

# COLLABORATIVE EUROPEAN DIGITAL ARCHIVE INFRASTRUCTURE

**Project Acronym:** CENDARI
**Project Grant No.:** 284432
**Theme:** FP7-INFRASTRUCTURES-2011-1
**Project Start Date:** 01 February 2012
**Project End Date:** 31 January 2016

| | |
|---|---|
| **Deliverable No. :** | D6.3 |
| **Title of Deliverable:** | Guidelines for Ontology Building |
| **Date of Posting to Basecamp/Confluence for Partner Review:** | April 2014 (Version 0.3) |
| **Date of Finalised Deliverable:** | July 2014 (Version 1.0) |
| **Revision No.:** | 1 |
| **WP No.:** | 6 |
| **Lead Beneficiary:** | King's College London (KCL) |
| **Author (Name and email address):** | Deliverable developed by the CENDARI WP6 team |
| **Dissemination Level:** | PU = public |
| **Nature of Deliverable:** | O = other |
| **Abstract:** | This document provides guidelines for ontology building within CENDARI. It is based on an assessment of existing methodologies and approaches to ontology development, the creation of ontologies within similar projects, the ontological requirements of the CENDARI users, and the technical limitations of the tools that are available. |


**Table of Contents**

# 1. Introduction

This document provides guidelines for ontology building within CENDARI. It is based on an assessment of existing methodologies and approaches to ontology development, the creation of ontologies within similar projects, the ontological requirements of the CENDARI users, and the technical limitations of the tools that are available.

The document consists of three parts in addition to the final conclusions:

- The definition of an ontology.
- Existing methodologies for ontology development.
- The application of the CENDARI approach to ontology building.

It is important to be explicit about what is meant by an ontology within the context of CENDARI before appropriate guidelines can be created for the building of the CENDARI ontologies. The difference in the meaning ascribed to the term by different communities (e.g., information scientists and computer scientists) has implications for both the types of ontology that are developed and the way that they are developed.

There have been a number of methodologies that have been developed for building ontologies, each of which has focused on the development of different stages of the ontology building process. Based on the existing methodologies an eleven step meta-model has been created for the development of the CENDARI ontologies.

The eleven stage CENDARI meta-model provides the framework for the final (and largest) part of the research guidelines. Each of the eleven stages is discussed within the context of building domain ontologies for CENDARI and the requirements of the different stakeholders. The different stakeholders each place constraints on the ontologies, impacting both how they are created and how they represent knowledge.

# 2. Defining an ontology

Ontologies form the backbone of CENDARI, bringing disparate resources together and enabling the discovery of new knowledge. Although the way that it brings these different resources together and the nature of the new knowledge that is discovered depends heavily on the definition of ontology that is used.

The definition of an ontology that is most often referenced within the information science community is that of Gruber (1993): “an explicit specification of a conceptualization”. Such a definition is extremely inclusive, and may be used as a catch all term for the various types of formalized vocabularies that exist to express concepts and the relationships between them, e.g., controlled vocabularies, taxonomies, and metadata schemas. In a more restrictive sense it may be used to refer only to those sets of conceptualizations with a richer variety of relationships or those that are expressed in accordance with a particular model that allows new relationships to be understood through inference. This is more similar to the W3C definition of an ontology in its overview of OWL 2 Web Ontology Language (2012):

"Ontologies are formalized vocabularies of terms, often covering a specific domain and shared by a community of users. They specify the definitions of terms by describing their relationships with other terms in the ontology."

The CENDARI infrastructure requires a wide range of ontologies, from small controlled vocabularies associated with one particular attribute of a metadata element to highly integrated domain ontologies incorporating the concepts and entities associated with the archives for medieval and First World War scholarship and the relationships between them. It is important to recognise that the development of the ontologies should not be considered wholly distinct from the development of a common metadata schema; not only are the metadata elements themselves clearly defined concepts, but so potentially are the contents of many of these elements (e.g., subject terms, geographic descriptions, and source type).

Within CENDARI there are many different sorts of "ontology":

- Metadata schemas for describing cultural heritage institutions, their collections, and items within the collections.
- Controlled vocabularies associated with the metadata records at the institutional, collection, or item level: lacunae causes; certainty of dates; role of person associated with collection; material type; impediments to use etc.
- Ontology of sources, broadening understanding of the types of sources that are available within archives, and their limitations.
- Domain ontologies incorporating concepts and entities associated with the two research communities (i.e., medieval and First World War scholars), and the relationships between them.

The focus of this document is on building the domain ontologies: representing the concepts and entities associated with each of the two research areas. These ontologies should include a far richer set of relationships than more traditional formalized vocabularies. Authority lists may have provided authorized versions of the names of people and places, and a thesaurus structure may have built a hierarchy of broader and narrower relationships between places, but an ontology could include information about the relationships between people and places (e.g., birthplace, deathplace) as well as relationships between people (e.g., married, father of, works for). Such relationships can turn a static vocabulary into to a database of knowledge to be queried for information that has never been brought together previously.

Confusion about the meaning of the term ontology is often compounded by the fact there are two parts to ontologies, the element sets and the instances adhering to the element sets, and 'ontology' may be used to refer to either or both. For example, an ontology of people may refer to the classes and attributes used to represent the concept of a 'person' and the relationships between them, or it may also include individuals adhering to those classes. Throughout this document clear distinctions are made between the **ontology element sets** and the **ontology instances**, with the phrase **domain ontology** used to refer to the combination of the ontology element set and the ontology instances for one of the two

research areas. A similar distinction has been made by the Library Linked Data Incubator Group (2011), where the distinction is between metadata elements sets and value vocabularies. Alternative terminology has been used within this document so there is less confusion when distinguishing between the ontology building described within this document and both the metadata records that already exist and the vocabularies that have not yet been incorporated as ontology instances.

