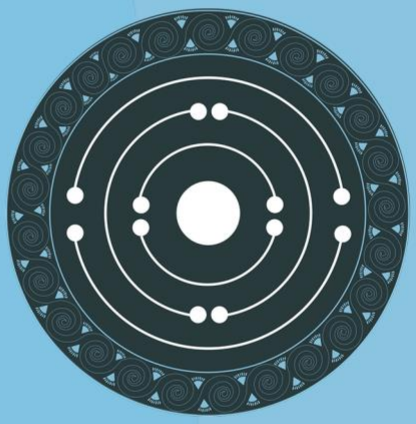

PARTHENOS

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

D5.7 Report on the Integration of Reference Resources

PARTNER(s) FORTH, CLARIN D, CNR, OAEW

DATE 31/01/2019

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

HORIZON 2020 - INFRADEV-4-2014/2015:

Grant Agreement No. 654119

PARTHENOS

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

Report on the Integration of Reference Resources

Deliverable Number	D5.7
Dissemination Level	[PUBLIC with CC-BY distribution]
Delivery date	31 January 2019
Status	Final
Author(s)	George Bruseker Alessia Bardi Felix Helfer Eleni Tsoulouha Elias Tzortzakakis Ksenia Zaytseva

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

Deliverable/Document Information

Deliverable nr./title	5.7
Document title	Report on the Integration of Reference Resources
Author(s)	George Bruseker, Alessia Bardi, Felix Helfer , Eleni Tsoulouha, Elias Tzortzakakis, Ksenia Zaytseva
Dissemination level/distribution	PUBLIC with CC-BY distribution

Document History

Version/date	Changes/approval	Author/Approved by
V 0.1 01.10.18	Update of 5.4 Strategy	George Bruseker
V 0.2 01.11.18	Update of Vocabulary Content	Felix Helfer
V 0.3 15.11.18	Update of BBT Work Content	Eleni Tsoulouha
V 0.4 01.12.18	Update of Implementation and Tools Section	Alessia Bardi, Eleni Tsoulouha, Elias Tzortzakakis, Ksenia Zaytseva
V 0.5 10.1.19	Review of controlled vocabularies decisions and update in document	Felix Helfer
V 0.6 15.1.19	Review, Enrichment and Update of BBT Content and Description	Eleni Tsoulouha
V 0.7 21.1.19	General Read Over and formatting control	George Bruseker
V 0.8 21.1.19	Additional formatting	George Bruseker
V 1.0 10.01.19	General Content Review	All authors, S. Bassett

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 normalised 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 the final report on the activities within PARTHENOS WP5 in collaboration with WP4 to adopt such a vocabulary management strategy and to identify high level standardised 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 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 harmonised 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 conceptualise 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 heterogeneous 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 harmonisation 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 harmonisation of results but at the risk, if carried out incorrectly, to stifle research by establishing inflexible canonical classifications unable to take into account new categorisations 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 catalysing 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 generalise 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

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

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 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, analyse previous efforts to meet the problem and their relative strengths and weaknesses, as a basis from which to learn and build. They analyse three major types of strategy that have been used to address this problem: the exhaustive subject classification system, taxonomic subject classification and the centralised 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 organisation 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

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

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 conclusions. The methodological reason that it is unfit for purpose as a top level meta-vocabulary is that, while it adopts hierarchical semantic organisation 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 organise 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 decentralisation 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 a bottom up analysis of existing classificatory systems, an open ended development of complete vocabulary including top terms and the ability to carry out this work as a distributed collective project. In brief, these principles can be explicated as follows.

semantic approach: refers to the framework of semantics, which lies at the heart of principled faceted classification. The resulting facets, then, are based on the intentional properties of terms –i.e. the essential characteristics expressing the substance of a concept, otherwise constructed, the necessary and sufficient conditions for belonging to a category.

The semantic approach of building a hierarchy of terms that spans disciplines and is based on the real world referents of terminologies is necessary to meet the integrative functionality envisioned for a meta-thesaurus. An approach that cannot critically analyse and integrate classification systems into a general system will not deliver the data integration capacity that a meta-vocabulary promises. That is to say, without a clear methodology for ascertaining the categorical semantics of classifications and aligning them to higher level agreed terms, the task of integration cannot be carried out since it will continuously be hampered by unexamined bias and ad hoc reasoning.

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

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

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 (Thesaurus Maintenance Working Group, VCC3 DARIAH EU, 2015).[7] As for the top-level terms, it follows that they should not be imposed by means of an *a priory* theory – as is the case in the Dewey Decimal System. Rather, they must be discovered through an analysis of existing sources, which ensures their functional and clear specification. Differently put, one must first perform an analysis of the classificatory systems that (s)he wants to integrate, before deciding on the most suitable top-level terms for the meta-thesaurus.

This bottom-up methodology guarantees that the definition of the high-level classes maintains their consistency as organising concepts relative to the scope provided by the domain of the classification systems that they aim to generalise. 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. In order to meet the needs of research, however, this bottom up approach must be left fundamentally open, meaning that higher level classes in the thesaurus are in principle open to revision. Therefore, it is a given that the overall classification will evolve over time, following the integration in BBT of new scientific domains and/or results of newly conducted research, resulting in enriched and expanded hierarchies.

This points to the final key element to the methodology propounded by the WG, namely that the construction and maintenance of the BBT is a collective effort carried out by a distributed group in an open, yet formal editorial process. In effect, what is proposed here is a federation of vocabularies that are brought together through an open-ended backbone and which are subject to tighter integration whenever deemed necessary. Such a situation meets the need of research communities for access to integrated classifications of more specific resources/research objects. 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 generalisation and in very specific communities of practice.

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

[7] Thesaurus Maintenance Working Group (VCC3, DARIAH EU). (2015). Thesaurus Maintenance; Methodological Outline.

activities
— disciplines
— human interactions
— intentional destructions
— functions
natural processes
— natural geneses
— natural destructions
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 generalisations 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. For an example of this case see the integration of the PARTHENOS Entities Vocabularies Place Types hierarchy described in Section 5.1.4.

Where a candidate vocabulary is a flat list with no declared top term, it may be necessary to introduce auxiliary intermediate generalisations in the source classificatory system which would then, in turn, link into the BBT in a semantically consistent way.[8] 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. End-users browsing the BBT will be making use of different classification systems for the same general class of things. The browsing of these rich interconnections can be supported by a SKOS vocabulary browser such as the SKOSMOS system deployed as ACDH Vocabularies.

In the data enrichment and development scenario, users of BBT may make use of BBTalk—formerly named Submission and Connection Management Tool—(Thesaurus Maintenance Working Group, VCC3, DARIAH EU, 2017),[9] developed by FORTH-ICS, within the framework of the Thesaurus Maintenance Working Group (VCC3, DARIAH EU, 2017). In the event that end users cannot find an appropriate high-level facet or hierarchy under which to place terms of their classificatory system, a process of discussing the extension and expansion of the BBT itself gets launched. The methodology for managing these discussions is discussed below.

Assuming that someone wishes to link his/her thesaurus to BBT and parts of this thesaurus do not integrate well with BBT, then (s)he can propose a new facet or a new hierarchy within one of BBT's facets which can accommodate the terms. It is possible that the BBT underspecifies semantic/conceptual distinctions that are particularly prominent in a specialist thesaurus. Integrating such a thesaurus with the BBT might call for changes in the overall scope of BBT –manifest by fine-grained distinctions, available even for high-level hierarchies. Aside proposing new terms, end users faced with this situation can propose to split – or otherwise modify – BBT facets and hierarchies. Conversely, there might be reason to broaden the scope of the BBT, in which case part of its structure may become irrelevant. It is possible then, that end-users request for facets/hierarchies to be deleted, merged, or otherwise modified. This is part of the open ended, revisability for the meta-thesaurus strategy.

[8] Examples of how that case was handled can be found throughout section 5 and in summary in table 32.
[9] Submission and Connection Management Tool (BBTalk):
[1] https://www.backbonethesaurus.eu/BBTalk/, see detailed functionality in:
[2] https://www.backbonethesaurus.eu/BBTalk/Manuals/BBTalk-UserManual.pdf

The functionalities mentioned so far are featured in BBTalk and correspond to connections (vocabulary integration) and submissions (new terms, splitting, merging and deleting terms, modifying terms). The latter are discussed among interested parties –i.e. (s)he who submits a proposed change, the curating team of BBT, specialist thesauri maintainers who have integrated their thesauri with BBT and whose thesauri will be affected by any changes implemented on the BBT. Domain experts in their respective fields willing to help settle pressing arguments can also be invited to participate in the discussion through BBTalk. Decisions in favour or against a specific request to change the structure of BBT are reached by vote.[10] The specifics of the implementation of BBTalk elaborated further in Section 2.6.

Overall, the BBTalk forms a communication system supporting discussions regarding (a) connections effected and (b) proposed changes on the current versions of the BBT, among specialist thesauri maintainers, the curating team of BBT and domain experts in their respective fields, willing to help settle pressing issues.

2.4. Creating a Reference Data Integration Workflow

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 query-able 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 standardisation. Such data values are usually ‘type’ fields such as ‘subject’ or ‘material’ or ‘object kind’ etc. Additional data values that are susceptible to standardisation 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 periodisation structure. Likewise, data values encoded in fields for expressing information such as ‘place’ which refer to well known geographic units can be standardised against

[10] https://www.backbonethesaurus.eu/BBTalk/Manuals/BBTalk-UserManual.pdf

well known gazetteers. This standardisation or matching of vocabularies and non-categorical reference resources ensures harmonisation of existing standards and 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.

To put the practice of standard reference data integration into practice, a general workflow had to be established in order to organise a consistent and sustainable process.

The PARTHENOS Project data integration scenario was taken as the test case upon which to build up a workflow scenario that would support a cohesive reference data integration strategy offering a sustainable process of reference data integration. In defining this workflow several key steps were identified as key for managing reference data integration. These steps include: identification, discovery, creation, registration, integration and implementation. The PARTHENOS Project test case presents all of the typical challenges present in a reference data integration scenario: heterogenous data inputs (including application of varied data standards, misspelled data, and incorrect data), extant and non-extant data standards, lack of top level terms for checking conceptual consistency etc. The overall integration of the contributing RI metadata for the establishment of a functional Joint Resource Registry in PARTHENOS required the mitigation of these factors by a standardising process to quality reference resources. For this reason, the activity of performing the integration of PARTHENOS data itself was able to form both a primary task in the implementation of the overall project but also as the test bed scenario for the formalisation of a standardised process supporting integration of the relevant reference resources under the general umbrella of the Back Bone Thesaurus integration system.

The main phases of reference data integration identified were:

Identification: this phase consists in an analysis of the data structure to be implemented and the standardised reference resources required for descriptors identified as susceptible

to standardisation. A full documentation of the kinds of reference resources is generated in order to create the requirements list for discovery of reference resources.

Discovery: this phase works from the identified reference resource requirements list in order to begin a research process to locate available, relevant and extant reference resources suitable to the domain of documentation and its standardisable data elements.

Creation: the discovery phase of reference resource research must be supplemented by the enabling of tools for the creation and management of new, standard reference resources in the case where suitable extant reference resources do not already exist. In many cases, terminology for standardisation does not already exist and must be generated from scratch. In this case, the generation of such lists can be made sustainable through the use of SKOS compliant thesauri management systems.

Registration: the sustainability of the discovered and created reference resources is only possible if they themselves, as data sources, and properly documented and their provenance documented. Registration of standard reference resources in a documentation system creates a corpus of available standard resources that can be reused throughout the life cycle of the project.

Integration: the BBT method for providing long-integration of reference resources amongst themselves provides the key novel feature to the PARTHENOS reference resource integration proposal. The BBT allows a check on the quality of reference resources by a) checking their conceptual consistency and b) enabling their harmonisation to higher level terms. This enables their potential discoverability through integration to common, high level agreed terms. This step requires the provision of tools enabling access to the BBT, proposal of new terms and of contributions of existing vocabularies as extensions of high level, canonical terms.

Implementation: the final function of the establishment of a register of reference resources is the ability to adopt the documented reference resources within a broader data integration workflow and use them for the standardisation of data values within this workflow. The step of implementation offers tools that support the standardisation of values in datasets adopted for integration in a traceable fashion.

