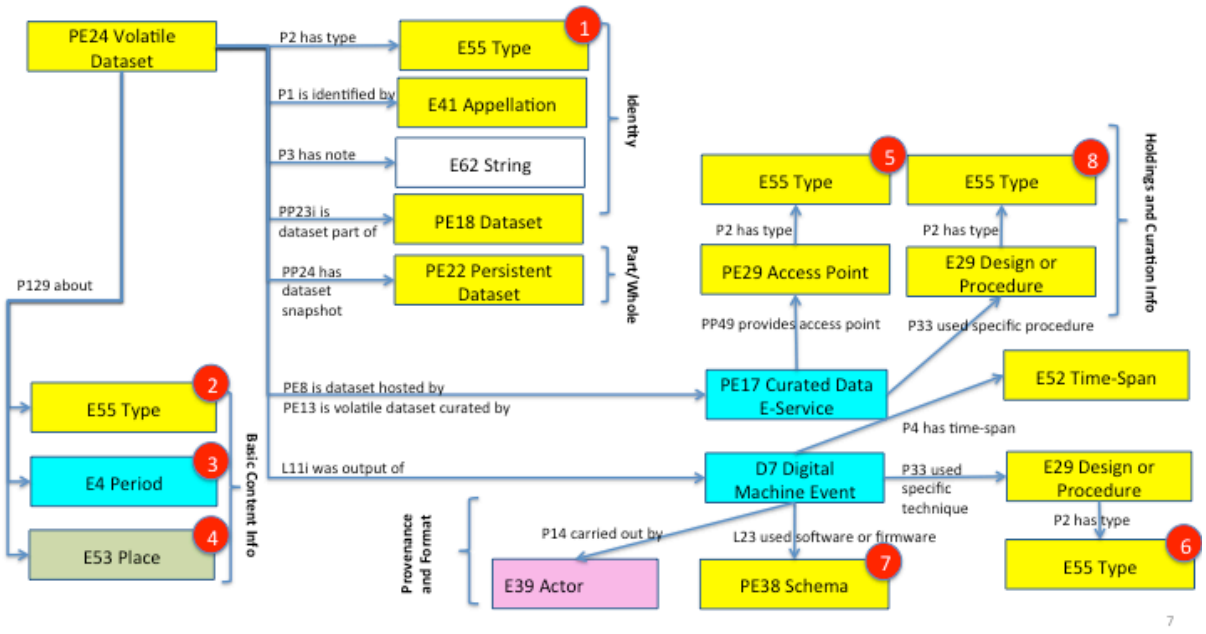

Figure 6: PE24 Volatile Dataset Minimal Metadata Application Profile Schema

The PE24 Volatile Dataset minimal metadata application profile makes reference to eight fields which require standardization according to common vocabularies. The following table summarizes the final results of chosen standards relative to these fields.

	Min Metadata Field Name	Path	Role	Recommended Standard
1	Type	PE24→P2→E55	Identity	CERIF - Output

				Types
2	Subject	PE24→P129→E55	Coverage	None
3	Temporal Coverage	PE24→P129→E4	Coverage	PeriodO
4	Spatial Coverage	PE24→P129→E53	Coverage	TGN
5	[E-Service] Access Point Type	PE24→PE8i→PE15→PP49→PE29→P2→E55	Holdings and Curation	CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details
6	Encoding Type	PE24→L11i→D7→P33→E29→P2→E55	Provenance	File Format Overview and Information
7	Schema/Format	PE24→L11i→D7→L23→PE38	Provenance	Metadata Standards
8	Curation Plan Type	PE24→PE13→PE17→P33→E29→E55	Holdings and Curation	None

Table 13: Recommended standards for PE24 Application Profile

3.2.4. Software

With the documentation of software, we also implement the ontological distinction provided by the PE model between volatile and persistent digital objects. In the context of software this corresponds to the software as a specific product which is developed over time (e.g. Word, Photoshop etc.) and its specific releases (v.1, 2 etc.). This distinction allows us to distinguish and relate a software product as a continuous object of development but also related it to its different expressions over time, which are the usable encodings that execute actual processes and can be distributed/used etc. An instance of volatile software is known through the development plan that holds for it and its releases. An instance of persistent software can be recognized over time by the bit level identity.

More analytically a list of questions that we wish to be able to support the user to ask and answer with regards to datasets includes:

- What is it? (Identity)
- What is it part of? (Identity)
- Who has it? (Holding Info)

- How do I access it? (Holding Info/Use)
    - Where can I download it? (Holding Info/Use)
    - Where can I run it? (Holding Info/Use)
- How was it made? (Provenance)
- How is it structured? (Provenance/Use)
- Who manages the software? (Curation Info)

This motivates the articulation of the following two basic profiles which in turn motivate a series of required vocabularies.

3.2.4.1. Persistent Software

The minimal metadata set profile proposed for Persistent Software is as follows:

Label	Mandatory (?)	Field Type	Description
ID	Y	String	The identifier used to indicate the object.
Other IDs	N	String (Multi)	Additional identifiers given to the object.
Type	Y	Controlled Vocabulary [1]	A typology for classifying the kind of software contained in this information object.
Title	Y	String	The name by which the object is known or referred to.
Description	N	Long Text	A textual description of the object
Executes Processes of Type	Y	Controlled Vocabulary [2]	The types of process that the software can exexcute.
Is/Was Part of	Y	Link	The digital object of which this digital object forms part.
Is Release of	Y	Link	The volatile software object of which this object is a release.
Run by	Y	Link	The digital e-service that offers to run a software service.
Available at	Y	String	The electronic address at which the software can be run.
Available at	N	Controlled	The type of access point at which the

Type		Vocabulary [3]	software has been made available.
Delivered by	Y	Link	The digital e-service that offers a download point for the software.
Available at	Y	String	The electronic address at which the software can be downloaded.
Available at Type	N	Controlled Vocabulary [3]	The type of access point at which the software has been made available.
Created by	Y	Link	The link of the dataset to its creator
Programming Language	N	Controlled Vocabulary [4]	The programming language used in creating the software.