## 3. Existing methodologies for domain ontology development

There have been a number of methodologies developed for ontology development, and as Noy and McGuinness (2001) have stated "there is no single correct ontology-design methodology". Table 1**Error! Reference source not found.** details the different steps associated with the development of a domain ontology according to four different methodologies. The steps associated with each of the methodologies may be read vertically, whilst similar stages in the methodologies have been aligned horizontally. As can be seen in Table 1**Error! Reference source not found.** there are a number of consistencies between the different methodologies, although they often emphasize different stages. Some methodologies focus more on establishing the scope of the ontologies, some on the building of the ontologies, and others on the evaluation and documentation of the ontologies.

Table 1 Comparison of different methodologies for domain ontology development

| Uschold and King (1995) | Gruninger and Fox's (1995) TOVE methodology | Fernández-López, Gómez-Pérez, & Juristo(1997) – METHONTOLGY | A simple knowledge-engineering methodology (Noy & McGuiness, 2001) |
|---|---|---|---|
| Identify purpose | Competency of the ontology – identify the form of questions an ontology must be able to answer. | Specification – expressing aspects such as the purpose, level of formality, and scope of the ontology. | Determine the domain and scope of the ontology. |
| | | | Consider reusing existing ontologies. |
| | | Knowledge acquisition | |
| Building the ontology<br><br>-Ontology capture | Define the terminology of the ontology – its objects, attributes, and relations. | Conceptualisation | Enumerate important terms in the ontology |
| Building the | Specify the definitions and constraints of the | | Define the classes and |

| | | | |
|---|---|---|---|
| ontology<br><br>-Ontology coding | terminology | | the class hierarchy |
| | | | Define the properties of classes-slots |
| | | | Define the facets of the slots |
| Building the ontology<br><br>-Integrating existing ontologies | | Integration | |
| | | Implementation | Create instances. |
| Evaluation | Test the competency of the ontology by proving completeness theories [logical tests] | Evaluation | |
| Documentation | | Documentation | |

From these four methodologies for developing a domain ontology, none of which provides more than seven steps to domain ontology creation, a preliminary eleven step meta-model was created for the development of the CENDARI domain ontologies. The eleven step meta-model combines elements from the four methodologies in Table 1 to produce a more detailed methodology than existed in any of the existing methodologies, whilst an additional element on the identification of appropriate software for domain ontology development was due to its importance in the development of large ontologies.

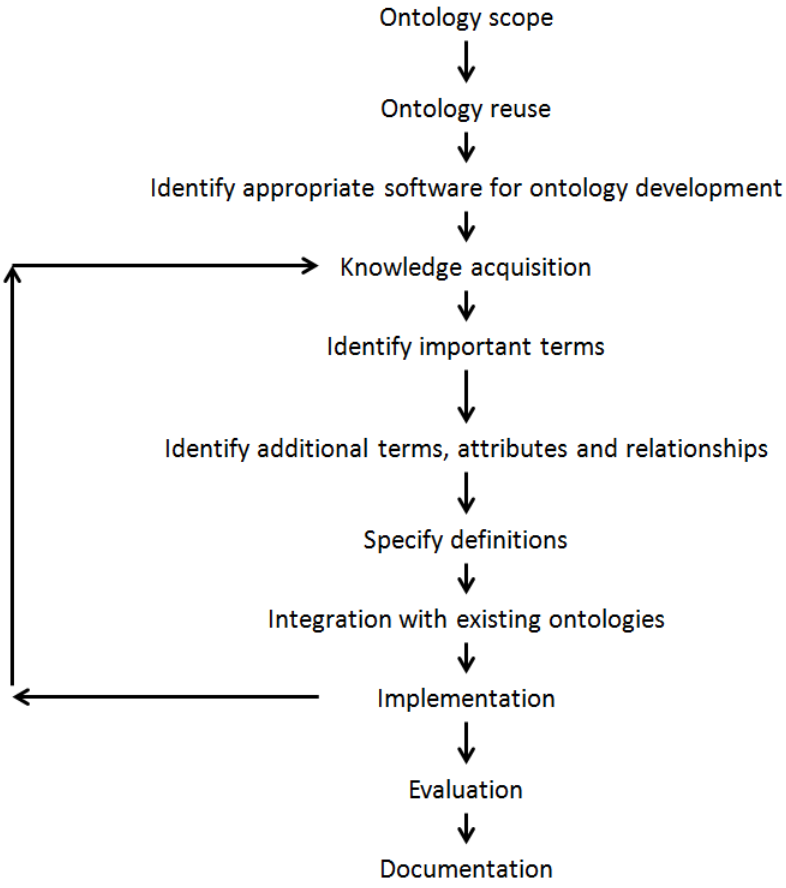


Figure 1 Eleven step meta-model to domain ontology development

Uschold and King (1995) point out that the 'building the ontology' sub-stages in their methodology (i.e., ontology capture, ontology coding, and integrating existing ontologies) do not have to be carried out in order, but may be reordered or merged into a single – iterative stage. The "iterative process" of ontology development is echoed by Noy and McGuiness (2001), and this notion is expressed in Figure 1 from knowledge acquisition through to implementation. This reflects the stages of iteration that would ideally be the focus of ontology develop. As is often the case, however, the iterative nature in fact transcends other stages as well. For example, the choice of technology for storing the ontologies has had to be changed to align with other CENDARI services, and the identification of ontologies for reuse is necessarily a gradual process due to the scope of the CENDARI domain ontologies and the many competing ontologies that cover similar areas.