The generic workflow presented here can be schematised as follows:

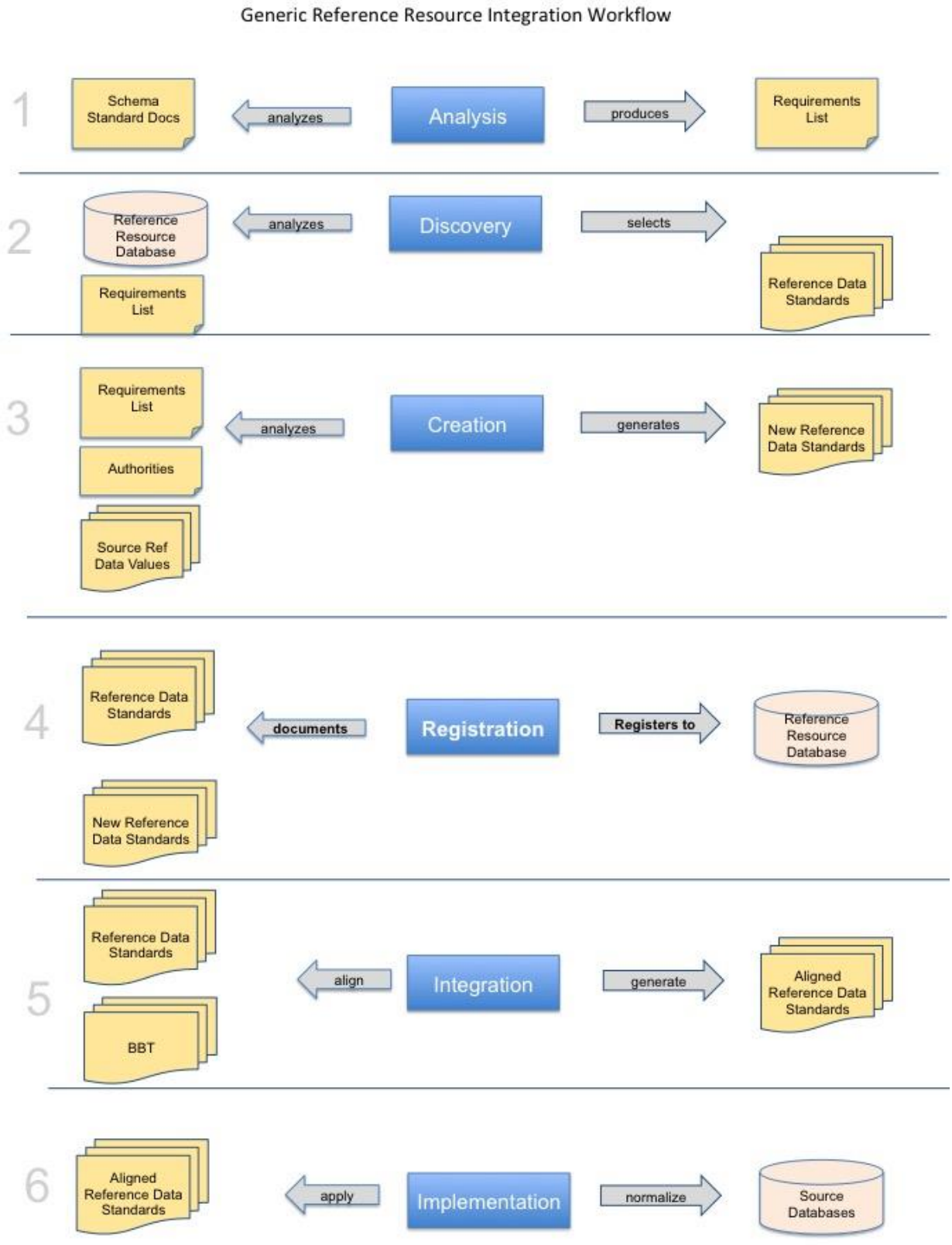

Figure 1: General Reference Resource Integration Workflow

The above workflow aims to integrate the Back Bone Thesaurus strategy into the broader framework of a large scale data integration. The scope for selecting, creating and curating a set of reusable reference data resources is given by the overall data integration project. The Back Bone Thesaurus is conceived as the control component which ensures that the selected resources are conceptually consistent and that they can be published to a wider audience for re-use. This general schema has to be put into particular practice through the selection of a set of tools which can be interrelated in a particular workflow for implementation.

2.5. Testing Reference Data Integration Workflow for PARTHENOS

Having organised the above generic workflow, its effectiveness was tested by setting up a working scenario for the integration of the thesauri necessary for the realisation of the Joint Resource Registry. In this section, we will expand on each of the identified steps, the task involved, the tools required, the output they generate and how they were handled within the PARTHENOS Project.

Identification

Task: the primary function of this task is to create the requirements list for reference data integration. It works from a target schema which will be used as the standardisation framework and performs an analysis of what categorical and particular reference resources are required to support this schema.

Tool: The basic tool necessary for this phase is the documentation of the target schema to be used for overall data integration.

Output: The required output of this task is a list of the names of the fields/descriptors requiring standardisation in the target data model.

Test Case: In the PARTHENOS project, the PARTHENOS Entities Model and PARTHENOS Minimal Metadata specification provide the target schema for overall data integration to the Joint Resource Registry. Because the PARTHENOS Entities data model is CIDOC CRM compliant, the task of identifying fields for standardisation centres on finding fields specified as holding instances of E55 Type, as well as looking for fields where well known and documented particulars are referenced such Period, Place and Schema. This phase was carried out by marking up the data standard documents with the additional requirements for standard reference resources.

The execution of this activity is documented in section 3.1 of this report.

Discovery

Task: the primary function of this task is to seek and document existing reference data standards relevant and adequate to the identified needs of the target schema and its application.

Tool: online reference resource databases provide the primary source for discovery of extant thesauri and vocabulary. These online resources provide the primary tool for the discovery stage.

Output: list of candidate reference resources for integration associated to requirements list generated by the identification phase.

Test Case: the online resources consulted for carrying out the discovery process in the PARTHENOS Project included: the Basel Registry of Thesauri, Ontologies & Classifications (BARTOC) [11], the Open Metadata Registry [12] and the Linked Open Vocabularies (LOV)[13]. Candidate standards for matching to the requirements list were gathered in a spreadsheet for discussion and consideration.

[11] https://bartoc.org/

[12] http://metadataregistry.org/

[13] http://lov.okfn.org/dataset/lov/

The execution of this activity is documented in section 3.1 of this report and the list of candidate standards given in the appendix.

Creation

Task: the primary goal of this task is to fill in gaps in available standard reference data resources by generating new resources, documenting them and making them curatable and enrichable.

Tool: the required tools are of two types. For the creation of new standardised reference resources, input data from publications of standardised terms and/or raw value input data from the list of extant data values used in data sources provides necessary primary material on which to generate a new standardised reference resource. Moreover, a reference resource data management software is required in order to generate, curate and create versions of these assets.

Output: new standardised reference resources usually in the form of SKOS

Test Case: in many cases, the PARTHENOS Minimal Metadata specification called for standardisation of values for which no extant reference resource was available. In this case, a consultation of literature relevant to the topic was undertaken and official terminology lists in publications were sought after. Barring the availability of such well defined extant lists, reference to the data values within the datasets to be integrated was made in order to carry out an analysis to find the scope of existing use and identify the most common terms. In order to generate the new standard data sources, the THEMAS management system was adopted.

This resulting new lists generated during this step are documented in Section 3.1 below while the use and application of the THEMAS tool is described in the next section.

Registration

Task: the function of this step is to provide provenance to the reference resource data integration process by generating appropriate metadata for each selected reference resource and publishing it.

Tool: any basic documentation medium, well formatted and published can be used for this step.

Output: register of reference resources

Test Case: the PARTHENOS Project used Google sheets and followed the PARTHENOS Entities Model in order to document each reference data source as a dataset asset. This spreadsheet was then mapped to PARTHENOS Entities Model and transformed into the Joint Resource Registry.

The results of this activity, the register of official sources, is included as an appendix to this document.

Integration

Task: in order to create a more broadly compatible and sustainable set of reference resources, all categorical reference data resources should be checked against and aligned to a higher level meta-thesaurus. Checking the adopted categorical reference resources against the Back Bone Thesaurus and aligning them to it, provides both a conceptual validity check but also a means to make reference resources more widely available through their integration into this upper level terminology.

Tool: required for this step is a meta-level thesaurus and tools by which to link specialist thesauri to it.

Test Case: In the PARTHENOS Project we adopted the Back Bone Thesaurus as the top level thesaurus against which to align, given its scope as a top level thesaurus for digital humanities applications. We adopted the BBTalk tool for carrying out the integration

process with the BBT. Publication of the resultant aligned SKOS was undertaken in the ACDH “Name Here” environment.

The process and results of this activity are documented in Section 5 of this report. A description of the BBTalk and ACDH Vocabularies tools are given in the section below.

Implementation

Task: the ultimate aim of the adoption of reference resources in data integration is to enable better interoperability and comparability at the level of data values. This task aims to harmonise data values in the overall data integration by adopting the chosen reference resources for data harmonisation on specific data fields in the target data schema.

Tool: required for this step is a tool that allows data value transforms based on lists and matching criteria. Ideally, the tool should enable a repeatable and modifiable process.

Test Case: In the PARTHENOS Project, we adopted the D-Net Record Cleaner tool in order to implement the cleaning of require data fields with the selected standards.

The description of the D-Net tool can be found in the section below. For a description of the process of its implementation see D6.2.

2.6. PARTHENOS Reference Resource Data Integration Implementation

In order to implement the generic workflow identified and tested above, a more specific workflow was devised to work with the set of tools that were chosen to carry out the work. In this section we document the tools chosen for carrying out the task, their role within the overall integration process, and the reasons for their selection. We then provide the basic workflow for carrying out this process with the selected set of tools.

Google Sheets

Function: Identification, Discovery, Registration

Description: the well-known commercial offering of Google offers online collaborative spreadsheet functionality.

Reasons for Adoption: Its chief advantage in this process is flexibility and share-ability. Documentation structures can be quickly generated and shared with partners for completion. Export facilities make it possible to output the data in XML. The flexibility is especially useful in the identification and discovery processes. The tool was also adopted to create the place for documenting and registering the selected reference resource datasets. Ideally, the registry could be made into a more formal data structure.

THEMAS

Function: Creation

Description: Thesaurus Management System – THEMAS is an open-source, workflow-based web application system used for the creation, development and management of thesauri following the guidelines of ISO 25964-1:2011 and ISO 25964-2:2013.

Reasons for Adoption: Domain specific terminologies can directly be created in THEMAS or loaded in a bulk mode following a quite simple XML Schema structure. The user-role based thesaurus development workflow followed, allows the simultaneous work of large user groups of different domain specific expertise on the same thesaurus, following an adjustable set of consistency rules, while the smaller higher expertise user-group is thus supported in the decision of the most suitable thesaurus structure. Hierarchical and associative semantic relations, translations, scope notes etc. extend and clarify the initial terminology set while the overall thesaurus can be aligned to the Back Bone Thesaurus, thus crossing the domain specific boundaries and connecting to cross disciplines queries and terminologies. The thesaurus development outcome can be exported in XML or SKOS

format for further offline usage or processing or directly integrated to heterogeneous systems after appropriate THEMAS policy configuration.

Indicative Screens:

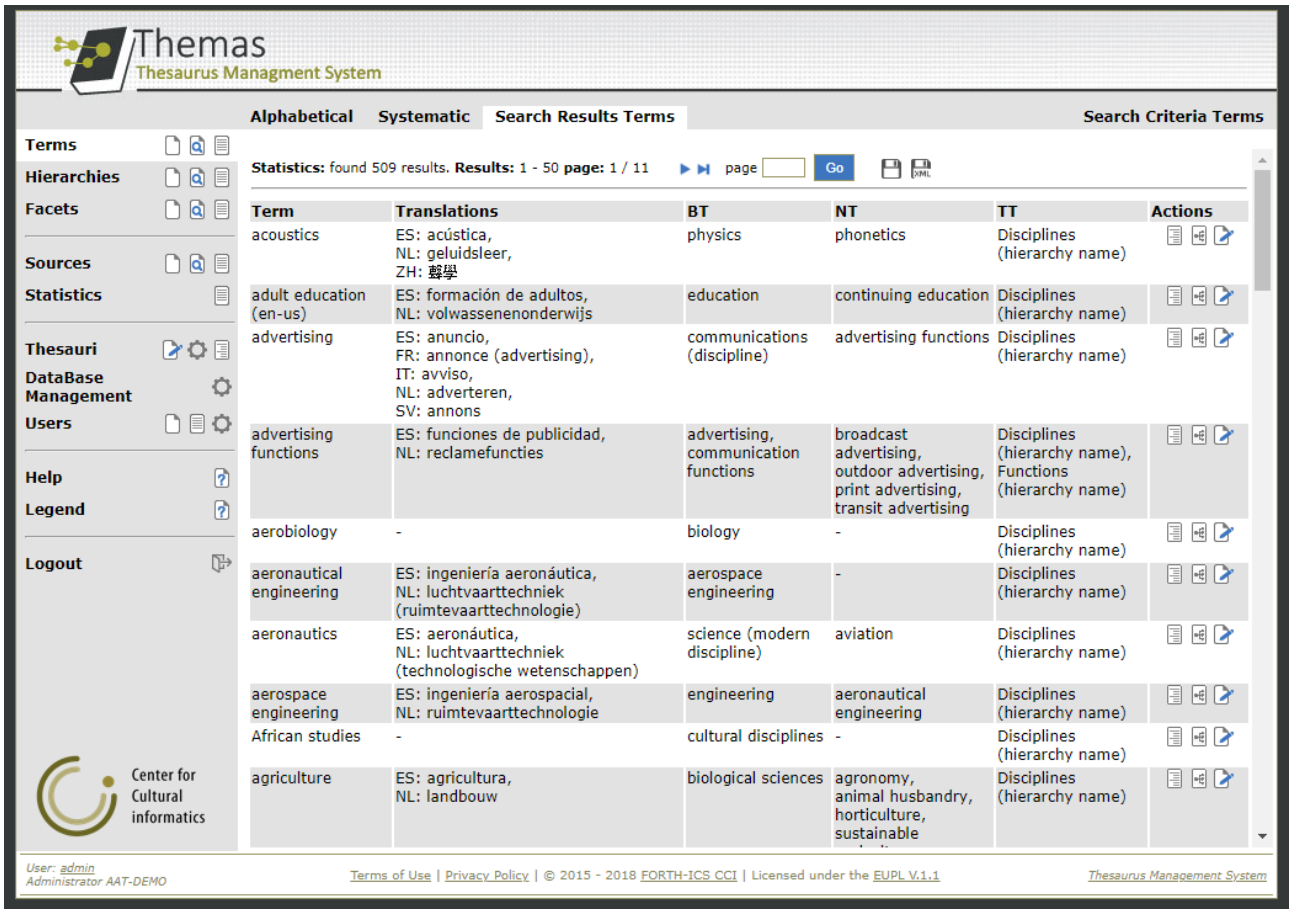

Figure 2: THEMAS Tool Interface

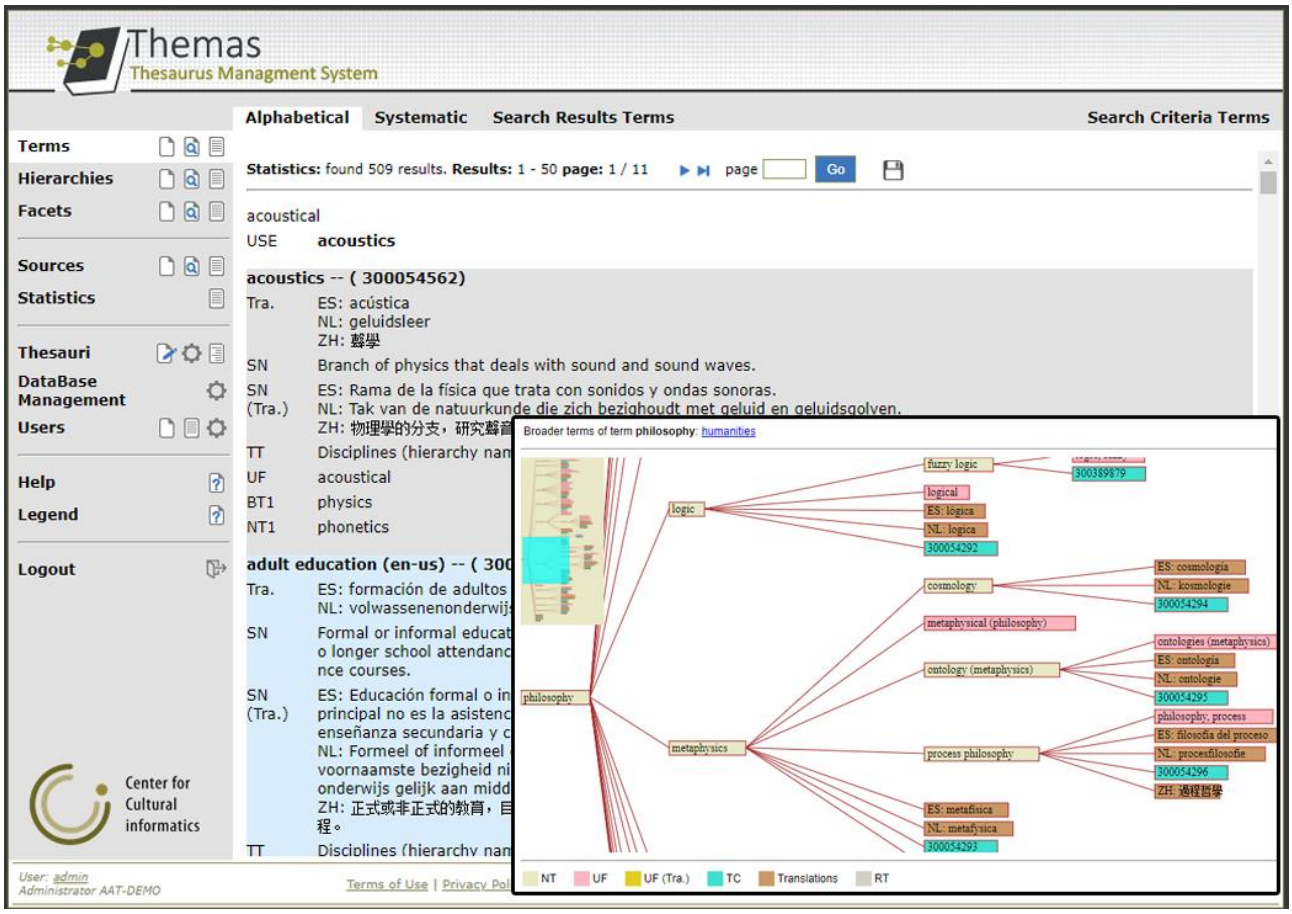

Figure 3: THEMAS Tool Interface, Visualisations

D-NET Metadata Inspector and Cleaner[14]

Function: Implementation

Description:

The Metadata Cleaner is a D-NET service that harmonises values in metadata records based on a set of thesauri. A D-NET thesaurus consists of a controlled vocabulary that is a list of authoritative terms together with associations between terms and their synonyms. Data curators – typically based on instructions from data providers and domain experts – are provided with user interfaces to create/remove vocabularies and edit them to add/remove new terms and their synonyms. Given a metadata format, the metadata cleaner service can be configured to associate the metadata fields to specific

[14] This partial description given of D-Net in this section is copied here for convenience sake from the full report on tools and services given in D6.2 Report on Services and Tools of the PARTHENOS Project.

vocabularies. The service, provided records conforming to the metadata format, processes the records to clean field values according to the given associations between fields and vocabularies. Specifically, field values are replaced by a vocabulary term only if the value falls in the synonym list for the term. If no match is found, the field is marked as 'invalid'. The 'invalid' marker is exploited by the Metadata Inspector to highlight non-cleaned records and suggest the update of D-NET vocabularies or the update of the values in the input record.

Reasons for Adoption:

The inclusion of the Metadata Cleaner was not initially planned, because the value cleaning can also be performed by defining specific rules in the X3ML mappings. However, the PARTHENOS Consortium agreed that a mechanism to ensure that all controlled fields (i.e. metadata fields whose values must comply to a controlled vocabulary) contain valid values was needed. At this goal, CNR-ISTI proposed to include in the D-NET instance of PARTHENOS the Metadata Cleaner so that, in the transformation phase of the aggregation workflow, each record is transformed by the X3ML Engine and, afterwards, the controlled fields are further cleaned by the Metadata Cleaner. If a controlled field cannot be harmonised according to the proper vocabulary, the record is marked in order to enable inspection via the Metadata Inspector.

Indicative Screens:

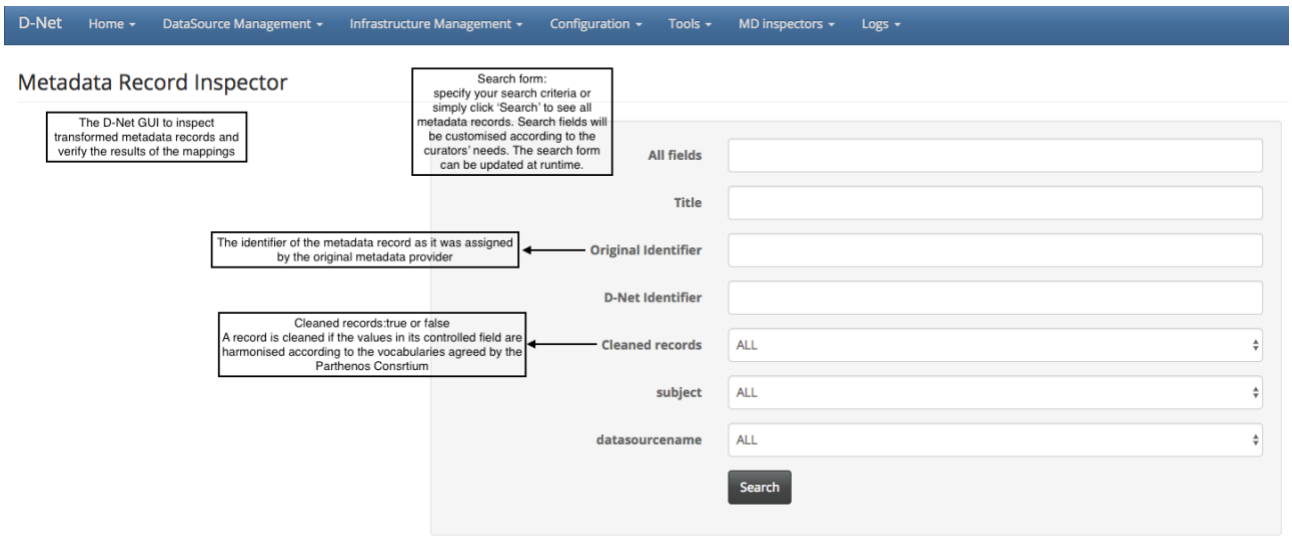

Figure 4: The main search form of the Metadata Inspector

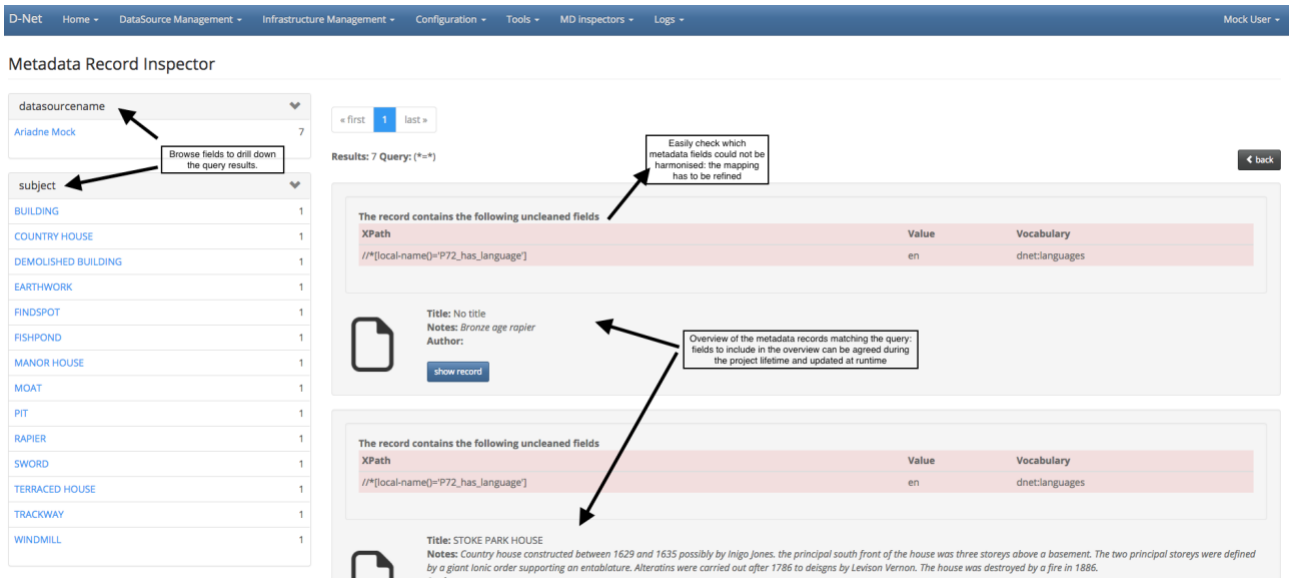

Figure 5: The Metadata Inspector shows metadata records with "uncleaned" fields

BBTalk

Description:

BBTalk is the software component that was developed by FORTH-ICS within the framework of the Thesaurus Maintenance Working Group (VCC3, DARIAH EU, 2017), in order to manage the functions of submission of new terms and changes in BBT, as well as to connect specialist thesauri into the federated system (Fig. 6). It is used as an alignment tool by researchers and institutions holding specialist thesauri that they want to link and publish to BBT (Fig. 7). It also serves as a communication system, supporting discussions between the curators of BBT and its users. Researchers can use BBTalk to submit requests for changes regarding the terms and hierarchies of the BBT (Fig. 8). BBTalk supports discussions between specialist thesauri maintainers and the curators of the BBT regarding proposed changes and the connections realised (Fig. 9). It further keeps track of the different versions of the BBT and the history of submissions and serves as a record of the relevant discussions related to the evolution of the thesaurus.

Reasons for Adoption:

BBTalk works as a maintenance system for the BBT, supporting the implementation of the proposed changes, based on the accepted submissions and releases of the new versions of the BBT. It further supports discussions regarding the said changes and enables the alignment of specialist thesauri to BBT.

Indicative Screens:

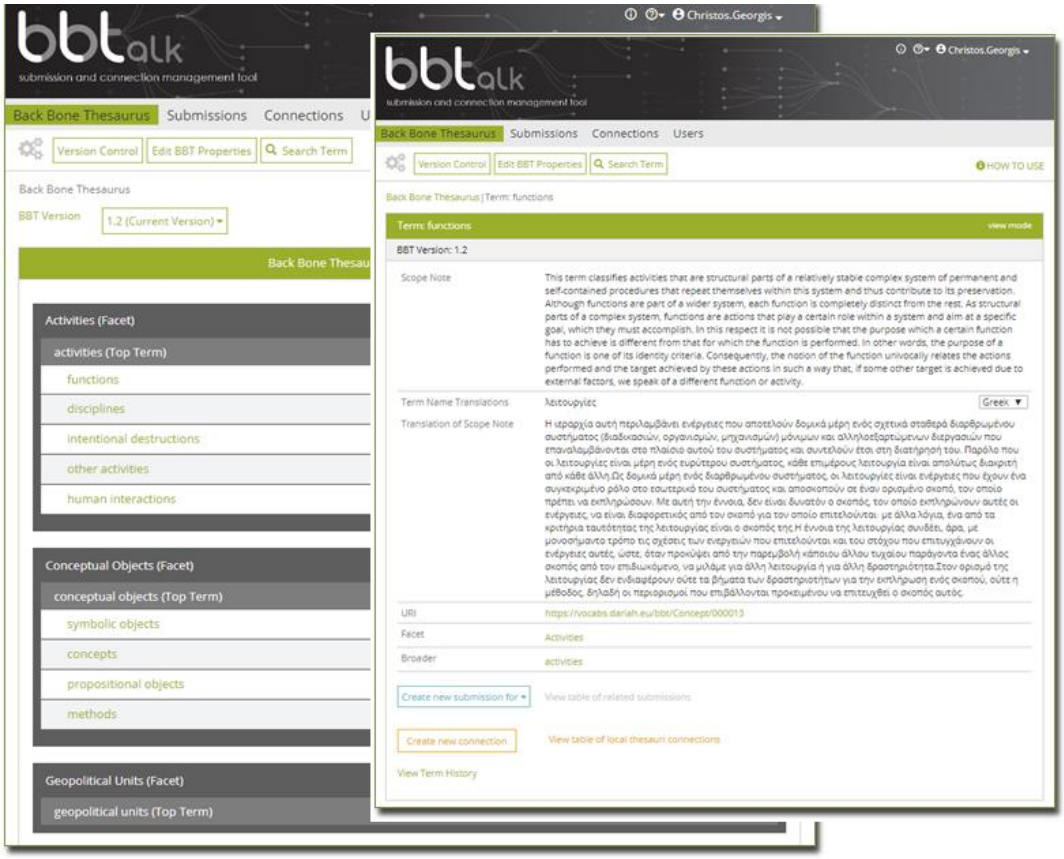

Figure 6: BBTalk, the BBT management system.