Table 14: PE21 Application Profile Minimal Metadata Configuration

The semantically encoded expression of the minimal metadataset for PE21 Persistent Software is as follows:

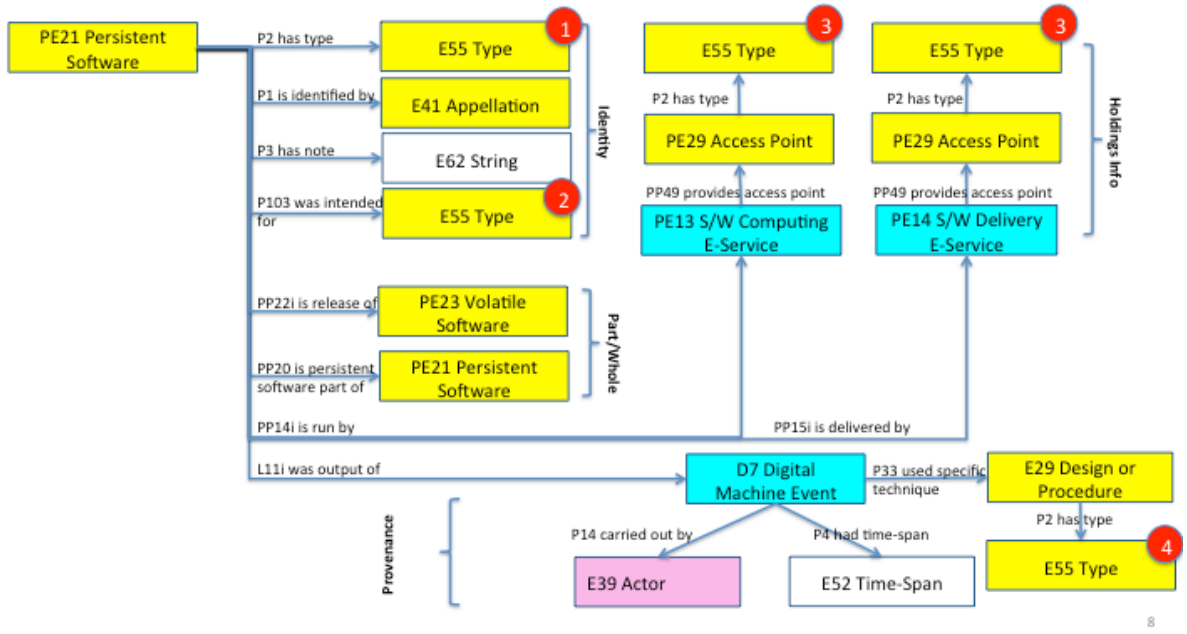

Figure 7: PE21 Persistent Software Minimal Metadata Application Profile Schema

The PE21 Persistent Software minimal metadata application profile makes reference to four fields which require standardization according to common vocabularies. The following table summarizes the final results of chosen standards relative to these fields.

	Min Metadata Field Name	Path	Role	Recommended Standard
1	Type	PE21→P2→E55	Identity	CERIF - Output Types

2	Process type	PE21→P103→E55	Identity	None
3	[E-Service] Access Point Type	PE21→PE14/5i→PE13/4→PP49→PE29→P2→E55	Holdings	CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details
4	Programming Language	PE21→L11i→D7→P33→E29→P2→E55	Provenance	Wikipedia list of programming languages

Table 15: Recommended standards for PE21 Application Profile

3.2.4.2. Volatile Software

The minimal metadata set profile proposed for Volatile Software is as follows:

Label	Mandatory (?)	Field Type	Description
ID	Y	String	The identifier used to indicate the object.
Other IDs	N	String (Multi)	Additional identifiers given to the object.
Type	Y	Controlled Vocabulary [1]	A typology for classifying the kind of software contained in this information object.
Title	Y	String	The name by which the object is known or referred to.
Description	N	Long Text	A textual description of the object
Executes Processes of Type	Y	Controlled Vocabulary [2]	The types of process that the software can execute.
Is/Was Part of	Y	Link	The digital object of which this digital object forms part.
Has Release	Y	Link	The volatile software object of which this object is a release.
Run by	Y	Link	The digital e-service that offers to run a software service.
Available at	Y	String	The electronic address at which the software can be run.
Available at Type	N	Controlled Vocabulary [3]	The type of access point at which the software has been made available.

Delivered by	Y	Link	The digital e-service that offers a download point for the software.
Available at	Y	String	The electronic address at which the software can be downloaded.
Available at Type	N	Controlled Vocabulary [3]	The type of access point at which the software has been made available.
Curated by	Y	Link	The service that cureates the digital object in question.
Created by	Y	Link	The link of the dataset to its creator
Programming Language	N	Controlled Vocabulary [4]	The programming language used in creating the software.

Table 16: PE23 Application Profile Minimal Metadata Configuration

The semantically encoded expression of the minimal metadataset for PE23 Volatile Software is as follows:

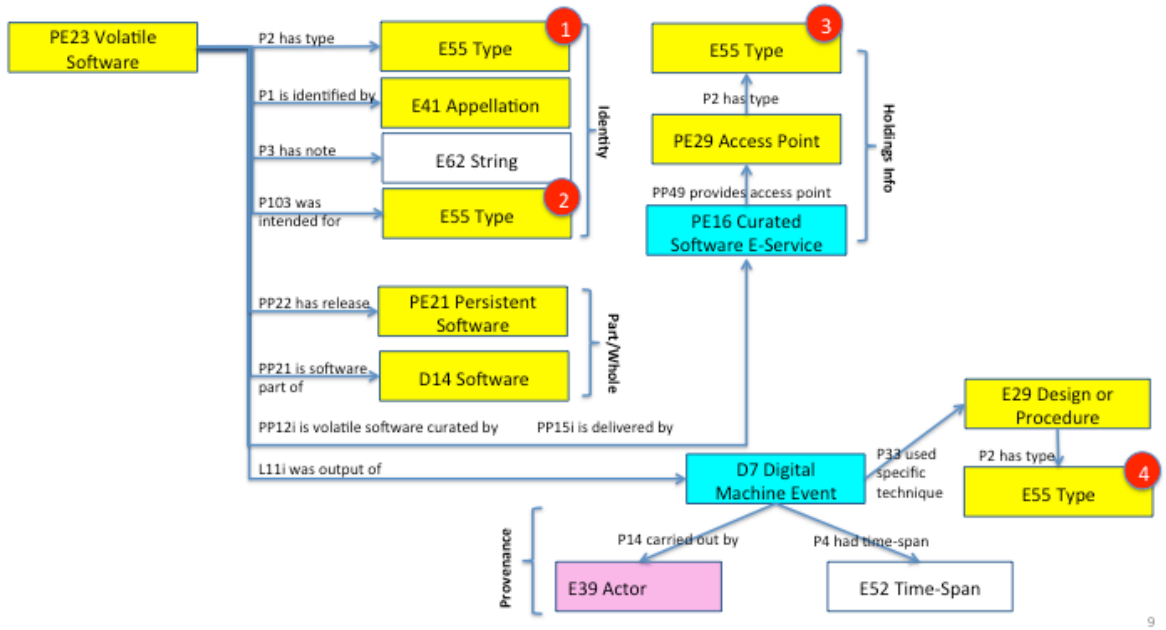

Figure 8: PE23 Volatile Software Minimal Metadata Application Profile Schema

The PE23 Volatile Software minimal metadata application profile makes reference to four fields which require standardization according to common vocabularies. The following table summarizes the final results of chosen standards relative to these fields.

	Min Metadata Field Name	Path	Role	Recommended Standard

1	Type	PE23→P2→E55	Identity	CERIF - Output Types
2	Process type	PE23→P103→E55	Identity	None
3	[E-Service] Access Point Type	PE23→PE14/5i→PE13/4→PP49→PE29→P2→E55	Holdings	CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details
4	Programming Language	PE23→L11i→D7→P33→E29→P2→E55	Provenance	Wikipedia list of programming languages

Table 17: Recommended standards for PE23 Application Profile

3.2.5. Actors

Keeping track of actors is an essential part of the PARTHENOS Entities model. Actors, be they teams or individuals, are the knowledge agents behind services and projects which have the final understanding of datasets and software that were generated or affected by them. They are also those to be contacted to know more about and make requests regarding projects and services generally.

WIth the actor classes we wish to support answering the following types of questions to the information model:

- Who is it? (Identity)
- How can they be contacted? (Communication)
- What groups have they been part of? (part/whole)
- What do they provide/maintain? (Activities)

Within the context of an application profile, one can reduce the actors classes to the documentation of teams (with RI Consortium a special subclass) and persons (individuals).

3.2.5.1. Team

The minimal metadata set profile proposed for Team is as follows:

Label	Mandatory	Field Type	Description

	(?)		
ID	Y	String	The identifier used to indicate the actor.
Type	Y	Controlled Vocabulary [1]	A typology for classifying the kind of actor.
Appelation	Y	String	The name by which the actor is known or referred to.
Description	N	Long Text	A textual description of the actor
Address	Y	String	An address at which the team can be contacted or legal address..
Address Type	Y	Controlled Vocabulary [2]	A type for the address given.
General Email	N	String	An email address for the actor.
Contact Person	N	Link	A designated contact person for the actor in question.
Contact Person Address	Y	String	Address of the designated contact person.
Contact Person Address Type	Y	Controlled Vocabulary [3]	A type for the address given.
Maintainer of	N	Link	The project which is maintained by this actor.
Provides	N	Link	Services offered by the actor.

Table 18: PE34 Application Profile Minimal Metadata Configuration

The semantically encoded expression of the minimal metadataset for PE34 Team is as follows:

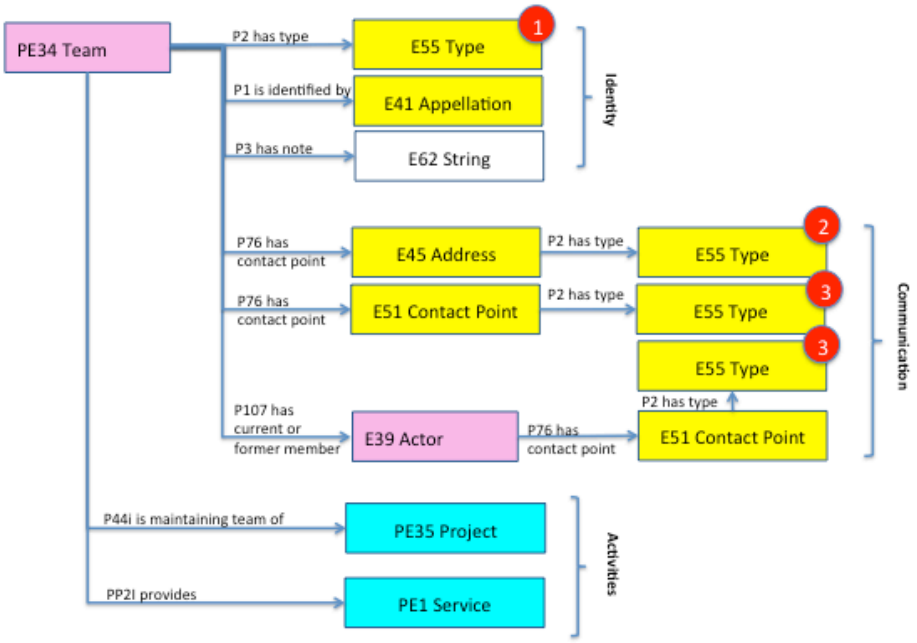

Figure 9: PE34 Team Minimal Metadata Application Profile Schema

The PE34 Team minimal metadata application profile makes reference to three fields which require standardization according to common vocabularies. The following table summarizes the final results of chosen standards relative to these fields.