# 4. Application of the CENDARI Meta-Model

## 4.1 Ontology Scope

The domain ontologies within CENDARI will not form a static database of knowledge separate from the rest of CENDARI, but will need to form a dynamic knowledgebase integrated into a wide range of other CENDARI services. The CENDARI domain ontologies will be:

- Integrated with institutional, collection, and item-level metadata records within the CENDARI environment, enabling researchers to navigate between related records.
- Integrated with researchers' notes in the virtual research environment and any associated annotation tools (e.g., Pundit), facilitating the organization of personal research notes as well as the discovery of other researcher's work (where permission is given).
- The foundation for additional CENDARI services, such as the development of Pineapple, a CENDARI service that is designed to answer questions rather than queries through the use of semantics and the formalization of concepts that may not have been explicit in the ontology.
- Enhanced as a Named Entity Recognition service identifies entities and relationships from additional metadata records.
- Enhanced as researchers using the CENDARI virtual research environment build new entities and relationships into the domain ontologies.
- A knowledgebase to be queried in its own right.

It is also important that the domain ontologies are available for use beyond the CENDARI project, and can be aligned with other data models.

Figure 2 provides a data flow diagram of the CENDARI domain ontologies with other CENDARI data stores and services. Following the Yourdon notation, the rectangles represent external entities, the circles are processes, the parallel lines are data stores, and the arrows show the direction of data flows.

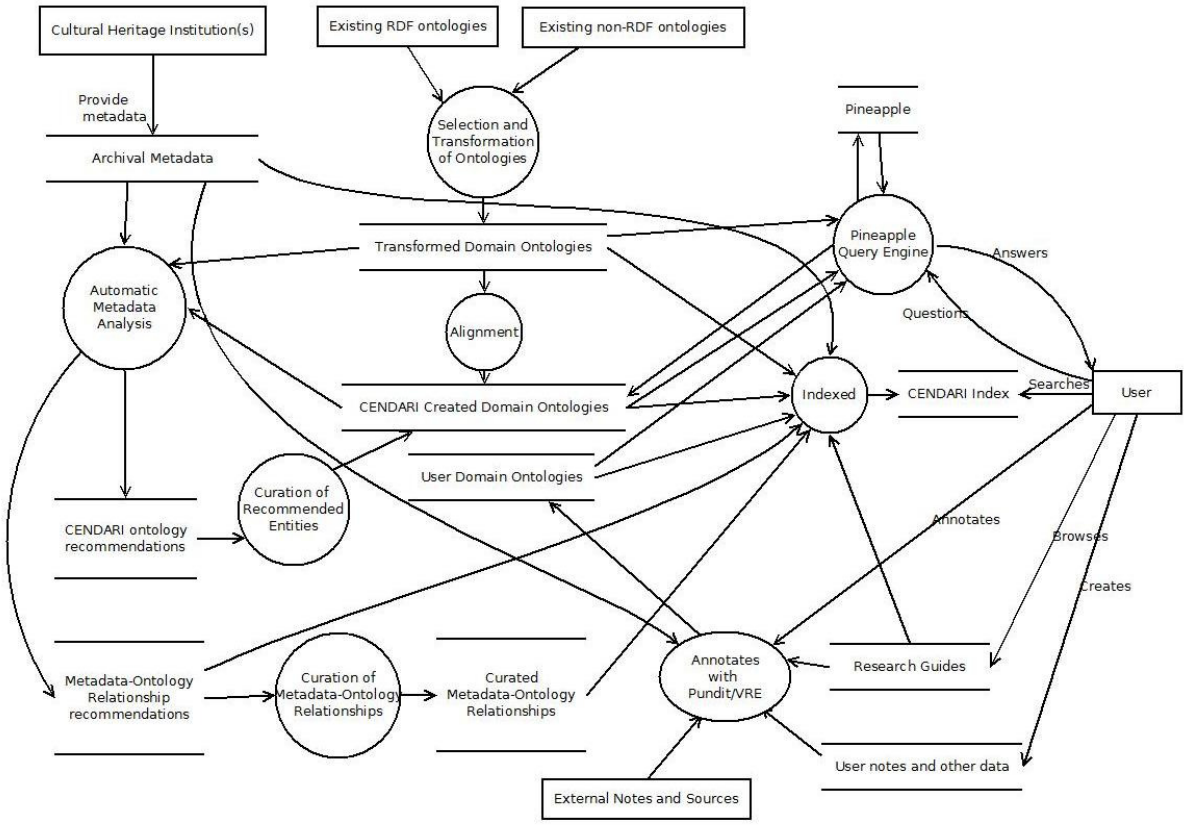


**Figure 2 Data flow diagram of CENDARI ontologies with other CENDARI data stores and services**

There are five types of external entities considered in the diagram:

- Cultural Heritage Institutions: The diagram shows the provision of metadata records by the CHIs. This is an over-simplification for clarity. The metadata records are created both by CENDARI researchers at different levels, as well as being automatically harvested. The model also ignores any necessary transformations to the metadata before it conforms to the CENDARI metadata model.
- Existing RDF Ontologies: There are a number of existing ontologies, already in an RDF format, that will be incorporated into CENDARI domain ontologies and linked to from the CENDARI domain ontologies.
- Existing non-RDF Ontologies: It is also important to recognize the non-RDF vocabularies that need to incorporated into the CENDARI domain ontologies.
- Users: 'Users' represents the imagined external user of the CENDARI infrastructure, that will engage with the content in various ways: questioning, searching, browsing, creating, and annotating.
- External notes and sources: Not all the content that will be used by CENDARI users will be hosted or indexed by CENDARI. Rather it will be incorporated through tools such as Pundit.