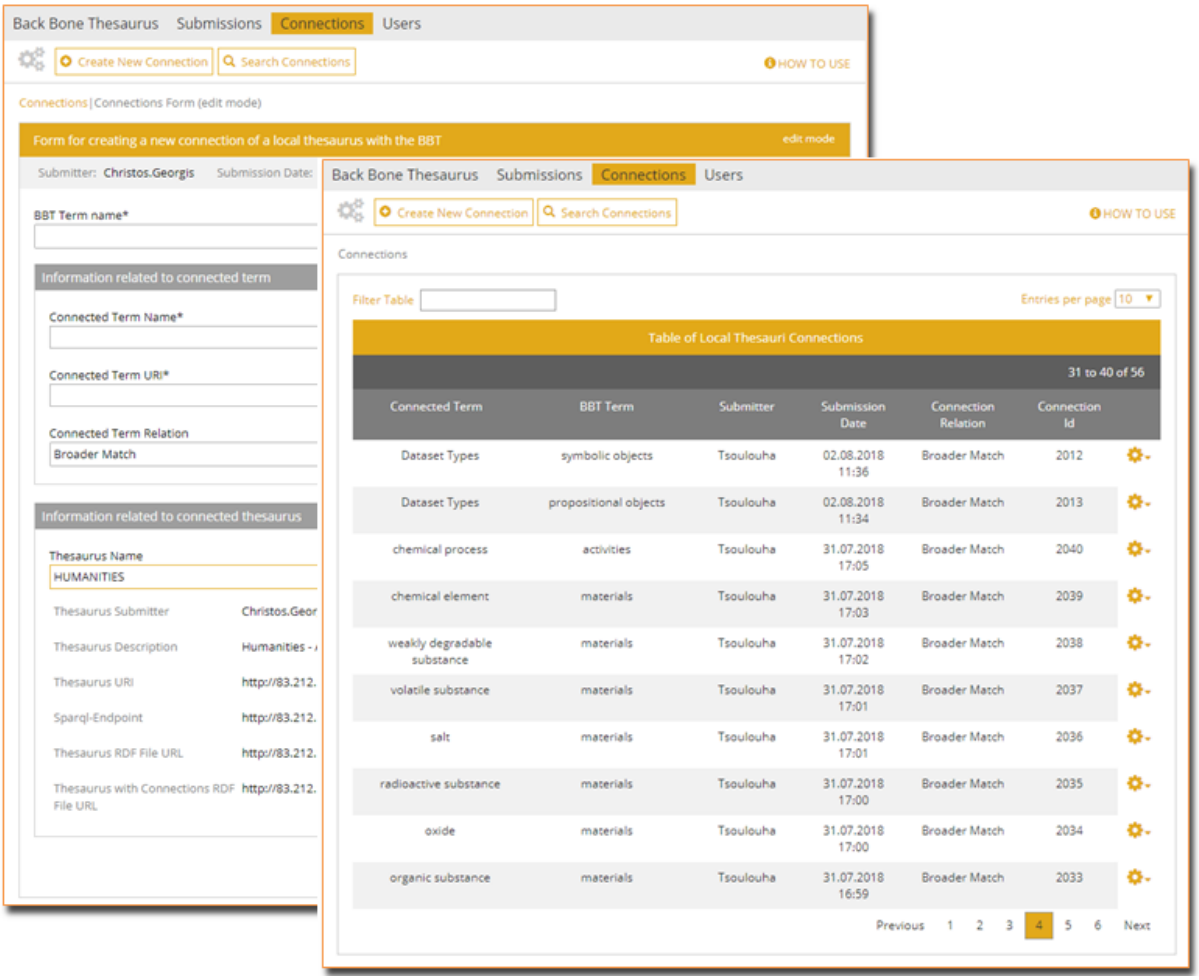

Figure 7: BBTalk, the connections management interface

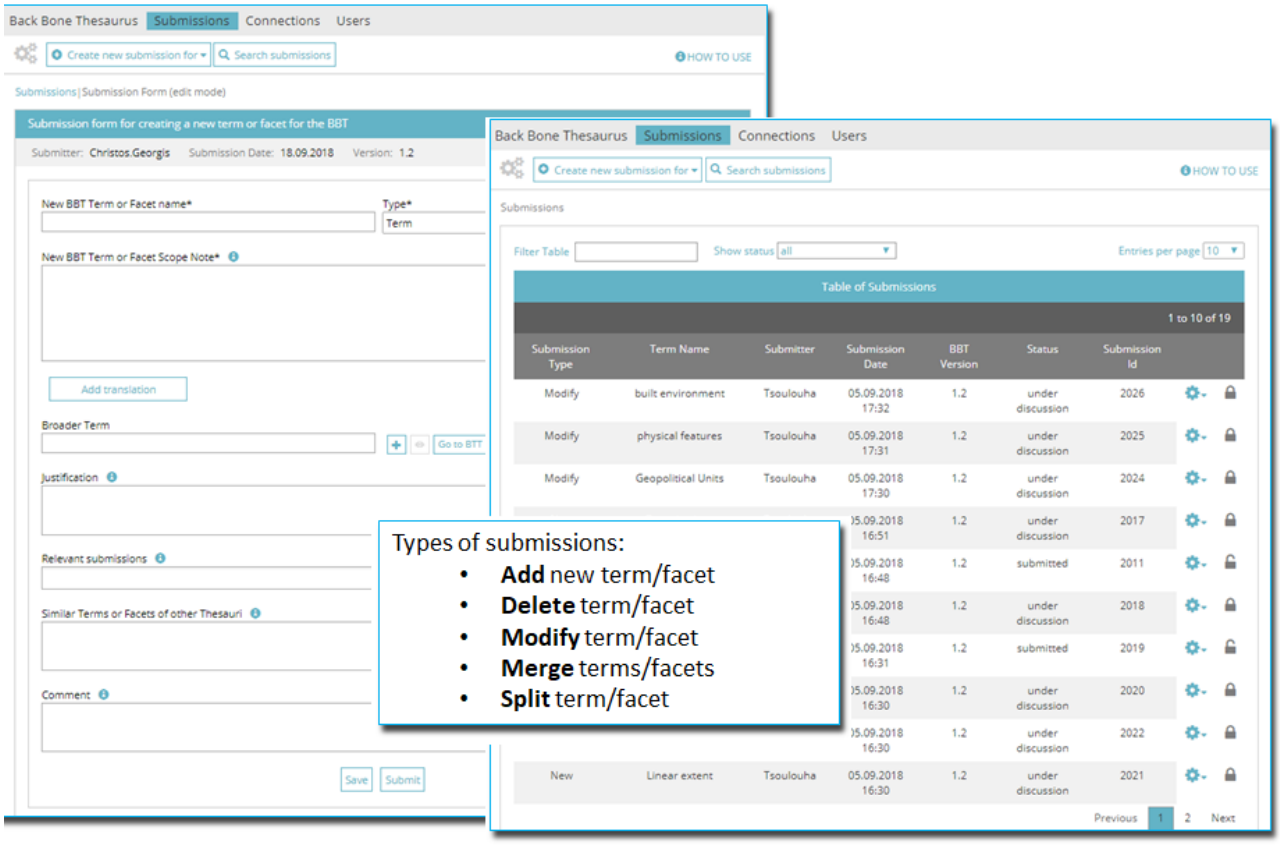

Figure 8: BBTalk, the submissions and management tab.

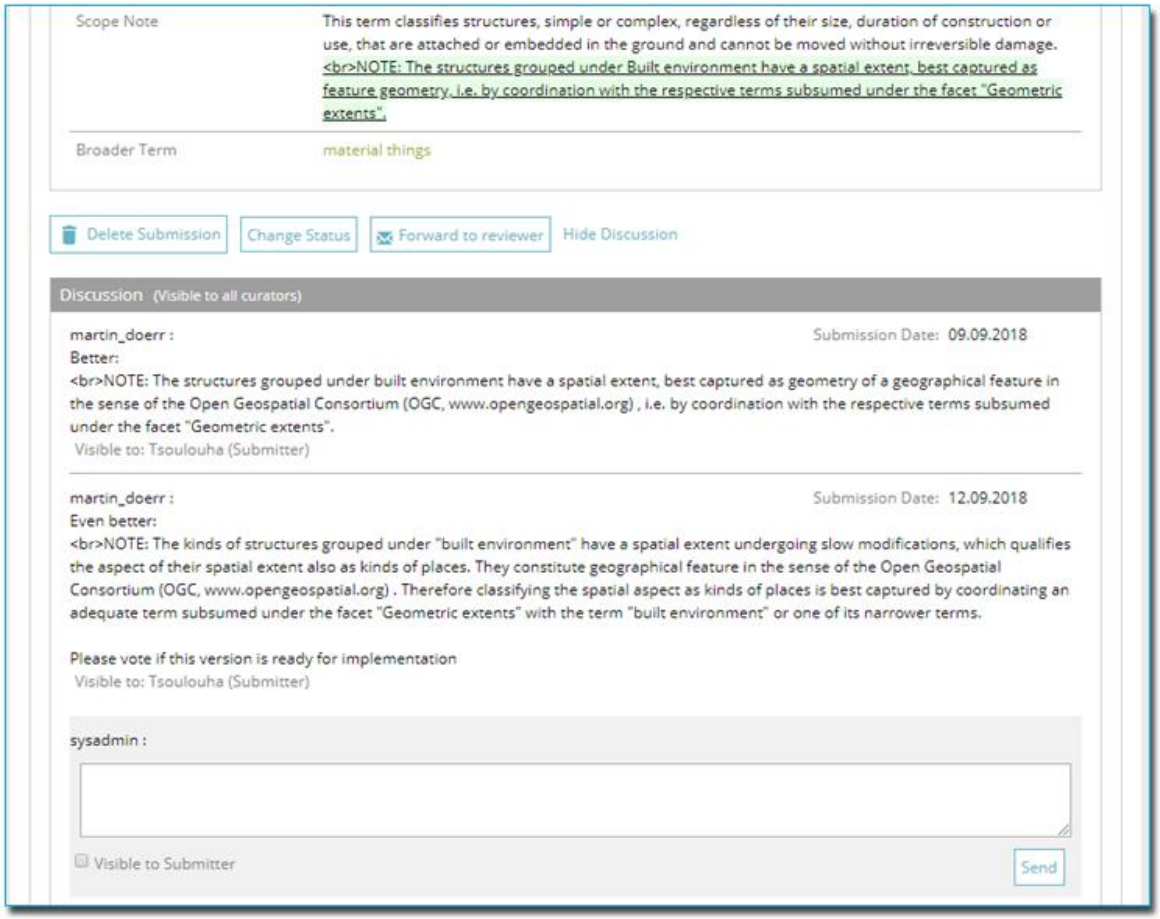

Figure 9: BBTalk, the discussion management system.

ACDH Vocabularies

Function: Integration, Discovery

Description: The ACDH provides a vocabulary repository service that allows for collaborative maintenance and publication of vocabularies and taxonomies of any kind.
The system is based on the open-source software Skosmos which uses SKOS as the underlying data model. Skosmos offers browsing of vocabularies with structured concept displays and visualisation of concept hierarchies. Each concept has a unique and resolvable URI. Vocabularies can be searched with a search interface or by consulting an alphabetical or thematic index. Vocabularies can be accessed via a REST API, to allow for Linked Data.

Reasons for Adoption: ACDH Vocabularies is a long-term project within ACDH infrastructure. This ensures a stable workflow and maintenance of all controlled vocabularies published in service and guarantees the URIs resolvability for Semantic Web. The service provides RDF/XML, Turtle and JSON-LD serialisation for individual concepts and download for a whole vocabulary in RDF/XML or Turtle. ACDH Vocabularies as a service suite is still expanding already providing SPARQL endpoint to query all vocabularies and aiming in future to provide a visualisation component to analyse the relationships among linked concepts.

Indicative Screens:

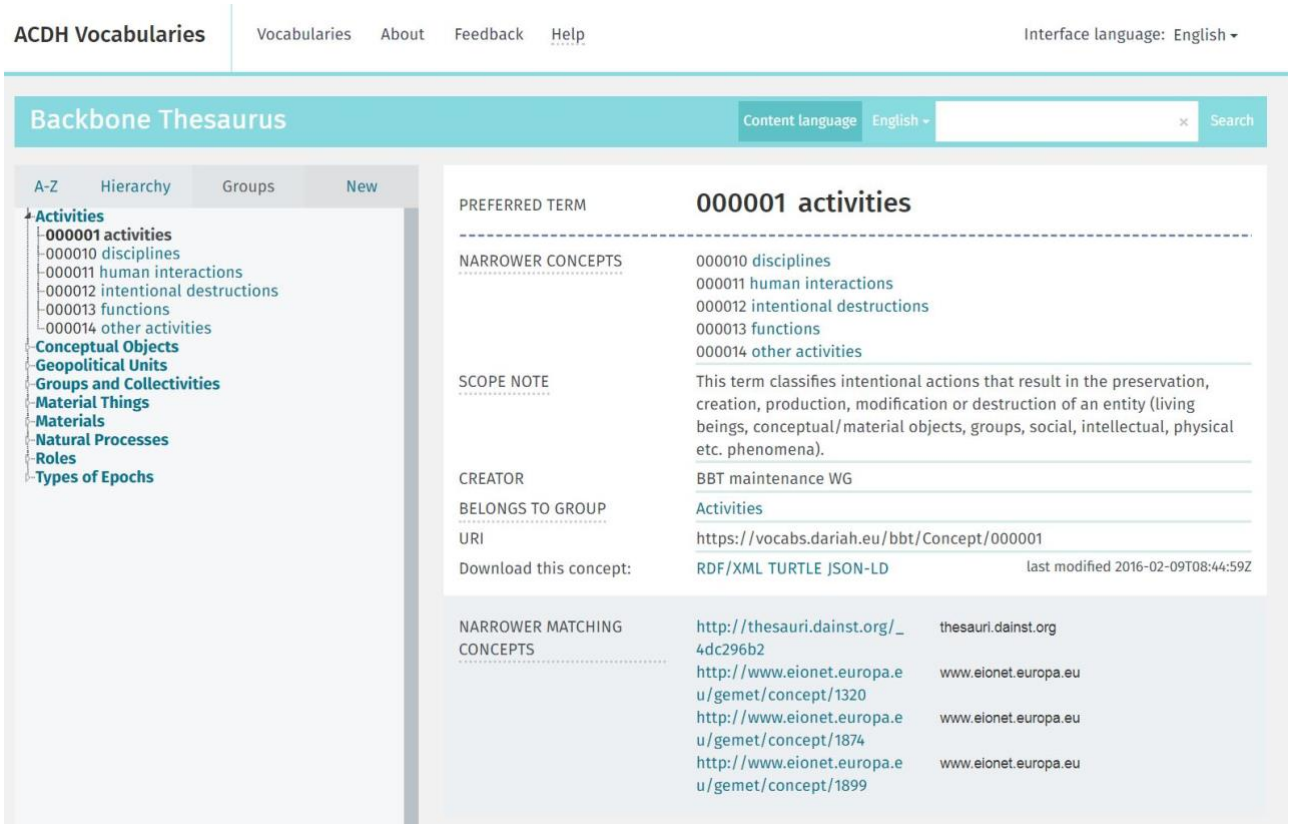

Figure 10: ACDH Vocabularies, browsing facility.