	Min Metadata Field Name	Path	Role	Recommended Standard
1	Type	PE34→P2→E55	Identity	None
2	Address Type	PE34→P76→E45→P2→E55	Identity	CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details
3	Contact Point Type	PE34→P76→E51→P2→E55	Access	CERIF Electronic Address Type & Person Contact Details & Organisation Contact Details

Table 19: Recommended standards for PE34 Application Profile

3.2.5.2. Person

The minimal metadata set profile proposed for Person is as follows:

Label	Mandatory (?)	Field Type	Description
ID	Y	String	The identifier used to indicate the actor.
Type	Y	Controlled Vocabulary [1]	A typology for classifying the kind of actor.
Appelation	Y	String	The name by which the actor is known or referred to.
Description	N	Long Text	A textual description of the actor
Address	Y	String	An address at which the team can be contacted or legal address..
Address Type	Y	Controlled Vocabulary [2]	A type for the address given.
Email	N	String	An email address for the actor.
Part of Team	N	Link	Link to team of which actor is a part.
Provides	N	Link	Services offered by the actor.

Table 20: E21 Application Profile Minimal Metadata Configuration

The semantically encoded expression of the minimal metadataset for E21 Person is as follows:

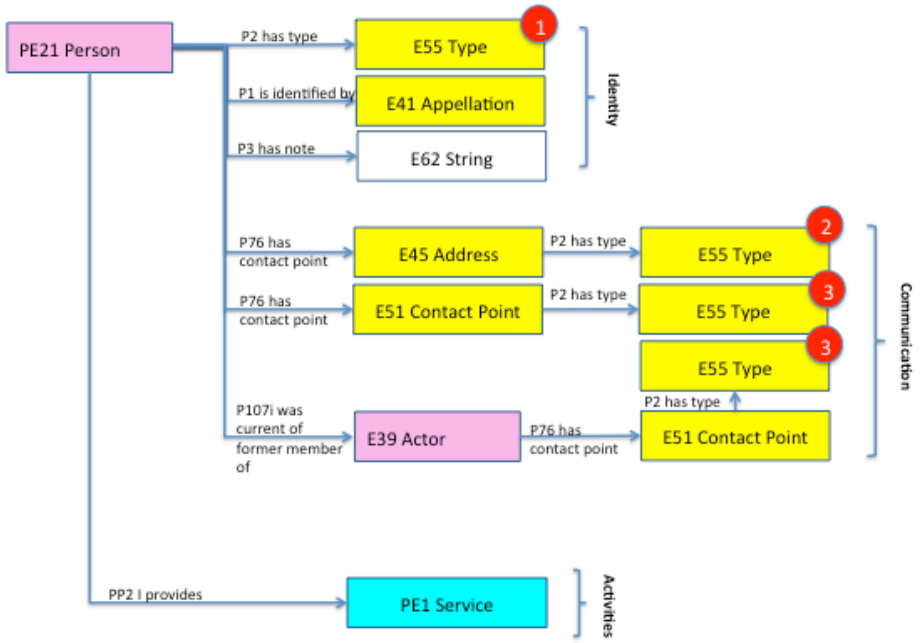

Figure 10: E21 Person Minimal Metadata Application Profile Schema

The E21 Person minimal metadata application profile makes reference to three fields which require standardization according to common vocabularies. The following table summarizes the final results of chosen standards relative to these fields.

	Min Metadata Field Name	Path	Role	Recommended Standard
1	Type	PE21→P2→E55	Identity	None
2	Address Type	PE21→P76→E45→P2 →E55	Identity	CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details
3	Contact Point Type	PE21→P76→E51→P2 →E55	Access	CERIF Electronic Address Type & Person Contact Details & Organisation Contact Details

Table 21: Recommended standards for PE21 Application Profile

4. Vocabularies Research

In line with the principles of both the conceptual modelling taken up to form the PARTHENOS Entities model and the methodology proposed by the BBT, research into required vocabularies was driven by a ground up process. In the process of populating the PARTHENOS Joint Research Registry through the mapping of RI registries to the PARTHENOS Entities Model in the X3ML Suite and using the D-Net Aggregation Infrastructure,[9] the required vocabularies to properly standardized data at the registry level was derived inductively. The above application profiles represent instantiations of the minimal metadata standard proposed in PARTHENOS. Actual data arriving from RIs varied in richness of detail, have more or less information about the different basic entities. Therefore, the complete list of vocabularies collected goes beyond the types identified relative to the minimal metadata. In what follows we will look at the need for standards identified from RI sources and comment why different standards were chosen, dropped or created for PARTHENOS' needs.

As there is not a singular place or institution to refer to when researching a standardized vocabulary for a particular field or topic, research broadly extended in all directions. Most helpful were several vocabulary collections hosted online, like the Basel Registry of Thesauri, Ontologies & Classifications (BARTOC)[10], the Open Metadata Registry[11] the Linked Open Vocabularies (LOV)[12], and the CERIF data model[13] which served to provide a with a wide range of different candidates, from very compact, focused vocabularies, to large term collections with thousands of entries. However, identifying suitable candidates often proved a difficult task: for many subjects, a well-defined standardization does simply not exist. The more potential for heterogeneity a subject has, the slimmer the chances for a standard to fit the desired values or even be conceivable. For other topics, one or a few vocabularies could be identified, but were too narrow in scope for the more heterogeneous nature of the data provided by the RIs. Other areas, often those in focus of multiple fields of research, are better covered and offered multiple extensive options to chose from.

[9] http://www.d-net.research-infrastructures.eu/node/22
[10] https://bartoc.org/
[11] http://metadataregistry.org/
[12] http://lov.okfn.org/dataset/lov/
[13] Used in a number of European projects, this data model includes also lists of controlled vocabularies that are empriicially derived and provide a rich resource for meta-metadata: http://www.eurocris.org/cerif/feature-tour/cerif-15

We will look at the standards according to their use within the ontology.

4.1. Activities Related Vocabularies

Data from RIs contained richer information with regards to certain types of general activities outside of the description of services. Some RIs documented different types of publishing activities while others documented, at least in principle, digitization activities. Of relevance to document for many RIs was also the role that actors played in a given activity. The model predicted that part of the documentation would cover the manner of preserving data. This was not borne out by the data retrieved. Research did not reveal strong relevant candidates for standard vocabularies for these identified fields. Therefore, in general we chose to create PARTHENOS specific vocabularies for the fields that we decided should be covered.