CENDARI will host a wide range of information, and there are eleven data stores represented in the diagram:

- Archival Metadata: Metadata will be created at institution, collection, and item level. These metadata records will be processed in three ways: they will be indexed, so that users can search; annotated by users using the VRE and Pundit; and be subject to automatic metadata analysis.
- CENDARI ontology recommendations: The automatic metadata analysis will produce to types of data, the first of which is ontology recommendations. NER will produce a set of suggested concepts and things may be fed into the domain ontologies, although this will need to be a curated process.
- Metadata–Ontology Relationship recommendations: The automatic metadata analysis will also combine data from the archival metadata and the domain ontologies to produce a list of suggested connections between the metadata records and the ontologies. As metadata records may continue to be ingested beyond the current funding of CENDARI, and the relationships may be of value irrespective of whether they have been curated, these will be indexed.
- Curated Metadata-Ontology Relationships: Those metadata-ontology relationships that have been curated and confirmed will be stored and indexed separately, so that users can identify the authority that may be ascribed to the relationships.
- Transformed Domain Ontologies: Existing RDF and non-RDF ontologies and other forms of knowledge organization system will be selected and transformed into a standardised data model (see section 4.2.1 for more details). These transformed domain ontologies will be indexed, queried by the Pineapple Query Engine , and inform the automatic metadata analysis.
- CENDARI Created Doman Ontologies: Although CENDARI domain ontologies will primarily reuse existing ontologies, some additional information will be created. For example, during the process of aligning existing knowledge organization systems and through ontology entity recommendations from NER (see section 4.6 for more details).
- User Domain Ontologies: CENDARI users may also create their own additional ontology instances, and relationships between existing entities, as they create notes and annotations with Pundit and the virtual research environment.
- CENDARI Index: All this information will be indexed so that it can be quickly searched.
- Pineapple: Pineapple will have its own data store for storing its formalizations of concepts that may not have been explicit in the domain ontologies. It will make use of the Transformed Domain Ontologies, the CENDARI Created Domain Ontologies, and a user's own User Domain Ontology.
- Research Guides: CENDARI will have a number of research guides, both created by CENDARI, and CENDARI users. It will be possible to browse, search (via the index), and annotate these guides with new and existing entities that will be stored within a user's personal User Domain Ontology.

- User notes and other data : Users will create a wide range of personal notes and content in the virtual research environment that can be annotated with concepts and entities and stored in a user's User Domain Ontology.

The different user communities can have different requirements from the domain ontologies. Whilst the virtual research environment and data extraction technologies provide the opportunity for the ongoing enhancement of the ontologies, there are also challenges in creating domain ontologies that can be used by historians and computer scientists. On the one hand historians are likely to be resistant to the idea of too strictly defining many of the concepts and entities that are the focus of their research, with even an event as well-known as World War 1 being open to debate as to when it started and finished. In comparison, using the ontologies as the basis for more sophisticated interfaces, such as Pineapple, may require more explicit terms. The CENDARI methodology starts with the primacy of the historians' ontological requirements, before considering solutions so that the ontologies can meet the requirements of other user communities.

The domain ontologies required by the two research communities supported by CENDARI are at one level quite distinct. The individuals and events that are of interest to historians of the medieval period will inevitably differ from those individuals and events that are of interest to those investigating World War 1. Even where there can be expected to be some level of continuity in a particular class of entities across the two periods, for example, in geographic places, there will inevitably be significant differences. Cities and historic states will have inevitably changed over the two periods, but so too will the aspects which the researcher is interested in; whereas researchers of the medieval period may find towns and cities the principle geographic terms of interest, the First World War scholar may also be interested in fronts, lines, or even specific trenches. Nevertheless, at another level, there are clear commonalities in the type of entities that are of interest to the researchers. In discussions with researchers about the types of entity that would need to be encoded within the CENDARI ontologies, six common types were identified:

- Places/spaces
- Persons/role
- Institutions
- Dates
- Events
- Topics

In addition to which it was recognized within the medieval domain that there are a finite number of manuscripts/shelf-marks that should also be incorporated, although this will be included within the associated metadata records. The commonalities allow for a shared upper ontology of broad classes of entities, with subclasses and relationships appropriate to the specific areas of historical research.

## 4.2 Ontology Reuse

The reuse of existing ontologies is particularly important within CENDARI due to the amount of resources already available in this two research areas and the extent of the research topics. The reuse of existing ontology instances also offers the possibility for greater multilinguality as existing vocabularies in multiple languages can be included. There are three types of ontology for reuse within CENDARI: a common upper ontology for structuring the common classes of entities; subclasses relevant to specific research areas; and ontology instances adhering to the ontology element set. So far, there has been an inconsistent approach to development of ontologies in similar projects.