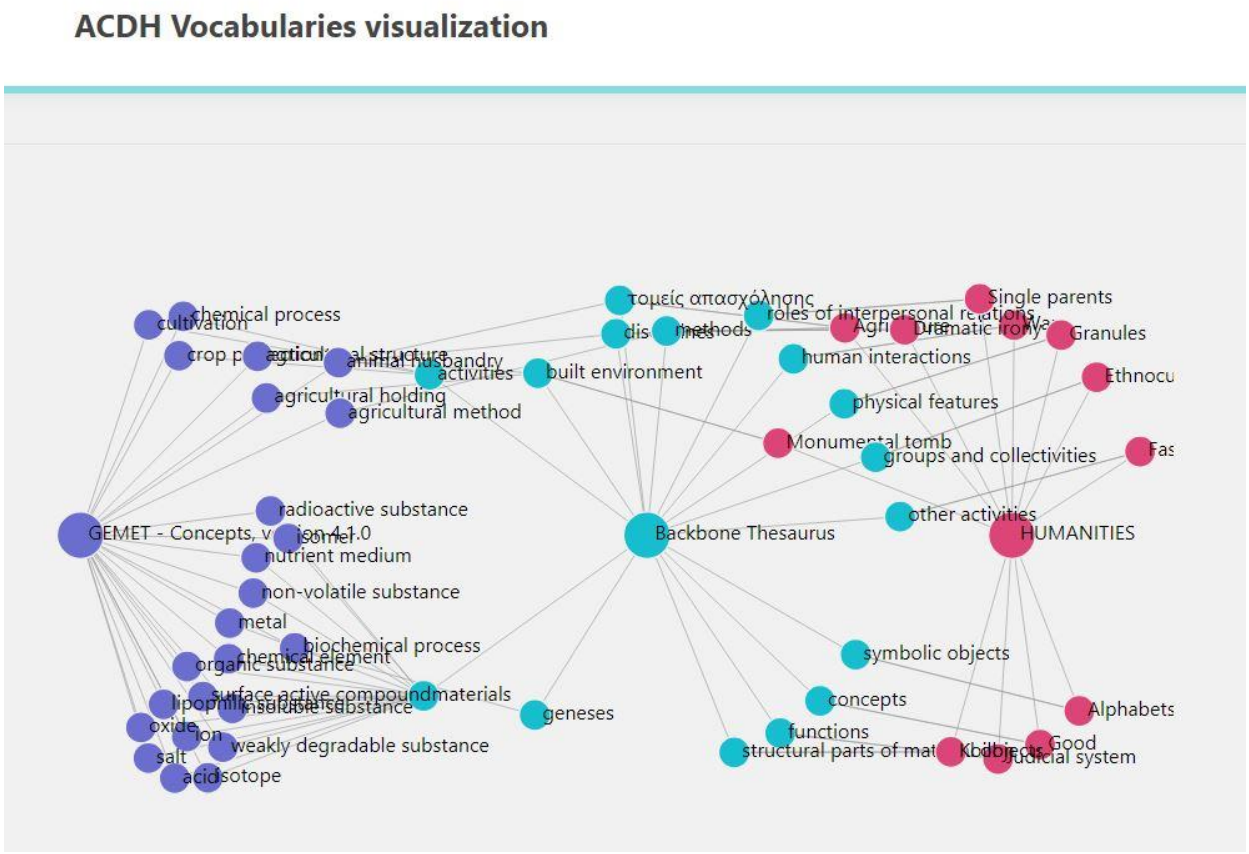

Figure 11: ACDH Vocabularies, visualisation functionalities.

Bringing the tools together effectively to create a functional workflow involved creating a path from the moment of selection and creation of vocabularies to their encoding in a terminology management system and their application for data harmonisation in the overall aggregation process but also their implementation into the overall BBT scheme. In order to execute this, the following workflow was devised:

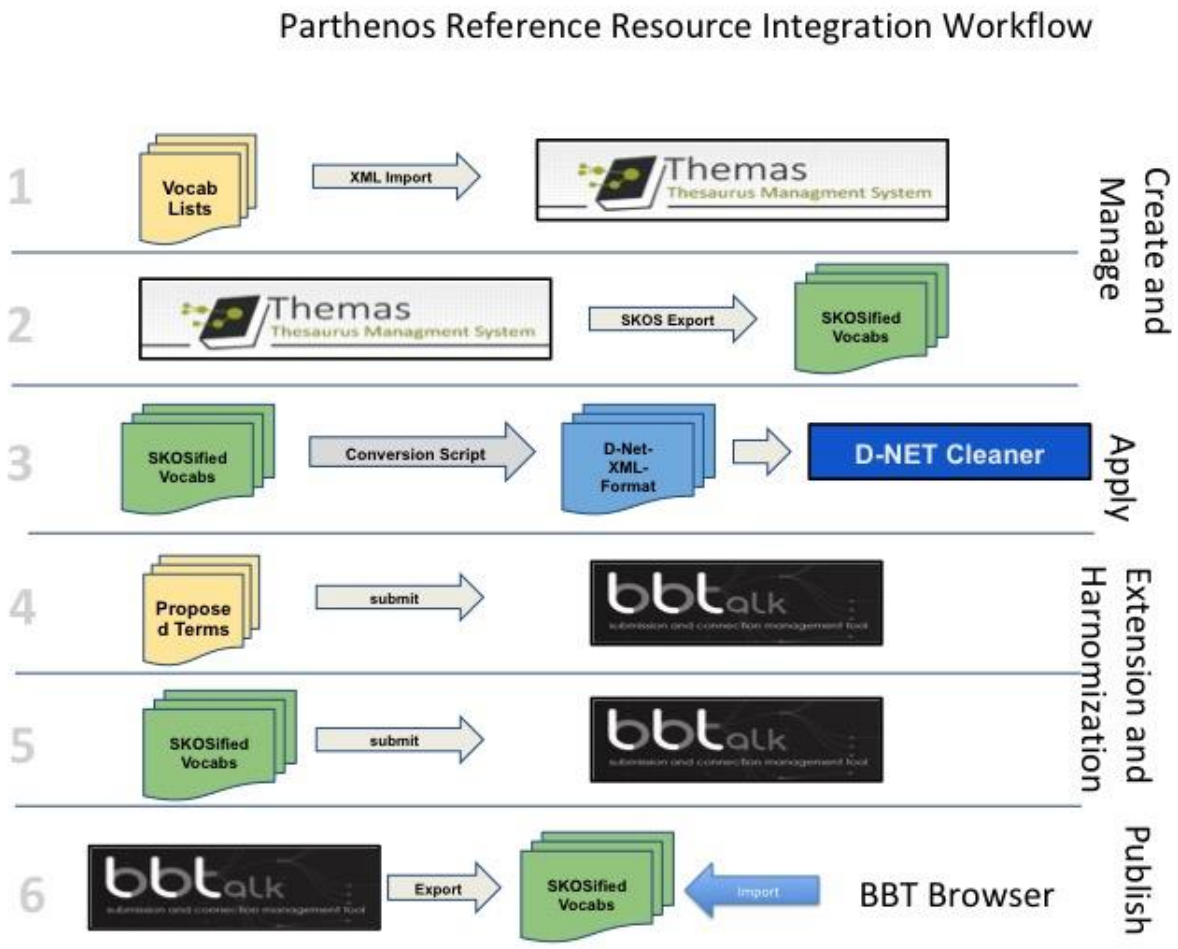

Figure 12: Implemented PARTHENOS Reference Resources Integration Workflow

In this implementation workflow, we execute the general workflow plan from the point of having selected existing sources and decided which new sources to create. We adopt the THEMAS tool in step 1 to import existing vocabularies for management and in order to enable the creation and curation of the new vocabularies. Step 2 create SKOS exports of the vocabularies for use in the following steps. The SKOS exports generated in step 2 are transformed in step 3 into an xml format used by the D-Net cleaner and imported for use in

that service for data cleaning on the aggregate data products of the overall data integration process. In step 4, after the analysis of the vocabularies and their relation to the BBT meta-thesaurus, new terms are proposed, where necessary, in order to extend the scope of BBT to the function of the data particular data integration process. In step 5, all adopted vocabularies are proposed as extension of the BBT top level terms. These new terms and submissions are reviewed and curated by a curation committee. More on this is described in section 2.3 and section 5. At the final step, the product of the BBT and its connected vocabularies can be exported to and published in the SKOS browser service setup by ACDH.

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 organisation 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 (JRR) 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.[15] It is at this point that the need for a set of standardised 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 harmonised 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

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

standard vocabularies to which to harmonise. Therefore, the research described in section 3.2 and continuing in Section 4 implemented the identification, discovery and creation steps of the general reference resource integration workflow described in section 2.4 This first research with regards to building integrated reference resources, is necessary to find appropriate reference resources for the overall data integration to the JRR.

In what follows, we will describe 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 Standardised 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 standardised vocabularies in order to move beyond schema matching to integrated ways of classifying and identify individual resources that will enable tightly integrated and highly query-able data.

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 standardised 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: Colour 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 specialisations 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 metadata set for PE35 Project is as follows:

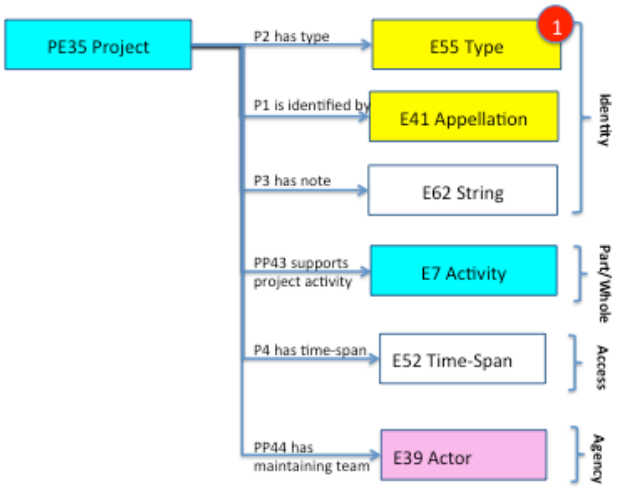

Figure 13: PE35 Project Minimal Metadata Application Profile Schema

The PE35 Project minimal metadata application profile makes reference to one field which require standardisation according to common vocabularies. The following table summarises 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, licence-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.

Table 4: PE1 Application Profile Minimal Metadata Configuration

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

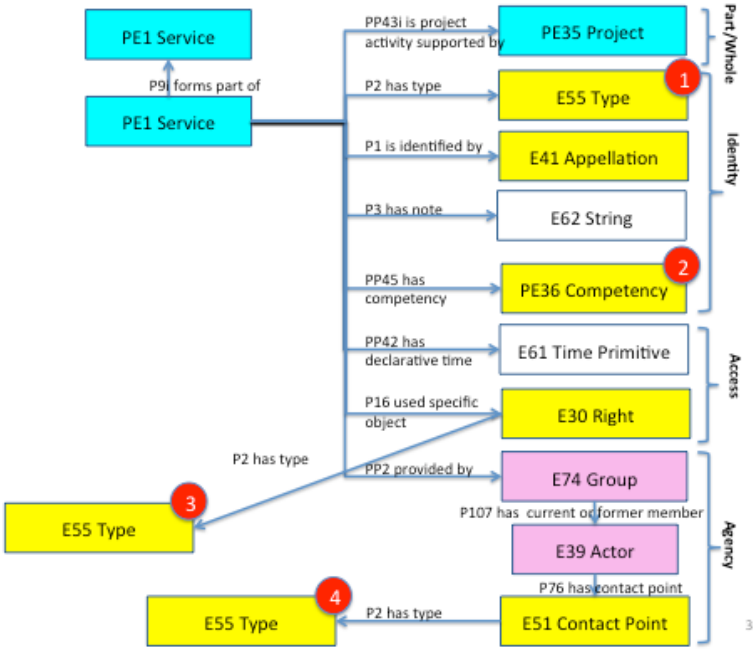

Figure 14: PE1 Service Minimal Metadata Application Profile Schema

The PE1 Service minimal metadata application profile makes reference to four fields which require standardisation according to common vocabularies. The following table summarises 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	Communication 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, licence-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.
Communicatio n Address	Y	String	The contact address for this contact person, any type.
Communicatio n 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 metadata set for PE17 Curated Data E-Service is as follows:

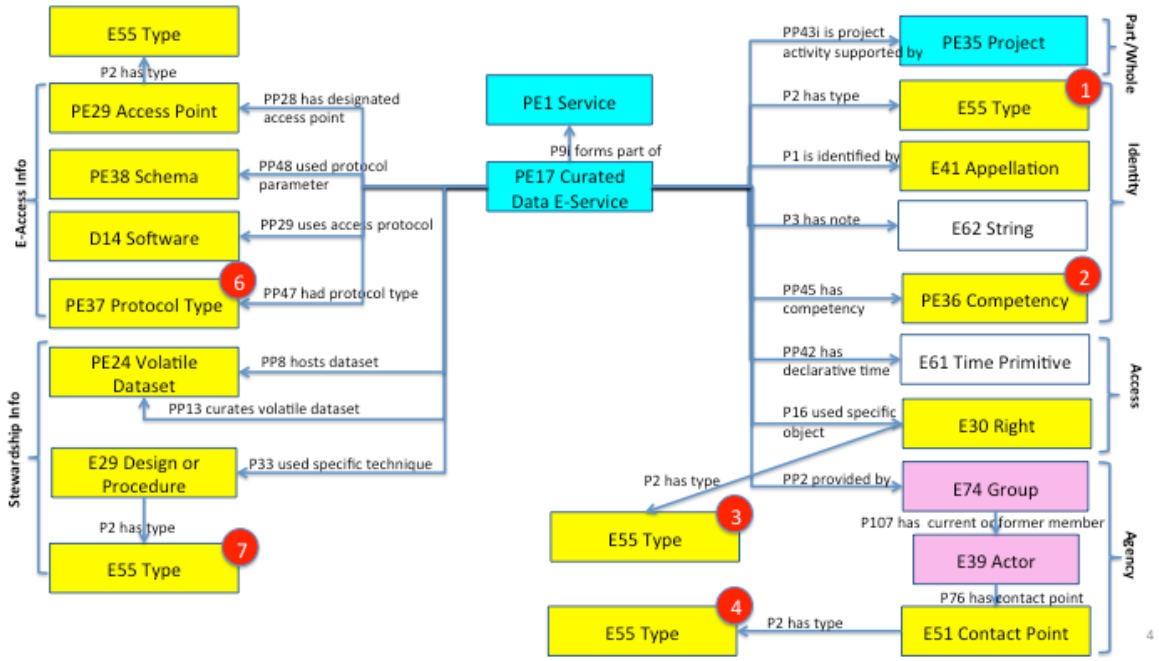

Figure 15: PE17 Curated Data E-Service Minimal Metadata Application Profile Schema