Activities				
Vocab Needed	**Function**	**Standards Considered**	**Decision**	**Rationale**
Activity Type	Classify activities generically	CERIF Activity Types PAV	PARTHENOS Publishing Activities List	No applicable standards with satisfying coverage
Digitization Process Types	Classify types of digitizing activities	Yale University Digitization Standards and Guidelines	Dropped	Not present in the data / recorded by any RI
Digital Machine Event Type	Classify types of intentionally activated digital events	PAV	PARTHENOS Publishing Activities List	Strong thematic overlap with Activity Type
Actor Roles in Activities	Classify actor roles of creating an intellectual product	CASRAI Contributor Roles Taxonomy Publishing Roles Ontology Scholarly	PARTHENOS Publishing Roles List	Broad concept combined with a more constricted selection of used values in the data makes a custom vocabulary the

		Contributions and Roles Ontology CERIF Person Organization Roles		most feasible
Preservation Activity Type	Classify types of preservation activities	PAV	Dropped	Not present in the data / recorded by any RI
DateTime Norms	Standardization of date & time values	ISO 8601 Standard	ISO 8601 Standard	Well-known standard with good representation of values

Table 22: Summary of standard vocabularies considered for Activities

4.2. Services Related Vocabularies

For services, the minimal metadata set proposed a number of basic descriptors for understanding what a service is and when it can be used. Research did not reveal well known standards for either of these descriptors and therefore necessitated the elaboration of a self generated list.

Services				
Vocab Needed	**Function**	**Standards Considered**	**Decision**	**Rationale**
Availability	Classify types of resource availabilities	Document Availability Information Ontology	PARTHENOS Availability List	Few values in the data warrent custom liste more than speculative third part option
DateTime Norms	Standardization of date & time values	ISO 8601 Standard	ISO 8601 Standard	Well-known standard with good representation of values
Service Competency Types	Classify types of competencies of services (e.g. LP)	None	PARTHENOS Service Competency List	No applicable standards with satisfying coverage

Table 23: Summary of standard vocabularies considered for Services

4.2.1. Curating Service Related Vocabularies

The PARTHENOS Minimal Metadata places an important emphasis on the documentation of the curation plan for the identity of a curated item. Therefore it recommends the documentation of a curation plan. This could be an official document or just a reference to the kind of plan followed. In practice, it would seem no one documents this, so no vocabulary could be chosen based on the data. In the same vein, archives seem to normally record accrual method type and accrual policy type. These could be considered also as curation plans. While some data were mapped to such fields in practice they were empty and therefore no vocabularies could be selected. However, some of the considered candidates could become relevant at a later date, with potentially more data getting integrated covering some of those typifications.

Services - Curating				
Vocab Needed	**Function**	**Standards Considered**	**Decision**	**Rationale**
Curation Types	Classify types of resource curations	DPCVocab	Dropped	Not present in the data / recorded by any RI
Curation Plan Types	Classify types of curation plans	None	Dropped	Not present in the data / recorded by any RI
Accrual Method Type	Classify types of accrual methods	Dublin Core Collection Description Frequency Vocabulary Dublin Core Collection Description Accrual Method Namespace CERIF Person Output Contributions &	Dropped	Not present in the data / recorded by any RI

		Person Project Engagements		
Accrual Policy Type	Classify types of accrual policies	Dublin Core Collection Description Accrual Policy Namespace	Dropped	Not present in the data / recorded by any RI

Table 24: Summary of standard vocabularies considered for Curating Services

4.2.2. E-Service Related Vocabularies

In order to gather important information to facilitate automatic integration of services that offer e-platforms, the PARTHENOS minimal metadata model suggests the gathering of a number of basic fields describing the means by which to establish electronic communication with a certain e-service. Again, fields necessary for doing this were often not actually documented in the source. Where they were, research was able to find some standard vocabularies.

Services - E-Service				
Vocab Needed	**Function**	**Standards Considered**	**Decision**	**Rationale**
Authorization Policies	Classify types of authorization policies	None	PARTHENOS Rights List	Not present in the data / recorded by any RI
Contact Point Types	Classify types of points of contact	CERIF Electronic Address Type & Person Contact Details & Organisation Contact Details International Contact Ontology NEPOMUK	CERIF Electronic Address Type & Person Contact Details & Organisation Contact Details	Best fit for present data values

		Contact Ontology Contact: Utility concepts for everyday life		
Access Point Type	Classify types of access points	See Contact Point Types	CERIF Electronic Address Type & Person Contact Details & Organisation Contact Details	Very strong overlap of classifications

Table 25: Summary of standard vocabularies considered for E-Services

4.3. Dataset Related Vocabularies

Datasets mapped to the PARTHENOS Entities model not surprisingly turned out to have the greatest amount of additional data going beyond the minimal metadata requirements and requiring a reflection on appropriate standards which would allow their global query.

It was quite typical for the dataset to refer to the form of its content, for example book or list or journal etc. Therefore, a typology for this was sought and found.

Datasets				
Vocab Needed	**Function**	**Standards Considered**	**Decision**	**Rationale**
Dataset Types	Classify types of datasets	CERIF - Output Types	CERIF - Output Types	Only relevant candidate and good fit for present data values

Table 26: Summary of standard vocabularies considered for Datasets

4.3.1. Dataset: Aboutness Related Vocabularies

Many datasets carried relatively accurate high level information concerning the subject or referent of their content. This usually broke down into place, period and subject referent, causing a search for appropriate vocabularies. The subject referent is the most complicated and will be left to the second part of the project for scholarly research together with the RIs.