The centenary of the start of the First World War in 2014 has seen a number of projects associated with the publication of First World War ontologies that provide useful insights into the different approaches that can be taken to the development of ontologies in the same field: WW1 Linked Data (Törnroos, 2013); Out of the Trenches (Pan-Canadian Documentary Heritage Network, 2012); and Trenches to Triples (Aim 24, 2012). Each project has adopted a different data model for representing the ontologies. WW1 Linked Data data model closely follows a subset of the more extensive CIDOC Conceptual Reference Model (CIDOC-CRM), which provides a structure for concepts and relationships used in cultural heritage documentation. Out of the Trenches created a bespoke data model which is not only informed by CIDOC-CRM but also the FRBR (Functional Requirements for Bibliographic Records) model from the International Federation of Library Associations and Institutions and ISAD(g) (General International Standard Archival Description) and ISAAR(CPF) (International Standard Archival Authority Record for Corporate Bodies, Persons and Families) from the International Council on Archives. In comparison the Trenches to Triples data does not adhere to some grand overarching integrated data model, but comprises of four distinct types of data: concepts, people, places, and organizations.

### 4.2.1 EDM as an upper ontology

Whilst the usefulness of adopting an existing upper ontology for CENDARI was recognized, it was decided to adapt the Europeana Data Model (EDM) in the first instance rather than CIDOC-CRM, or another upper ontology designed primarily for cultural heritage institutions. The EDM is less complex than CIDOC-CRM, but nonetheless aligns closely with the requirements of CENDARI, can be mapped to CIDOC-CRM, and already contains many equivalences to CIDOC-CRM (https://github.com/europeana/corelib/blob/master/corelib-solr-definitions/src/main/resources/eu/rdf/edm.owl). Importantly it is designed to be extensible, which means that it is possible to avoid too great a specialization at a high level.

Version of 5.2.4 of the *Definition of the European Data Model* (Europeana, 2013) has six classes for contextual resources: agents, places, timespans, concepts, events, and physical things. The implementation of the full set of classes is on an incremental basis, and version 2.0 of the *EDM Mapping Guidelines* currently only provide detailed descriptions of four of the classes (agents, places, timespans, and concepts), although EDM Object templates are available for events and physical things. The contextual classes identified by the CENDARI researchers can be aligned closely with those of the EDM model (see Table 2**Error! Reference source not found.**).

Table 2 Suitability of EDM to meet the needs of CENDARI researchers

| CENDARI contextual classes | EDM contextual classes |
|---|---|
| **Places/spaces**: geographic locations relevant to research topics or other contextual entities. | Places |
| **Persons/role**: individuals associated with the research topics or other contextual entities, and their associated roles. | Agents |
| **Institutions**: organizations associated with the research topics or other contextual entities. | Agents |
| **Dates**: specific dates or periods of time associated with specific events, people, or organizations. | Timespans |
| **Events**: notable events associated with the particular topics, as well as with other contextual entities. | Events |
| **Topics**: subjects associated with the two research area. | Concepts |

Although physical things were not identified as being of primary interest to the historians of the medieval and First World War periods, it will nonetheless adopted within CENDARI so that any appropriate existing ontologies may also be included. The EDM also has classes for describing Cultural Heritage Objects, although within CENDARI these will be replaced by the schemas identified within the metadata section (http://www.cendari.eu/public-project-deliverables/metadata/).

EDM is designed to be extensible, with the use of terms from RDFs to provide greater specialization through the use of rdfs:subClassOf and rdfs:subPropertyOf. This enables the creation of a finer level of granularity for a particular research community. For instance, an extension has already been created for manuscripts in the Digital Manuscripts to Europeana project (DM2E) (http://dm2e.eu/). Whilst DM2E makes use of the EDM it has added subclasses and subproperties to provide more information. For example, whereas EDM allows for the people to be related in an unspecified way through the use of the dc:relation property, DM2E allows the expression of a relationship more specific to the world of manuscripts, that of one person being influenced by the work of another (dm2e:influencedBy). Equally the edm:hasMet relationship, has been extended to express the more specific relationship of one person being the student of another (dm2e:studentOf).

Extensions to the EDM for each of the two research communities will be created, where possible making use of existing vocabularies that can be aligned to EDM. The necessary requirements of the extensions will be identified through analysis of existing domain ontologies and discussions with researchers within each of the domains.

The extensions will be able to contain all the necessary contextual instances, even if users wish to add a finer level of granularity with additional sub-classes and sub-properties at a later date. The instances, however, adhering to the ontology element set will necessarily be

more fluid, developing as the CENDARI infrastructure is used. In addition to those instances identified during the process of creating the knowledge framework, it is important that the ontologies are added to both as records are ingested into CENDARI, and as researchers make use of the CENDARI virtual research environment.

It is equally important that there are guidelines for the application of the EDM and the extensions for use within CENDARI. As is often the case with elements sets, the same information may be encoded in multiple different ways.

### 4.2.2 Limitations of EDM: SKOS v. OWL

Whilst the EDM includes the contextual classes that need to be included within the domain ontologies, that does not mean these classes are necessarily structured in a fashion that is suitable for all data and uses. The potential for difficulties most noticeably arises through the use of the SKOS vocabulary for encoding concepts and the relationships between them. Whilst SKOS can be used to encode many of the traditional knowledge organization systems, the traditional knowledge organization systems do not adhere to the stricter requirements of more the formal ontologies required by additional semantic services.

The SKOS vocabulary (http://www.w3.org/TR/skos-reference/) is designed for encoding traditional knowledge organizations systems (e.g., thesauri, classifications schemes, subject heading systems) in a manner that is suitable for the semantic web. A concept can not only be provided with a label (skos:prefLabel), but also express relationships with other concepts through the use of properties such as skos:broader, skos:narrower, and skos: related. Whilst SKOS is sufficient for encoding the traditional knowledge organization systems, it does not translate easily to OWL (the Web Ontology Language) that underpins CENDARIs proposed Pineapple service and other more complex semantic services. Hepp and de Bruijn (2007) identify two main problems inherent in the vocabularies themselves (as opposed to problems inherent in SKOS representations of the vocabularies): labels are not necessarily context-neutral and hierarchical relations are often not a strict subClassOf. At first glance the rdfs:subClassOf property may seem very similar to which is used by the skos:narrower property. However, whereas rdfs:subClassOf is transitive and all instances may be included within the larger classes, skos:narrower is not transitive and the relationship may be between very different types of thing.