The PE17 Curated Data E-Service minimal metadata application profile makes reference to seven fields which require standardisation according to common vocabularies. The following table summarises 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, licence-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 metadata set for PE16 Curated Software E-Service is as follows:

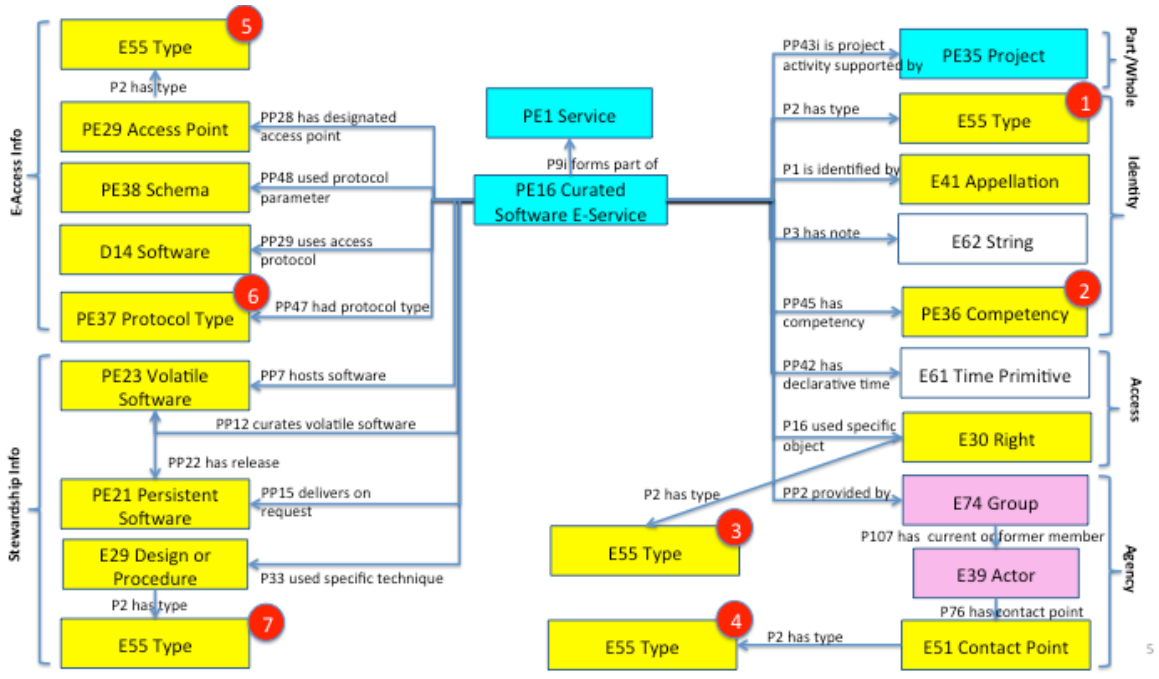

Figure 16: PE16 Curated Software E-Service Minimal Metadata Application Profile Schema

The PE16 Curated Software E-Service minimal metadata application profile makes reference to seven fields which require standardisation according to common vocabularies. The following table summarises 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 data stream 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 metadata set for PE22 Persistent Dataset is as follows:

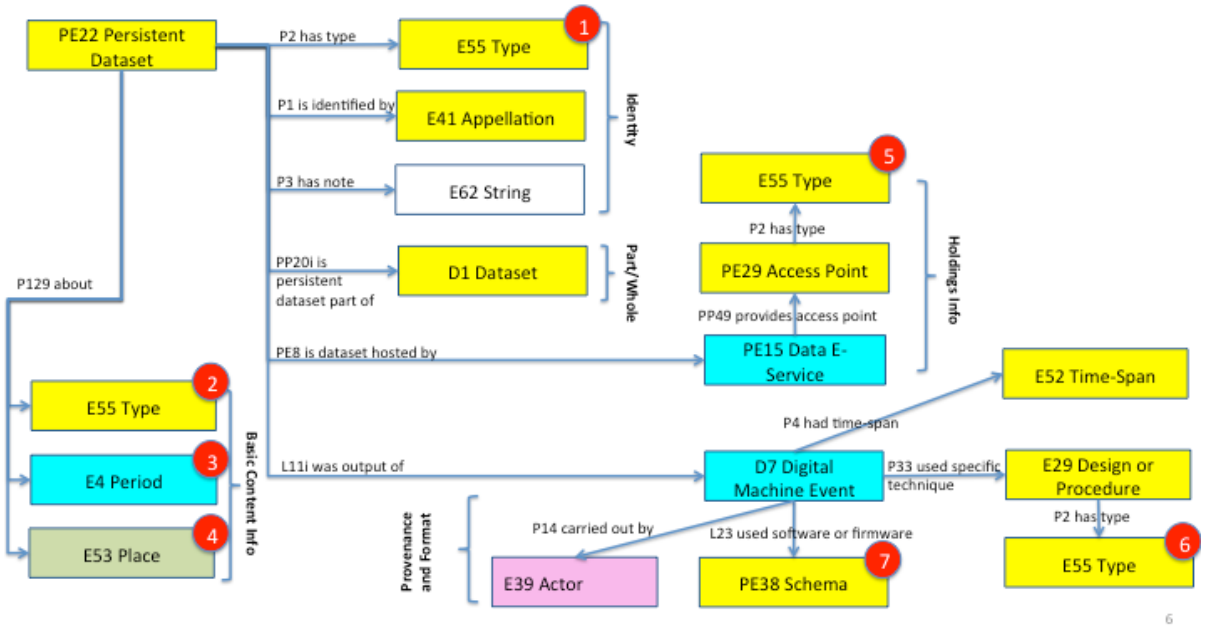

Figure 17: PE22 Persistent Dataset Minimal Metadata Application Profile Schema

The PE22 Persistent Dataset minimal metadata application profile makes reference to seven fields which require standardisation according to common vocabularies. The following table summarises 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 metadata set for PE24 Volatile Dataset is as follows:

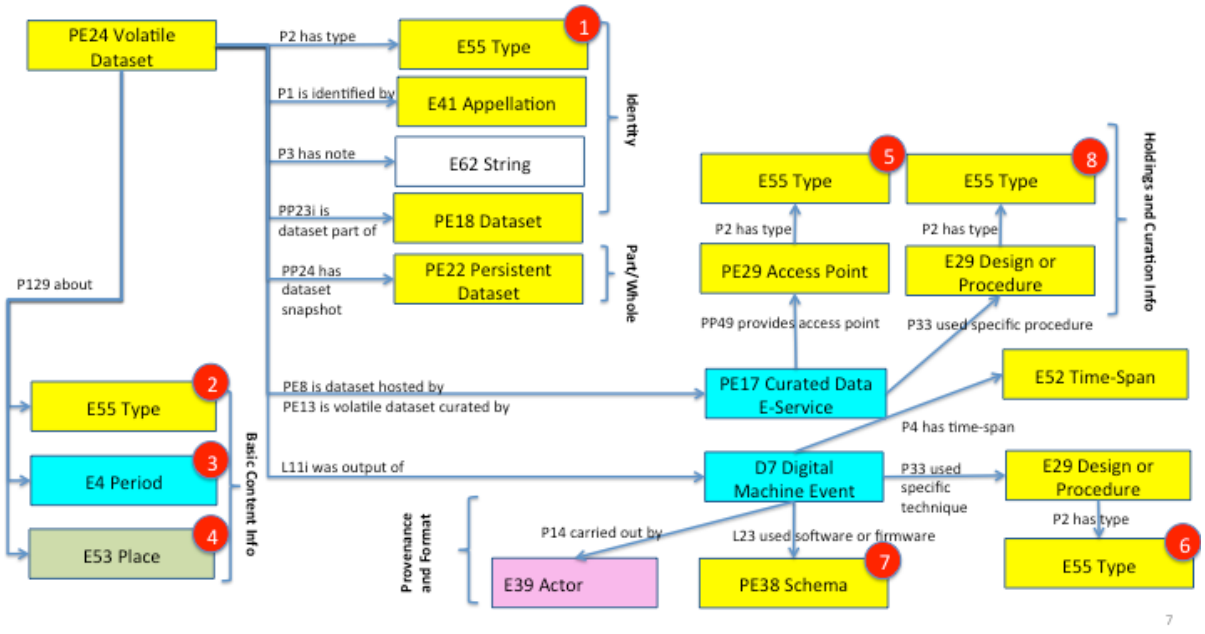

Figure 18: PE24 Volatile Dataset Minimal Metadata Application Profile Schema

The PE24 Volatile Dataset minimal metadata application profile makes reference to eight fields which require standardisation according to common vocabularies. The following table summarises 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 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.
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 metadata set for PE21 Persistent Software is as follows:

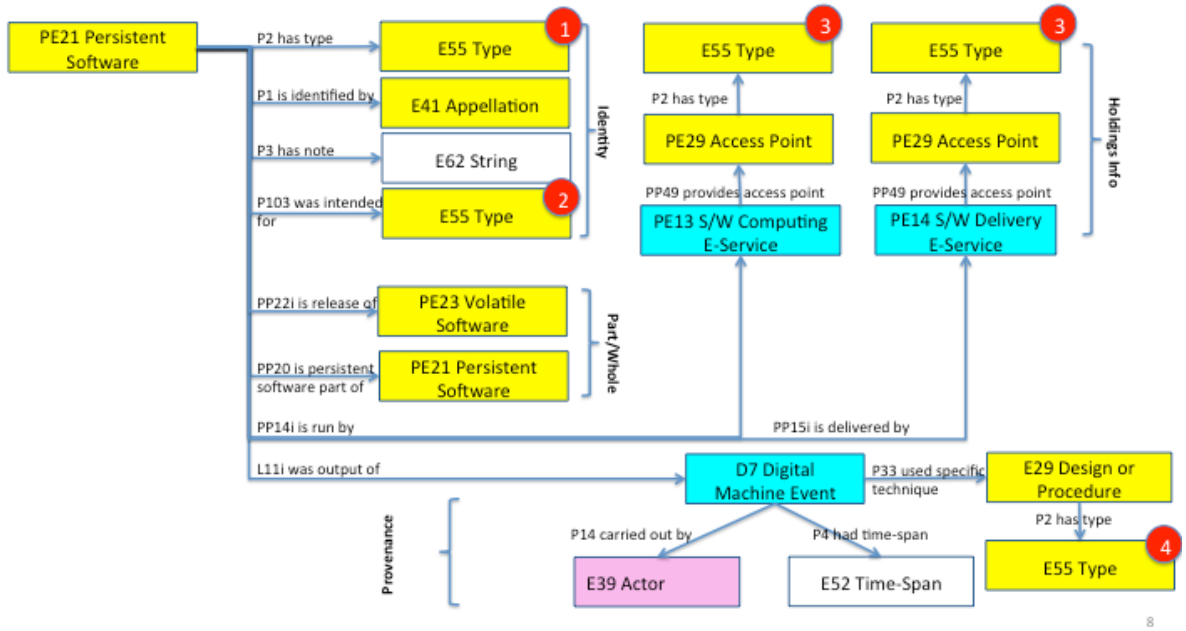

Figure 19: PE21 Persistent Software Minimal Metadata Application Profile Schema

The PE21 Persistent Software minimal metadata application profile makes reference to four fields which require standardisation according to common vocabularies. The following table summarises 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 metadata set for PE23 Volatile Software is as follows:

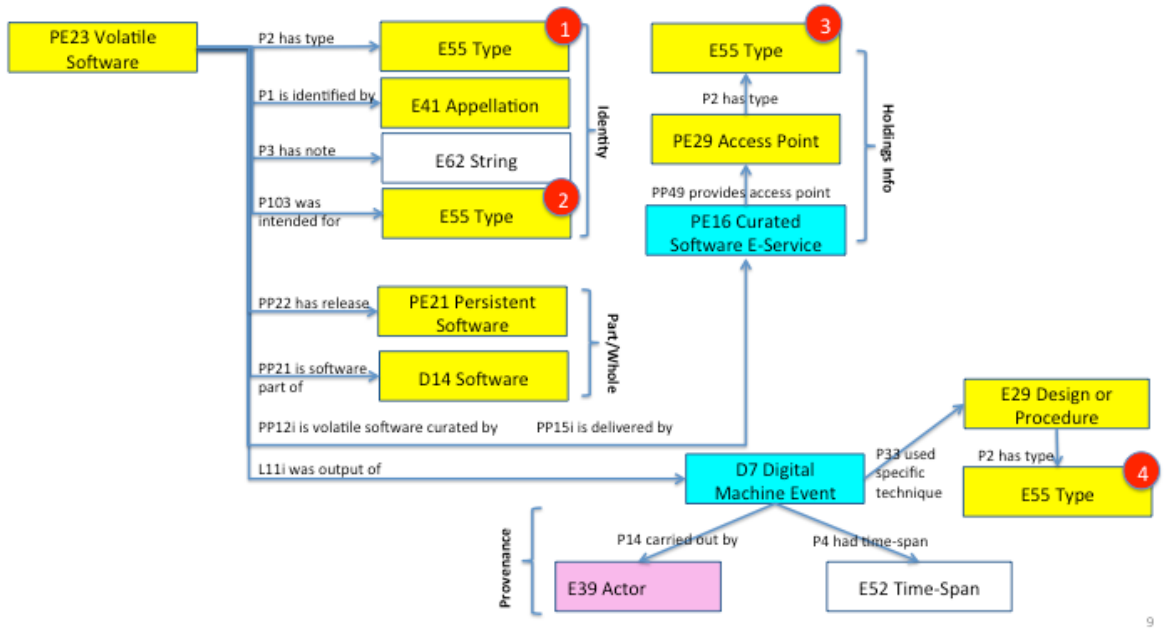

Figure 20: PE23 Volatile Software Minimal Metadata Application Profile Schema

The PE23 Volatile Software minimal metadata application profile makes reference to four fields which require standardisation according to common vocabularies. The following table summarises 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 metadata set for PE34 Team is as follows:

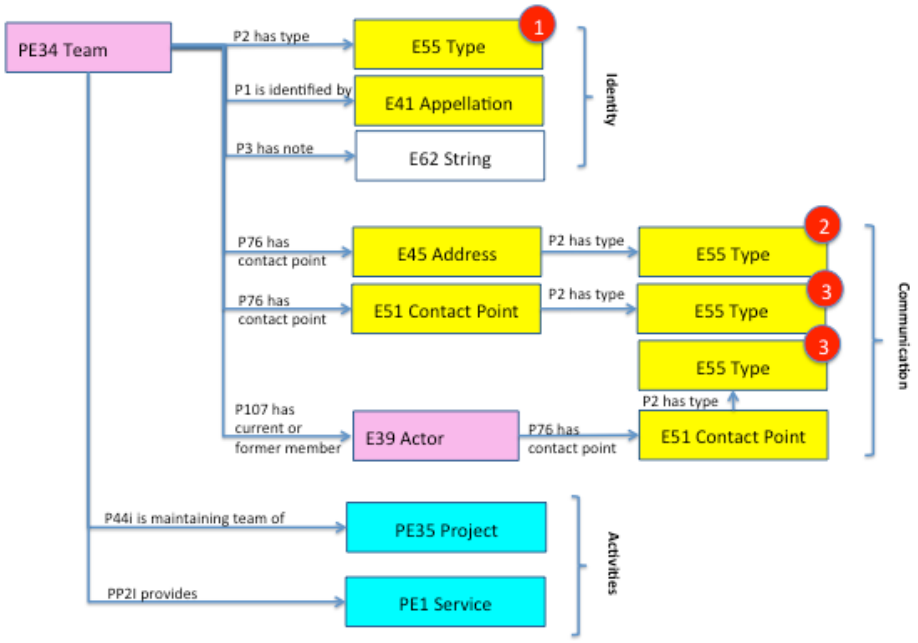

Figure 21: PE34 Team Minimal Metadata Application Profile Schema

The PE34 Team minimal metadata application profile makes reference to three fields which require standardisation according to common vocabularies. The following table summarises 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 metadata set for E21 Person is as follows:

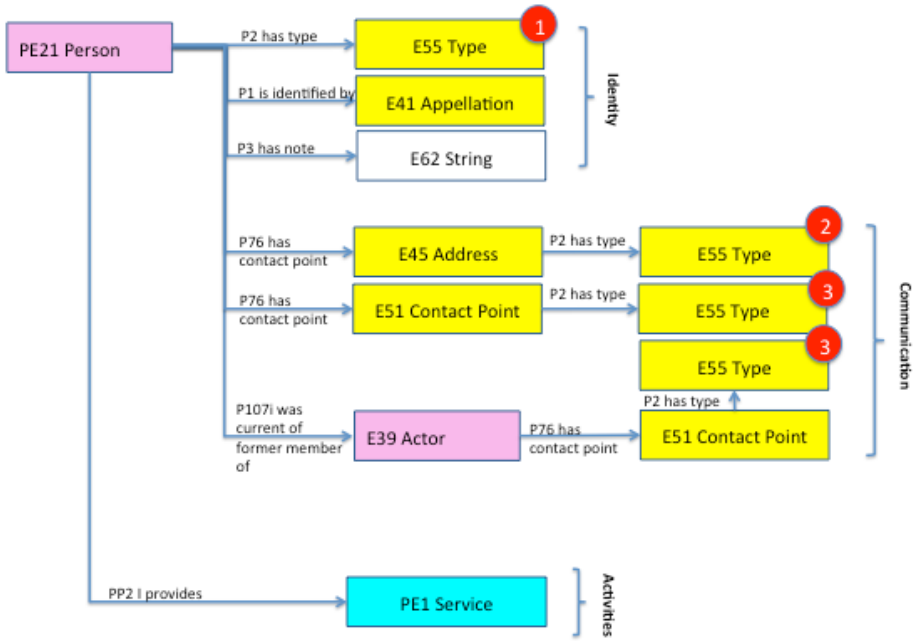

Figure 22: E21 Person Minimal Metadata Application Profile Schema

The E21 Person minimal metadata application profile makes reference to three fields which require standardisation according to common vocabularies. The following table summarises 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,[16] the required vocabularies to properly standardised 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 standardised 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)[17], the Open Metadata Registry[18] the Linked Open Vocabularies (LOV)[19], and the CERIF data model[20] 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 standardisation 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.

[16] http://www.d-net.research-infrastructures.eu/node/22
[17] https://bartoc.org/
[18] http://metadataregistry.org/
[19] http://lov.okfn.org/dataset/lov/
[20] 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, digitisation 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
Digitisation Process Types	Classify types of digitising 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	PARTHENOS Publishing Roles List	Broad concept combined with a more constricted selection of used values in the data makes a custom

		Scholarly Contributions and Roles Ontology CERIF Person Organization Roles		vocabulary the most feasible
Preservation Activity Type	Classify types of preservation activities	PAV	Dropped	Not present in the data / recorded by any RI
DateTime Norms	Standardisation 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 - E-Service				
Vocab Needed	**Function**	**Standards Considered**	**Decision**	**Rationale**
Authorisation Policies	Classify types of authorisation policies	None	PARTHENOS Rights List	Not present in the data / recorded by any RI, but reasonable fit for the already required list
Contact Point Types	Classify types of points of contact	CERIF Electronic Address Type & Person Contact Details & Organisation	CERIF Electronic Address Type & Person Contact Details & Organisation	Best fit for present data values

		Contact Details International Contact Ontology NEPOMUK Contact Ontology Contact: Utility concepts for everyday life	Contact Details	
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 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 & Person Project Engagements	Dropped	Not present in the data / recorded by any RI
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**
Authorisation Policies	Classify types of authorisation 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 Contact Ontology Contact: Utility concepts for everyday life	CERIF Electronic Address Type & Person Contact Details & Organisation Contact Details	Best fit for present data values
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. After review of the values present concerning places, it was established that a normalisation was unsuitable for this field, as it covers actual instances of places, rather than types or concepts thereof. As those values were discovered to be highly heterogenous and often messy, first steps have been taken for exploring approaches of instance matching, which could be expanded upon in future works.

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	Dropped, candidate for possible instance matching	As the observed values described instances rather than types, a vocabulary normalisation was deemed unsuitable
Spatial Coordinates	Standardise spatial coordinate values	ISO 6709	Dropped	Ideally, the standards used by the RIs omit the need for further normalisation
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	Dropped/Delegated to BBT	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 ARIADNE Data Collection PeriodO subset Historic England Periods Authority File iDAI.chronontology	ARIADNE Data Collection PeriodO subset	Best fit for present data values and quite exhaustive, while not as heterogenous and redundant as the full PeriodO collection

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	Standardised 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	Iana Media Types	Very exhaustive, highly curated list
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 Types	Classify types of dimensions	Units of Measurement Ontology	Dropped	Not present in 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 the former, the task of normalisation was discovered to be nonapplicable, as discussed in Section 4.3.1. For the latter, a good solution 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	Dropped, candidate for possible instance matching	As the observed values described instances rather than types, a vocabulary normalisation was deemed unsuitable
Spatial Coordinates	Standardise spatial coordinate values	ISO 6709	Dropped	Ideally, the standards used by the RIs omit the need for further normalisation
Legal 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 2.3 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 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 results of this activity can be seen in the re-expressed BBT now with the PARTHENOS Entities vocabularies integrated within the general framework. In what follows, BBT facets are marked in **boldface**, new BBT hierarchies that have been proposed to integrate the PARTHENOS Entities vocabularies are marked as [BBT NEW], whereas any additional structure imposed on the PARTHENOS Entities Vocabularies that serves as a hook by which the relevant terms will be connected to the BBT, is marked as [*PARTHENOS hierarchy top-term*]. Finally, the

label “Intermediate Generalisation” refers to cases where the vocabulary to be aligned to BBT was a flat list, with no declared top-term, and the intermediate node was introduced to support its hierarchical integration.

activities
— disciplines
— — ***PARTHENOS Disciplines***
— human interactions
— intentional destructions
— functions
— service competency [BBT NEW]
— data management activities [BBT NEW]

---

natural processes
— natural geneses
— natural destructions

---

materials

---

material things
— mobile objects
— built environment
— — ***PARTHENOS Place Types***
— physical features
— — ***PARTHENOS Place Types***
— structural parts of material things

---

types of epochs

---

conceptual objects
— symbolic objects
— — identifiers [PARTHENOS hierarchy top-term]
— — — contact point types [Intermediate Generalisation]
— — — — ***CERIF –Electronic Address Type, Person Contact Details and Organization Contact Details***
— — dataset types [PARTHENOS hierarchy top-term], [Intermediate Generalisation]
— — — ***CERIF Output Types***
— propositional objects
— norms [BBT NEW]
— — Intellectual Property Rights [PARTHENOS hierarchy top-term]
— — — Copyrights [Intermediate Generalisation]
— — — — ***PARTHENOS Rights List***
— — — Industrial Property Rights [Intermediate Generalisation]
— methods
— — ***PARTHENOS Data Policy Functions***
— — encoding types [PARTHENOS hierarchy top-term], [Intermediate Generalisation]

— — — ***File Format and Overview Information***
— languages [BBT NEW]
— — natural languages [PARTHENOS hierarchy top-term], [Intermediate Generalisation]
— — — ***Languages Name Authority List***
— — formal languages [PARTHENOS hierarchy top-term]
— — — programming languages [Intermediate Generalisation]
— — — — ***Wikipedia Programming Language List***
— concepts
groups and collectivities
— — ***PARTHENOS Audience***
roles
— offices
— roles of interpersonal relations
— — publishing roles [PARTHENOS hierarchy top-term], [Intermediate Generalisation]
— — — ***PARTHENOS Publishing Roles***
geopolitical units
— ***PARTHENOS Place Types***
geometric extents [BBT NEW]
— points [BBT NEW]
— linear extents [BBT NEW]
— surface areas [BBT NEW]
— 3D volumes [BBT NEW]

Table 32: Summary of BBT Organization after Integration of PARTHENOS Reference Resource Datasets

In total we integrated eleven vocabularies discovered in the effort to find robust and sufficiently wide but accurate control terms. The following section gives an outline of the results of the integration divided by facet and by function. “By function” refers to whether the hierarchies created are treated as BBT new terms or as top-terms of PARTHENOS-particular hierarchies, aligned to BBT.

Assuming a bottom-up approach, we will first be presenting the PARTHENOS-particular hierarchies aligned to BBT by facet, before examining the new BBT terms that they motivated. The scope notes for the BBT new terms and for the PARTHESNOS particular hierarchies can be found in APPENDIX III.

5.1. Activities Vocabularies

The only relevant vocabulary that has been connected under BBT-activities is ***PARTHENOS Disciplines***, which lists the types of professional and scientific domains involved in the PARTHENOS project infrastructure.
Aside that, it has been proposed that two new hierarchies be introduced to BBT, namely (i) service competency and (ii) data management activities. Despite the fact that we have found no formalised terminology to integrate under the relevant nodes, one can envisage a situation where relevant vocabularies will be recovered/generated. Hence, we have decided to maintain these two distinct activity types for the moment.

5.2. Conceptual Objects Vocabularies

The symbolic objects facet is designed to capture types of immaterial but identifiable mental products.

Among the PARTHENOS vocabularies that were integrated to BBT under conceptual objects, two fall within the scope of symbolic objects; the hierarchy Identifiers encompasses all sorts of symbols that aim to univocally name an item through a certain elaborated identification system. It is further specialised by the subhierarchy “Contact Point Types”, i.e. identifiers used for all kinds of addresses. “Dataset types” were also connected to BBT under symbolic objects. What motivated this decision is that the purpose of the dataset necessarily reflects on its form. Hence, types of datasets are to be classified according to their forms, rather than their contents (or the combination thereof).

Finally, the ***vocabularies PARTHENOS Data Policy Functions*** and ***File Format and Overview Information*** –the latter forming the hierarchy Encoding Types –were connected under BBT-methods.

The rest of the PARTHENOS Entities vocabularies aligned under conceptual objects called for the declaration of new terms in BBT.

The hierarchies (a) Formal languages –and its subhierarchy Programming languages –and (b) Natural languages alike describe systems of communication comprising a finite set of elements and a set of recursive rules to combine them into a potentially infinite array of

discrete expressions. They form specialisations of conceptual objects in the sense that (i) they are products of human activity that may –but need not –be supported by the use of technical devices, (ii) their essence remains the same regardless of the carrier, and (iii) they have the ability to exist on more than one particular carrier at the same time. However, they do not fit under any of the existing BBT hierarchies particular to conceptual objects –whence the need to declare a new BBT branch within the conceptual objects facet to deal with systems of communication (as opposed to their products in symbols, propositions and information objects), namely "Languages".

The ***PARTHENOS Rights List*** vocabulary revolved around copyrights and licences. Integrating it to BBT required declaring a number of additional hierarchies within the PARTHENOS Entities Vocabularies, namely Intellectual Property Rights and its children, Copyrights and Industrial Property Rights,[21] all defined by is*A* relations. The BBT hierarchy Norms [BBT NEW], which covers all sorts of systems of regulation, can adequately accommodate the types of copyrights and licences recovered from the data, whereas the intermediate nodes between Norms and the ***PARTHENOS Rights List*** ensure that the classification of copyright types are not considered artificial/ad hoc.