Datasets - Aboutness				
Vocab Needed	**Function**	**Standards Considered**	**Decision**	**Rationale**
Places	Classify types of places/locations	Getty Thesaurus of Geographic Names (TGN) GeoNames geographical database Free World Cities Database	Getty Thesaurus of Geographic Names (TGN)	Most extensive list of terms, with the best chance of covering types present in the data
Spatial Coordinates	Standardize spatial coordinate values		**TBD**	**TBD**
Subject Types	Classify types of subjects	CERIF Person Output Contributions & Person Project Engagements UNESCO Thesaurus Library of Congress Subject Headings (LCSH) Zine Thesaurus of Subject Terms	None	As this field is highly dependent on the actual content of the data sets, further input from the RIs is required, especially as they might already have vocabularies of their own

Periods	Classify historic time periods	PeriodO Historic England Periods Authority File iDAI.chronontology	PeriodO	Best fit for present data values and very exhaustive

Table 27: Summary of standard vocabularies considered for Dataset Aboutness

4.3.2. Dataset: Properties Related Vocabularies

The dataset properties found in the actual sources were richer in description of descriptors not specified by the minimal metadata. It was, for example, extremely rare to find documentation of encoding type or schema type, something which will make it fundamentally difficult to work with this data. The identification of the language in which the information is presented was relatively well documented and things like dimensions (even file size) were documented. Where possible appropriate general vocabularies were identified and recommended.

Datasets - Properties				
Vocab Needed	**Function**	**Standards Considered**	**Decision**	**Rationale**
Languages	Standardized language identifiers	Languages Name Authority List (NAL)	Languages Name Authority List (NAL)	Only relevant candidate and very exhaustive list
Encoding Types	Classify types of file encodings	QaamGo Media File format overview and information Iana Media Types	**TBD**	**TBD**
Schema Types	Classify types of schemata	Metadata 2nd Edition (2016) - Metadata Standards	Metadata 2nd Edition (2016) - Metadata Standards	Only relevant candidate and very exhaustive list
Dimension	Classify types of	Units of	Dropped	Not present in

Types	dimensions	Measurement Ontology		the data / recorded by any RI
Material Types	Classify types of materials	FISH Building Materials Thesaurus Art & Architecture Thesaurus Materials Facet	Dropped	Not present in the data / recorded by any RI Recommendation for AAT

Table 28: Summary of standard vocabularies considered for Dataset Properties

4.3.3. Dataset: Rights Related Vocabularies

The PARTHENOS minimal metadata recommendation sought to link rights to services. Actual practice as indicated from the incoming RI data suggests that it is much more typically and more assiduously documented on the dataset level. The issue of rights is quite complicated and there are many different types to take account of. We took advantage of the many views on rights across RIs to make a high level tree of types of rights, information we could not otherwise find elsewhere in a suitable format. While many different types of rights were documented, we felt they could be functionally collated in a single rights type hierarchy of use at a general level.

Datasets - Rights				
Vocab Needed	**Function**	**Standards Considered**	**Decision**	**Rationale**
Rights Types	Classify types of rights	None	PARTHENOS Rights List	Too broad of a field, with too few and heterogeneous values in the data
Condition of Use	Classify conditions of use	None	PARTHENOS Rights List	See Rights Types
Access Policies Types	Classify types of access policies	None	PARTHENOS Rights List	See Rights Types

Access Rights	Classify types of acces rights	None	PARTHENOS Rights List	See Rights Types
Use Restriction	Classify types of use restrictions	None	PARTHENOS Rights List	See Rights Types

Table 29: Summary of standard vocabularies considered for Dataset Rights

4.4. Software Related Vocabularies

The PARTHENOS minimal metadata model suggested documenting the programming language used to create a software item and the kinds of processes that it could execute. This latter would enable linking software to potential datasets. In fact, the incoming data revealed these are rarely recorded in our case. For programming languages, well known lists can be found anyhow. With regards to process types, the lack of empirical data to work with made a decision on adopting or creating some standard impossible.

Software				
Vocab Needed	**Function**	**Standards Considered**	**Decision**	**Rationale**
Programming Language	Classify programming languages	Wikipedia list of programming languages	Wikipedia list of programming languages	Only valid candidate and very exhaustive list
Process Types	Classify types of software processes	None	Dropped	Not present in the data / recorded by any RI

Table 30: Summary of standard vocabularies considered for Software

4.5. Actors Related Vocabularies

For actors, the minimal metadata model made few requirements. The idea of legal statuses suggested in the model turned out to be highly theoretical against the actual data. It was not documented in source and therefore no vocabulary could be selected. Most important were descriptors connecting actors to places and addresses. For these, good solutions could be discovered.

Actors				
Vocab Needed	**Function**	**Standards Considered**	**Decision**	**Rationale**
Actor Types	Classify types of actors	None	Dropped	Not present in the data / recorded by any RI
Contact Point Types	Classify types of points of contact	CERIF Electronic Address Type & Person Contact Details & Organisation Contact Details International Contact Ontology NEPOMUK Contact Ontology Contact: Utility concepts for everyday life	CERIF Electronic Address Type & Person Contact Details & Organisation Contact Details	Best fit for present data values
Places	Classify types of places/locations	Getty Thesaurus of Geographic Names (TGN) GeoNames geographical database Free World Cities Database	Getty Thesaurus of Geographic Names (TGN)	Most extensive list of terms, with the best chance of covering types present in the data
Spatial Coordinates	Standardize spatial coordinate values		**TBD**	**TBD**
egal Statuses	Classify types of legal statuses	CERIF cfOrgUnit	Dropped	Not present in the data / recorded by any RI

Table 31: Summary of standard vocabularies considered for Actors

4.6. Vocabularies as Curated Datasets

The investment of time and effort to find effective and potentially sustainable thesauri for use as controlled vocabularies in the PARTHENOS Joint Resource Registry is a solid empirical validation of the utility and yet inaccessibility/invisibility of such resources to a wider public. In fact the creation and maintenance of a thesaurus and particularly its maintenance is a long term investment in a curatorial project that has significant knock on effect and impact beyond the immediate collation of data. The importance of these resources and the difficulty of finding them, led to the decision that they should not only be used in PARTHENOS but documented as resources in their own right and offered within the Joint Research Registry as resources for the overall users of the PARTHENOS services.