The lack of a strict hierarchy and the inherent context associated with many terms is generally not a problem to human readers, but rather is a natural part of human language. The SKOS ontology has been specifically designed for the representation of these types of vocabularies, which is used in the EDM for representing concepts. It is, however, a problem for automatic reasoning and the inference required by additional technological services. For example, the computer scientists in CENDARI who wish to build services on top of the ontologies that make use of automatic reasoning and inferences will require more formal conceptualised terms than can be represented in SKOS.

The lack of a strict hierarchy can be seen within the Trenches to Triples concepts (see Appendix 1). For example, the concept “1917 Offensive (April-May)” has “Battle of Doiran, 1917 (24-25 April and 8-10 May)” as a narrower concept and “Macedonia” as a broader

concept. The 1917 Offensive and the Battle of Doiran may both be considered conceptualizations of events, and the Battle of Doiran is part of the 1917 Offensive. But whilst the 1917 Offensive may have been a battle on the Macedonian front, it is should be considered in any meaningful sense part of Macedonia.

This challenge is widely recognized, and a number of alternatives have been suggested for using both SKOS and OWL together (Bechhofer & Miles, 2008). The most practical solution seems to be to have two ontologies, or rather one ontology with a growing set of transformation rules. The original ontology will be created according to SKOS, with a set of transformation rules creating OWL. Although even with the use of additional tools and methodologies such as SKOS2OWL (http://www.heppnetz.de/projects/skos2owl), it will necessarily be resource intensive and the extent to which it can be applied across all SKOS vocabularies will have to be investigated.

A simpler alternative is to make use of two additional skos terms, skos:broaderTransitive and `skos:narrowerTransitive`. These are super-properties to `skos:broader` and `skos:narrower`, allowing for a presumption of transitivity between two concepts. In the example of the Battle of Doiran:

**'Battle of Doiran' --skos:broader---> '1917 Offensive' --skos:broader--> 'Macedonia'**

This can also be represented by a less strict relationship `skos:broaderTransitive`:

**'Battle of Doiran' --skos:broaderTransitive---> '1917 Offensive' --skos:broaderTransitive--> 'Macedonia'**

As `skos:broaderTransitive` is a transitive relationship, the fact that 'Battle of Doiran' has `skos:broaderTransitive` 'Macedonia' can be inferred, whereas 'Battle of Doiran' has `skos:broader` 'Macedonia' cannot.

By using `skos:broaderTransitive`, rather than breaking down the original vocabulary so that those relationships that should be transitive are recognized as such, it is a sleight of hand that allows relationships to be treated as transitive even if they are not. The usefulness of such an approach depends heavily on the original vocabularies, and will only become apparent after testing.

### 4.2.3 Instances

As well as existing ontology element sets for describing classes and properties, there are also many vocabularies of instances to adhere to these element sets; from general knowledge organization systems or knowledge bases that may have a proportion of instances that are appropriate to CENDARI's needs (e.g., Library of Congress Subject Headings (LoCSH) or DBpedia) to a system specific to a particular domain (e.g., the taxonomy of the 1914-1918 Encyclopaedia).

Some of the existing vocabularies will already be in a suitable machine readable format. For example, both LoCSH (http://id.loc.gov/authorities/subjects.html) and DBpedia (http://dbpedia.org) are available according to linked data principles:

> "1. Use URIs as names for things.
> 2. Use HTTP URIs, so that people can look up those names.
> 3. When someone looks up a URI, provide useful information…
> 4. Include links to other URIs…"
>
> (Heath & Bizer, 2011, p.7)

Other sources of data, however, may only be available in a paper format, or online in a format that needs to be converted (e.g., a PDF of authorised names).

An extensive literature search is necessary to identify many of the potential sources. This includes investigating the vocabulary and ontology libraries that are available online from around the world, for example:

- The Finnish Ontology Library Service - ONKI (http://onki.fi/)
- Basel Register of Thesauri, Ontologies & Classifications - BARTOC.org (www.bartoc.org)

It also includes general searching of online resources, and making use of the investigative and descriptive work done by WP5 in identifying the resources of the various cultural heritage institutions.

As well as identifying suitable resources, it is necessary to take into consideration the costs and benefits associated with a particular source. For example, a highly comprehensive source may be prohibitively expensive due to copyright restrictions or the amount of work necessary to transform it from an analogue into a digital format.

### 4.3 Identify appropriate software

There are three main software requirements for building the CENDARI domain ontologies: an ontology editor for building the EDM extensions; a data store for storing the instances adhering to the EDM extensions; and an interface for ontology alignment (i.e., matching equivalent instances from the different vocabularies). Although ontology editors can be used to create ontology element sets, populate them with instances, and align instances from different vocabularies, there was no single ontology editor that met all the requirements of an ontology editor for CENDARI.

An all-in-one solution would have three main requirements: be web-based to allow for distributed editing of the ontologies; be able to merge existing ontologies; and be either free or at least not prohibitively expensive. Separating the storage of the ontology instances, the alignment of different vocabularies, and the development of the ontology element sets means that the ability to merge existing ontologies is no longer essential, and WebProtege (http://protegewiki.stanford.edu/wiki/WebProtege) seems to provide a suitable free distributed editing tool.