5.3. Roles Vocabularies

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

5.4. Vocabularies split among different BBT facets.

Of the vocabularies that were integrated to BBT, two required to be split across multiple BBT facets and/or hierarchies. The relevant vocabularies were the ***PARTHENOS Subjects List*** and the ***PARTHENOS Place Types List.***

The ***PARTHENOS Subjects List*** conveys information regarding the research objects that are deemed relevant for the Research Infrastructures participating in the PARTHENOS

[21] Industrial Property Rights have only been added to the classification for the sake of completeness; in fact, the ***PARTHENOS Rights List*** makes no reference to inventions, patents, trademarks and/or industrial designs.

project. These research objects express propositional objects in essence –subjects convey an aboutness topic, which needs be expressed by a proposition. However, integrating the relevant vocabulary under the respective BBT term was not an option at this stage of the project; not only such an approach would generate a parallel hierarchy within BBT propositional objects, but it would also create ambiguity –for instance, is criminology considered an interdisciplinary field or a subject that one can talk about? Hence, unless we have explicit scope notes on the designated subjects, we cannot really proceed with integrating the said vocabularies.

The PARTHENOS Subjects hierarchy that was actually integrated in BBT was the outcome of the mappings undertaken in the context of the project –whereby propositional objects were assigned to their corresponding CRM entities: *E89 Propositional Object –P129 is about –E1 CRM Entity*. To avoid the creation of a parallel hierarchy, the resulting subjects were split across facets, as indicated in the table below:

Common Policies	conceptual objects -> norms [BBT new]
Communication	activities -> human interactions
Research agenda, foresight studies	conceptual objects ->methods
Standards	conceptual objects -> norms [BBT new]
Training	activities -> human interactions

The ***PARTHENOS Place Types List*** was compiled based on the TGN place types –i.e. controlled terms –describing the TGN entities (e.g., nation, empire, caliphate, inhabited place, village, archaeological site, cave dwelling, peak).

The place type terms are linked to AAT and their meaning is defined as a spatial projection of the spatiotemporal extents of observable and/or measurable real world phenomena.
The TNG place type list was extracted directly from the SPARQL endpoint of the Getty Research Institute[22] via a query. The resulting j*son* file, containing terms and URIs, was then parsed into an XML schema for easy import via a Python script.

These place types terms were integrated to BBT without problems, where they split into two separate facets –geopolitical units and material things. Depending on the inherent

[22] http://vocab.getty.edu/

properties of the entities they denote, the terms falling within the material things facet were classified as built environment or physical features.

5.4.1. Geometric Extents

Integrating the Place Types vocabulary to the BBT motivated the declaration of a new facet. This particular branch aims at defining places based on types of geometric expressions that may be used to represent them. Geometric extents can be coordinated with terms listed as built environment, physical features or geopolitical units, to refer to their shapes and/or representations on a given reference space, aside their nature.

The scope notes of the relevant terms can be found in Appendix III of this document.

5.5. Non-Categorical Reference Resources

Worthy of note are three standardised sources that we did not integrate to the BBT, namely the ARIADNE system for standardising periods, TGN for standardising place references, 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 expressly designed for organisation information and the categorical level.

6. Conclusion

The PARTHENOS project attempts a broad, cross-disciplinary aggregation of basic data regarding information management at the RI level. The aggregated data is presented in the Joint Resource Registry. Aside from schema level integration, integration at the level of data values is a basic requirement in order to make the aggregate data tractable to query and research. This aggregation effort provided an ideal environment for designing, testing and implementing a generic workflow for reference data integration, adopting the methodology of the BBT as a conceptual check and long term sustainability tool for this work.

The result of research on this topic was the design of a general workflow for reference data integration taking into account a data integration project as the context for creating a sustainable and compatible set of reference resources. The general workflow suggests six documented steps and management points: identification, discovery, creation, registration, integration and implementation. These are seen as essential parts of a complete and scoped cycle of data integration. The initial steps cover the documentation of needs for reference resources and the steps for finding, creating and registering these. The BBT methodology is applied in step 5, to integrate the reference resources amongst themselves and into an overall compatible model. All of this is tooled towards application in a data aggregation scenario where the resultant vocabularies can be used for data normalisation on controlled fields in the aggregate data sources.

The general plan was implemented in a specific workflow, adopting the best available tools to carry out the tasks envisioned. The identification of needs, discovery and creation processes are documented in the present report in sections 3.2 and 4. An analysis of the PARTHENOS Entities Data Model was undertaken to study the required fields for implemented categorical standards. Section 5 describes the process the intellectual process that was undertaken to align the selected vocabularies to the BBT. The overall implementation workflow is described in section 2.5.

The experience of implementing the above general workflow in PARTHENOS revealed the sparsity of standard reference resources for use at the level of information management for RIs. There is a lack of investment in the creation of reference resources which would

support standardisation of data values in such data structures. It is a positive outcome of this process, that our research was able to uncover sources and derive lists from data values within the aggregated datasets. These have been published in the Joint Resource Registry. The adoption of the BBTalk tool to integrate the selected resources into a broader framework was generally successful. As expected, the introduction of vocabulary from a new domain necessitated an expansion of the base terms of BBT to accommodate new areas of research. This allowed the testing of the evaluation and curation methodologies developed for controlling the BBT in order to assess the suitability and correctness of new terms and term extensions into the BBT. The curation process is on-going, the results of which will be published in the ACDH Vocabularies.

Further research would need to investigate those areas where RIs used competing and equally correct reference resources for the same field or descriptor, for example the concept of ‘subjects’ to determine to what extent a deeper alignment beyond agreement on a top term could be carried out.

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/digitisationguidelines/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: Standardised Vocabularies

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

Appendix III: BBT NEW & PARTHENOS Hierarchies Top-Terms.

1. BBT NEW terms

a. Service competency

This term classifies processes or actions that a service is designed to carry out and should deliver/accomplish upon request. The concept serves to classify services not according to what they are, but according to the type of outcome they offer to their respective beneficiaries –the latter serves to determine the identity of the service competency.

b. Data management activities

This term classifies the kinds of activities undertaken at each of the different stages in the lifecycle of data, from their creation to re-use, and the activities undertaken to make data usable and available for the long term.

Examples of data management activities regarding the creation of data, relate to planning and resolving issues regarding the ownership of the data to be collected, designing the appropriate methods for its collection and its enrichment with metadata, decisions as to its preservation, as well as decisions regarding the circumstances under which access will be granted to the data collection. The activities involved in this stage have to do with the planning and the collection per se, as well as the creation of the metadata to describe the collection. Examples of data management activities regarding data processing, involve data entry, digitization, check, validation, cleaning etc. Examples of data management activities regarding data preservation include data migration to best format and suitable medium, back-up and storage. Finally, examples of data management activities regarding the dissemination of the data take place, preceded by establishing a controlled access to the data. At this stage, the data can be reused for follow-up studies.

c. Norms

This term classifies official standards, usually presented in a formal document written by a recognized organization (such as ISO, ANSI, AFNOR, DIN, etc.) that establishes uniform criteria, rules, methods, processes and practices to be used as references for an activity, a subject, a result.

d. Languages

This term classifies types of communications systems comprising of a finite set of elements and a set of recursive rules to combine them into a potentially infinite array of discrete expressions.

e. Geometric extents (facet)

This facet comprises kinds of designations and definitions of spatial extents based on either geometric expressions or spatial properties of observable features -like mountains, lakes, buildings, cities, etc. -and social constructs -referring to the spatial extent of territories that fall within the jurisdiction of some geopolitical or other administrative unit.

NOTE: The terms and hierarchies of this facet can be coordinated with the suitable type of phenomenal place, in the sense of CRMgeo, classified accordingly under Physical Features, Built Environment or Geopolitical Units.

f. Geometric extents (top-term)

This term classifies kinds of designations and definitions of spatial extents based on either geometric expressions or spatial properties of observable features -like mountains, lakes, buildings, cities, etc. -and social constructs -referring to the spatial extent of territories that fall within the jurisdiction of some geopolitical or other administrative unit.

NOTE: The terms listed as Geometric extents can be coordinated with the suitable type of phenomenal place, in the sense of CRMgeo, classified accordingly under Physical Features, Built Environment or Geopolitical Units.

g. Points

This term classifies zero-dimensional geometric primitives, representing the position of the centroid of a particular feature, on a given surface –irrespective of its actual spatial extent –depending on the scale of the representation (the smaller the scale, the more likely it is for a feature to be thus represented), convenience and the type of feature the points stand, for or some position on linear structure, such as a "border triangle".

NOTE: The terms listed as points can be coordinated with the suitable type of phenomenal place -in the sense of CRMgeo -classified under the hierarchies of Physical Features, Built Environment or Geopolitical Units.

h. Linear extents

This term classifies one-dimensional shapes on a surface that are either straight or curved and can be defined by a connected series of unique x,y coordinate pairs/points forming a continuous path[1]. The said points are all contained in it. Linear extents may be used to approximate the 2-dimensional extent of features much longer than wide, such as roads, rivers, contours, footpaths, flight paths and so on, or to describe declarations of borders.

NOTE: The kind of Physical feature or Built environment providing the geometric extent -i.e. a river, a coastline, a road or a bridge -can be specified by coordinating this term with the suitable feature type, such as “surface areas of Physical features/ Built environments”.

i. Surface areas

The term classifies expressions specifying the position and extent of a two dimensional feature, figure or shape. Such expressions may be numerically determined closed paths, such as a connected sequence of x,y coordinate pairs/points forming a polygon, or the geometric projection of a stationary two-dimensional feature on a surface, in particular that of earth. They may even be projections of temporally limited dynamic phenomena such as a flood area or a battlefield. Surface areas can be seen as contiguous projections onto some reference space. Examples of such areas are enclosed spaces like that of islands, cities, forests, lakes, country or real estate boundaries and so on.

NOTE: The kind of Physical feature, Built environment or Geopolitical unit providing the geometric extent -i.e. a lake, a stadium, a prefecture -can be specified by coordinating this term with the suitable feature type, such as “surface areas of Physical features/ Built environments/ Geopolitical units”.

j. 3D-volumes

This term characterizes physical features or material objects extending in three dimensions/ defined along three axes of a Euclidean space . They can –but need not –be solid and can be reduced to three-dimensional polyhedra.

NOTE: The kind of Physical feature, Built environment or Geopolitical unit providing the geometric extent -i.e. the bed of a lake filled with water, the volume occupied by a building, or the Exclusive Economic Zone of a sovereign state represented in terms of a 3D volume -can be specified by coordinating this term with the suitable feature type, such as “surface areas of Physical features/ Built environments/ Geopolitical units”.

2. PARTHENOS Hierarchies Top-terms

a. Identifiers

This term classifies strings or codes assigned to items/objects in order to identify them uniquely and permanently within the context of one or more organizations. Such codes are often known as inventory numbers, registration codes, etc. and are typically composed of alphanumeric sequences.

b. Contact point types

This term classifies identifiers employed or understood by communication-services by which services and/or service-providers can be accessed. These include addresses of all types, such as email addresses, telephone numbers, post office boxes, fax numbers, URLs etc.

c. Dataset types

This term classifies kinds of identifiable information objects that can be represented as sets of bit sequences and whose content contains propositions about some world.

d. Intellectual Property Rights

This term classifies legal privileges concerning material and immaterial things or their derivatives. They are like any other property right by allowing creators or owners of patents, trademarks or copyrighted works to benefit from their own work or investment in a creation.

As an example of Intellectual Property rights, consider copyrights, patents and trademarks protection.

e. Copyrights

This term classifies property rights ascribed to creators of intellectual creations. The domain of copyright protection is original works of authorship fixed in any tangible medium of expression. Works that may be copyrighted include literary, musical, artistic, photographic, architectural, and cinematographic works; maps; and computer software. For something to be protected it must be “original”—the work must be the author’s own production; it cannot be the result of copying. A further requirement that limits the domain of what can be copyrighted is that the expression must be “non-utilitarian” or “non-functional” in nature.

f. Industrial Property Rights

This term primarily classifies property rights related to inventions and industrial designs, i.e. new solution to technical problems and aesthetic creation determining the appearance of industrial products, respectively. In addition, it also covers trademarks, service marks, commercial names and designations, including indications of source and appellations of origin and protection against unfair competition.

g. Encoding types

This term classifies the kinds of algorithmic processes by which a file is generated, and indicates the means by which it can be read or displayed and operated on.

h. Natural languages

This term classifies the kinds of communication systems particular to humans and (possibly) no other species on the planet. Like all kinds of communication systems that can be referred to as ‘Language’, natural language also possesses a finite set of elements (sounds or gestures) and a recursively defined grammar (i.e. set of rules and principles) specifying the properties of its expression. What sets natural language apart from other communication systems is that it is passively, effortlessly acquired during early age, by mere exposure to linguistic input.

At the same time, natural language is a phenomenon deeply entrenched in human culture, that –aside communication ‘pure’ –is associated with other functions as well, like establishing relations, building identities (both individual and social), entertainment etc.

Examples of Natural Languages are English, Modern Greek, Turkish, Arabic, Chinese and their dialects (especially if the “building identities” part is to be considered) like BEV/AMEV (Black English/American English Vernacular), Cappadocian “Greek” (heavily Turkicized after the Ottoman conquest in the 111th century) etc.

i. Formal languages

This term classifies types of languages consisting of recursively defined collection of strings on a fixed alphabet (also referred to as a 'vocabulary'), by means of a number of explicit rules and constraints (also referred to as a 'syntax') that state which expressions (or 'words') combine with one-another into well-formed expressions, observing compositionality. Formal languages are designed by people for a clear, particular purpose. [LTF GAMUT, Partee et al. 1993, Stanford Encyclopedia of Philosophy -classical logic] Examples of formal languages are the language of Set theory, the language of FOPL, the language of ordinary arithmetic and others.

j. Programming languages

This term classifies kinds of formal languages comprising sets of instructions used to implement an algorithm or sets of statements that express facts and rules about some problem domain, which produce some valid output when executed on a computer.

Examples of programming languages are C, C++, Java, Perl, Python, R, etc.

k. Publishing roles

This term classifies the roles undertaken in the context of making the outcome of creative or academic work publicly available and they largely correspond to the different kinds of activities involved in the publishing process/business, namely (i) preproduction, (ii) production and (iii) dissemination/distribution.

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.

Manghi, P., Artini, M, Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela, L., Castelli, D. & Pagano, P. (2014). *"The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures"*, Program, Vol. 48 Issue: 4, pp.322-354, doi:10.1108/PROG-08-2013-0045

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.