To this end, the vocabularies identified for use in the Joint Research Registry have been documented as instances of PE24 Volatile Dataset following the minimal metadata model and will be merged into the Joint Research Registry. The official list of vocabularies described using the minimal metadata for volatiles datasets is also appended in Appendix II at the end of this document.

5. Matching Identified Vocabularies to BBT

In section 1.3 above, we introduced the idea of the BBT and how it aims to serve a broad interdisciplinary community of researchers by allowing an open ended expansion of federated thesauri through an open, revisable and methodologically clear hierarchy of vocabularies. The first test of this methodology in the PARTHENOS project comes with the integration of the vocabularies identified for use in the PARTHENOS Entities to the established facets and hierarchies of the BBT. The initial results of this activity can be seen in the re-expressed BBT now with the PARTHENOS Entities vocabularies integrated within the general framework.

activities
- disciplines

- human interactions
- intentional destructions
- functions
- service competency [BBT NEW]
    - ***PARTHENOS Service Competency List***
- data management activities [BBT NEW]
    -***PARTHENOS Publishing Activities List***

natural processes

- natural disasters
- geneses

materials

material things

- mobile objects
- built environment
- physical features
- structural parts of material things

types of epochs

conceptual objects

- symbolic objects
    - identifiers [BBT NEW]
        - contact points[BBT NEW]
            - ***CERIF - Electronic Address Type, Person Contact Details and Organisation Contact Details***
    - encoding [BBT NEW]
        - ***File Format and Overview Information***

- propositional objects
    - dataset [BBT NEW]
        - ***CERIF Output Types***
- norms [BBT NEW]
    - rights [BBT NEW]
        - ***PARTHENOS Rights List***

    - authorization policy
        - ***PARTHENOS Authorization Policies List***
- methods

- language [BBT NEW]
    - natural language [BBT NEW]
        - ***Languages Name Authority List (NAL)***
    - formal language [BBT NEW]
        - programming language [BBT NEW]

- ***Wikipedia Programming Language List***

- concepts

groups and collectivities

roles

- offices
- roles of interpersonal relations

-Publishing Roles [BBT NEW]

-***PARTHENOS Publishing Roles***

geopolitical units

In total we integrated ten vocabularies discovered in the effort to find robust and sufficiently wide but accurate control terms. The following were the results of the integration divided by top level facet.

5.1. Activities Vocabularies

Here we introduced two vocabularies: one for service competency (PARTHENOS List) and one for data management activities (PARTHENOS Publishing Activities List). As this facet is specifically designed for types of activity there was no difficulty in finding a home for these vocabularies, although they did necessitate the introduction of the new general hierarchical terms for these vocabularies.

5.2. Conceptual Objects Vocabularies

The symbolic objects facet is designed to capture types of immaterial but identifiable mental products. Into this category the integration of the vocabularies from PARTHENOS cause the need for a number of new hierarchies. To integrate our thesauri the following new hierarchies had to be declared.

Under symbolic object were declared: identifiers -> contact points, encoding and schema. The new hierarchy ‘identifiers’ was declared for all sorts of symbols that aim to univocally name an item through a certain elaborated identification system. Contact points are a sub-

hierarchy of this as identifiers used for addresses of all types. Encoding, not requiring any content, can be considered as kinds of symbolic object.

Under propositional object we declared dataset as a sub-hierarchy. Propositional objects are defined as some sort of informational content about the world which is in line with the PARTHENOS Entities understand of dataset.

The integration of rights as a hierarchy under conceptual objects required the declaration also of a hierarchy ‘norms’. The new norms hierarchy will cover all sorts of systems of regulation. Rights are a natural division underneath norms and fit well there. Likewise authorization policies fit well as a sub hierarchy of this new division.

Likewise, the effort to integrate both a vocabulary for natural languages and programming languages motivated the declaration of an entirely new branch within the conceptual objects facet to deal with systems of communication (as opposed to their products in symbols, propositions and information objects). Therefore a new hierarchy for language with sub-hierarchies for natural and formal languages was created.

5.3. Roles Vocabularies

Within the roles facet, a place was found for the publishing roles that are documented by PARTHENOS RIs with regards to the management of datasets.

5.4. Non-Vocabulary Style Standards

Worthy of note are three standardized sources that we did not integrate to the BBT, namely the PeriodO system for standardizing periods and the TGN system for standardizing geographic referents and a standard for describing schema types. None of these forms a vocabulary in the sense of the typologies that BBT handles. They are controlled knowledge systems about particulars and not types. Therefore, they are intentionally not mapped into the BBT system which is expressely designed for organization information and the categorical level. Rather, they will be addressed through continued methodological work on spatiotemporal gazetteers with regards to space-time

concepts and through the collection of an authoritative list of schemas in use by which to control the list of schema particulars in the knowledge base.

6. Conclusion

The preliminary integration work of the identified vocabularies into the BBT provided valuable experience for beginning to plan the broader integration of reference resources envisioned by PARTHENOS to be accomplished through this system. As foreseen by the PARTHENOS model, the integration of a broader set of resources initiates a process of revision of the model itself. The need for the declaration of new major subsections of the hierarchy and the elaboration of robust scope notes for these concepts is work in progress. Regardless, it gives a good sense of the need for a constant feedback loop between the vocabulary integrators amongst themselves. This provides good practical experience for how to actually integrated the BBT within the PARTHENOS services.

7. Analysis and Next Work

This deliverable looked at the integration of reference resources for the PARTHENOS community. The immediate need to this end was recognized as the need for the identification of standardized vocabularies for entities used within the PARTHENOS Entities Model. The needs were identified, different standards researched and final decisions reached as to which standards to implement in the project. The strategy for integrating such resources is to follow the BBT model proposed by DARIAH. Therefore, we engaged in a preliminary testing of the possibility of integrating these vocabularies to the BBT, an exercise ending with success. The experience and outcome of this interim report will serve as the basis for the work in the next phase of the project in which a tool for implementing this integration process will be integrated to the PARTHENOS services and RIs polled for important reference resources to be integrated under the common platform.