The obvious solution to the problem of storing the ontology instances is to make use of triple store. This would make the ontologies easily accessible to complex querying by both CENDARI services and external services. Other services used by CENDARI also need to

make use of a triple store, and in the case of the annotation tool Pundit (www.thepund.it) the original requires a specific triple store, i.e., Sesame (www.openrdf.org). The particular triple store is not particular important from the perspective of ontology building, as long as it has the requisite SPARQL endpoint and RESTful APIs.

Rather than entering the information directly into the triple store, however, an additional step was incorporated so that the provenance of the data could more easily be incorporated.

Rather than extending the EDM to include additional provenance information, it was decided to have an intermediate step, with the instances from a particular source (e.g., DBpedia) being stored as an RDF/XML within the data management system CKAN, from which it can be automatically ingested into the triple store (see Figure 3Figure 1).

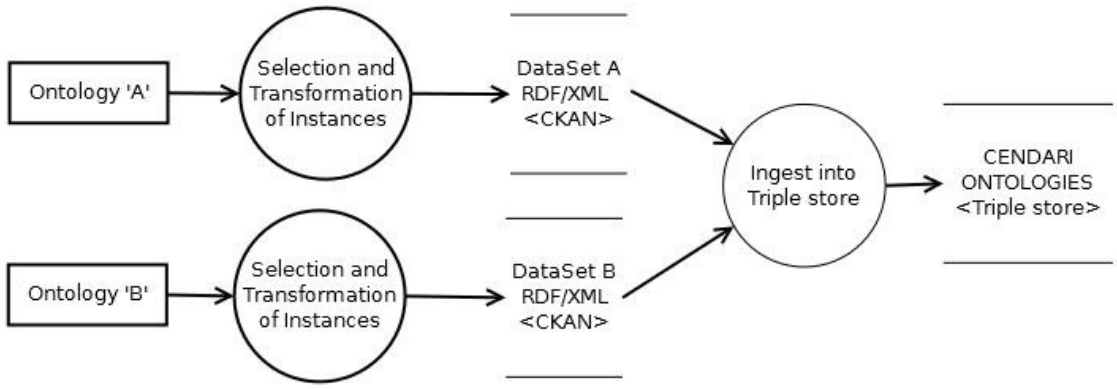


Figure 3 Incorporating existing instances into the CENDARI triple store

CKAN provides a certain amount of basic metadata functionality, allowing data sets to have descriptions, tags, a revision history, and the licence under which it is shared (if indeed it is being shared). It also allows for the additional customizable metadata name-value pairs. However, as each dataset will contain a variety of file types (e.g., original files, transformation files, and transformed files) it is important that the associated different file types are clearly identifiable. This may be achieved by creating a METS file in the root directory of each dataset using the same name. The <fileSec> element allows for the grouping of multiple elements into different groups (http://www.loc.gov/standards/mets/METSOverview.v2.html#filegrp).

When the RDF triples conforming to the EDM are uploaded to the triple store, each dataset will form its own named graph. Queries will then be run against the set of graphs that are open to the particular user. The CENDARI ontology of alignment instances, which matches

instances in different vocabularies, will form its own separate dataset. The topic of ontology alignment and the necessary software is returned to below (see section 4.6).

### 4.4 Knowledge acquisition

Knowledge acquisition refers to capturing the domain knowledge that doesn't already exist within an ontology that can be used. Within CENDARI the focus is primarily on the transformation and curation of ontologies that already exist. Although this may then be supplemented by instances created through named entity recognition (NER) and user contributions. It is unrealistic to expect a completely exhaustive ontology that meets all CENDARI users' needs to be created as part of the creation of the knowledge framework. This is most immediately obvious in the case of the First World War, where there are not expected to be comprehensive lists of every soldier or event, and the archives rarely adopt widely standardised vocabularies. Even within the medieval area of research, where the primary objects of study are a known corpus of medieval manuscripts that have been extensively studied for many years and have an extensive collection of associated vocabularies, the ontology will need to adapt to reflect new perspectives and approaches taken to investigating the manuscripts.

NER provides the opportunity for the automatic analysis of a large number of metadata records that would not be possibly otherwise. The NER process can suggest named entities within the records that can either be linked to existing instances within the CENDARI ontology, or form the basis of additional instances for the CENDARI ontology.

There will also be new instances and relationships that users wish to add to the CENDARI ontologies. This may be when users are making use of the virtual research environment, adding annotations to documents, or in the process of creating archival research guides.

NER and crowd-sourced suggested entities will not necessarily produce entities and relationships that can be considered as trustworthy as those produced by the CENDARI researchers. As such identifying important terms is not only part of selecting terms to be part of the CENDARI ontologies, but also enabling users to distinguish between the source of the instances.

### 4.5 Identify important terms

There are two parts of the identifying important terms within CENDARI: identifying important terms within existing ontologies, and identifying important terms identified by NER and CENDARI users.

Section 4.2 discussed the importance of reusing existing domain ontologies within CENDARI, however this does not necessarily mean CENDARI will want to incorporate all the instances within a particular ontology. For example, DBpedia is an important data source of millions of triples, and highly integrated into the Linked Open Data cloud, but whilst there a lot of information that CENDARI would wish to incorporate into the CENDARI ontology, there is also a lot of information that is highly unlikely to be of use to researchers using CENDARI

(e.g., the large quantity of data about the Start Trek series, films, characters, and spaceships!).