Appendix I: Vocabulary Candidates

Vocabulary Candidates		
Name	**Creator / Source**	**Link**
CERIF	VRE4EIC	http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/CERIF1.5_Semantics.xhtml
PAV	Paolo Ciccarese, Stian Soiland-Reyes	http://pav-ontology.github.io/pav/pav.rdf
Yale University Digitization Standards and Guidelines	Yale University	http://web.library.yale.edu/digitizationguidelines/guidelines
CASRAI Contributor Roles Taxonomy	CASRAI	http://dictionary.casrai.org/Contributor_Roles
Publishing Roles Ontology	David Shotton, Silvio Peroni	http://www.sparontologies.net/ontologies/pro/source.html
Scholarly Contributions and Roles Ontology	David Shotton, Silvio Peroni	http://www.sparontologies.net/ontologies/scoro/source.html
Document Availability Information Ontology	Jakob Voß	https://github.com/gbv/daia/
DPCVocab	Tiffany C. Chao, Melissa H. Cragin, Carole L. Palmer	https://www.ideals.illinois.edu/handle/2142/44032
Dublin Core Collection Description Frequency Vocabulary	Dublin Core Metadata Initiative	http://dublincore.org/groups/collections/frequency/2013-06-26/freq.rdf
Dublin Core Collection Description Accrual Method Namespace	Dublin Core Metadata Initiative	http://dublincore.org/groups/collections/accrual-method/2013-06-26/accmeth.rdf

Name	Creator / Source	Link
Dublin Core Collection Description Accrual Policy Namespace	Dublin Core Metadata Initiative	http://dublincore.org/groups/collections/accrual-policy/2013-06-26/accpol.rdf
International Contact Ontology	Mark S. Fox	http://ontology.eil.utoronto.ca/icontact.html
NEPOMUK Contact Ontology	Antoni Mylka, Leo Sauermann, Michael Sintek, Ludger van Elst	https://developer.gnome.org/ontology/stable/nco-ontology.html
Contact: Utility concepts for everyday life	Berners-Lee	https://www.w3.org/2000/10/swap/pim/contact
Getty Thesaurus of Geographic Names (TGN)	Getty Research Institute	http://www.getty.edu/research/tools/vocabularies/tgn/
GeoNames geographical database	Unknown	http://www.geonames.org/
Free World Cities Database	MaxMind	https://www.maxmind.com/en/free-world-cities-database
UNESCO Thesaurus	UNESCO	http://vocabularies.unesco.org/browser/thesaurus/en/index
Library of Congress Subject Headings (LCSH)	Library of Congress	https://www.loc.gov/aba/cataloging/subject/
Zine Thesaurus of Subject Terms	Anchor Archive Zine Library	http://robertsstreet.org/n/thesaurus/out.htm
PeriodO	Adam Rabinowitz, Ryan Shawn	http://perio.do/
Historic England Periods Authority File	SENESCHAL project	http://heritagedata.org/live/schemes/eh_period.html
iDAI.chronontology	iDAI	http://chronontology.dainst.org/
Languages Name Authority List (NAL)	EU	http://data.europa.eu/euodp/en/data/dataset/language
QaamGo Media File format overview and information	QaamGo Media	https://www.online-convert.com/file-type

Name	Creator / Source	Link
Iana Media Types	IANA	https://www.iana.org/assignments/media-types/media-types.xhtml
Metadata 2nd Edition (2016) - Metadata Standards	Marcia L.ei Zeng, Jian Qin	http://www.metadataetc.org/book-website/readings/appendixaschemas.htm
Units of Measurement Ontology	National Center for Biomedical Ontology	https://bioportal.bioontology.org/ontologies/UO
FISH Building Materials Thesaurus	SENESCHAL project	http://heritagedata.org/live/schemes/eh_tbm.html
Art & Architecture Thesaurus Materials Facet	Getty Research Institute	http://www.getty.edu/vow/AATHierarchy?find=&logic=AND&note=&english=N&subjectid=300000000
Wikipedia list of programming languages	Wikipedia	https://en.wikipedia.org/wiki/List_of_programming_languages

Appendix II: Standardized Vocabularies

Detailed documentation of the list of standardized vocabularies described according to the minimal metadata suggested for PE24 Volatile Dataset can be found in https://goo.gl/T5oe9D.

Bibliography

Confucius. (2016). *The Analects of Confucius*. (J. Legge, Trans.). CreateSpace Independent Publishing Platform.

Laertius, D. (1925). *Diogenes Laertius: Lives of Eminent Philosophers, Volume I, Books 1-5*. (R. D. Hicks, Trans.). Cambridge/Mass. London: Harvard University Press.

Plato. (1921). *Plato, VII, Theaetetus. Sophist*. (H. N. Fowler, Trans.) (Loeb Classical Library edition). Cambridge, Mass.: Harvard University Press.

Plato. (1927). *Plato: Charmides, Alcibiades 1 & 2, Hipparchus, The Lovers, Theages, Minos, Epinomis*. (W. R. M. Lamb, Trans.) (Revised edition). Cambridge, Mass.: Harvard University Press.

Thesaurus Maintenance Working Group, VCC3, DARIAH EU. (2015). *Thesaurus Maintenance Methodological Outline*. Greece. Retrieved from http://www.backbonethesaurus.eu/sites/default/files/workingpaperonthesaurusmaintenance29_05_2015.pdf

Thesaurus Maintenance Working Group, VCC3, DARIAH EU. (2016). *A model for sustainable interoperable thesauri maintenance* (No. 1.1). Greece.

Thesaurus Maintenance Working Group, VCC3, DARIAH EU. (2017). *BBT –Submission and Connection Management tool* (No. 3.0). Greece. Retrieved from http://backbonethesaurus.eu/sites/default/files/BBT_SubmissionAndConnectionManagementTool_v3.0%20%28draft%29.pdf

Zhuangzi. (2003). *Zhuangzi: Basic Writings*. (B. Watson, Trans.) (1st edition). New York: Columbia University Press.