It is anticipated that additional contributions will be to the ontologies both through the use of named entity recognition and users engaging with the CENDARI infrastructure. It is important, however, to distinguish between identified important terms and suggested important terms, and a verification process will be necessary before a suggested important term moves to an identified important term.

### 4.6 Identify additional terms, attributes and relationships & specify definitions

The creation of the CENDARI ontologies brings together many different sources of information about the same or similar entities, as well as new entities identified by NER and suggested by users. CENDARI needs to align the ontologies and manage potentially conflicting information about the same entities, and facilitate the addition of detailed new entities.

**Error! Reference source not found.** provides a more detailed version of the ingestion process than was originally represented in **Error! Reference source not found.**. Showing not only the ingestion process of two ontologies to the triple store, but also the relationship between these ontologies and the relationships that are made within CENDARI and by other users.

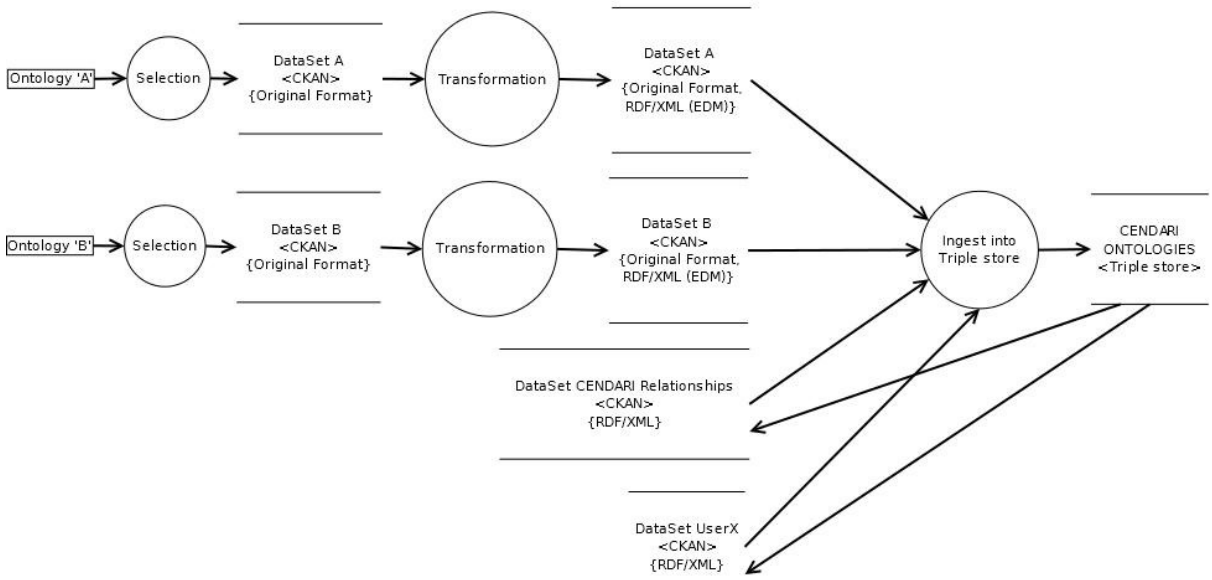


Figure 4 Existing ontologies, CENDARI relationships, and user entities in CENDARI triple store

#### 4.6.1 CENDARI ontology of alignment instances

Part of the CENDARI created domain ontologies (see Figure 2) will be the relationships between the different ontologies. This will contain universal identifiers for entities that may appear in multiple vocabularies, and build relationships with various incarnations of these

entities. This means that where different information about a particular entity is available from different sources the information can be aggregated into a single entity graph, whilst at the same time keeping the provenance of the data.

For example, it is quite conceivable a person, General X, will appear in multiple existing ontologies, albeit with only partial information in each. In one dataset information may be available about General X's birthdate, whilst another may contain information about the deathdate:

### Ontology A

```
…/persons/ontologyA_generalX  skos:prefLabel          "General X"
…/persons/ontologyA_generalX  rdaGr2:dateOfBirth      "1870"
```

#### Ontology B

```
…/persons/ontologyB_generalX  skos:prefLabel          "General X"
…/persons/ontologyB_generalX  rdaGr2:dateOfDeath      "1916"
```

**The CENDARI ontology expresses the relationship between the two records so that they can be brought together in one place.**

#### CENDARI ontology of alignment instances

```
…/persons/cendari_generalX    owl:sameAs  …/persons/ontologyA_generalX
…/persons/cendari_generalX    owl:sameAs  …/persons/ontologyB_generalX
```

Where there is conflicting information all variations are available, although the information that is displayed in the default record can be based on hierarchy of authoritativeness (e.g., Library of Congress triples take precedence over DBpedia triples, which in turn takes precedence over information identified by automatic data extraction tools). The CENDARI ontology of alignment instances is only one graph of the many that will be in the CENDARI triplestore, and like any of the graphs users will be able to choose not to incorporate a particular graph when querying the ontologies.

Ontology matching tools require different amounts of human input, which a lack of human input adversely affecting the quality of the matching. Fully automatic ontology matching tools may achieve a result quality of 70% for single language matching and 40% for multi-lingual matching (Paulheim, Hertling, & Ritze, 2013). As the CENDARI domain ontologies are designed to be of high quality, and may be multilingual, it is important that an interactive approach is taken to ontology development. There are number of interactive tools available for the matching of instances, for example:

- LogMap: http://www.cs.ox.ac.uk/isg/tools/LogMap/
- YAM++: Ngo & Bellahsen (2012)
- CODI: https://code.google.com/p/codi-matcher/
- WeSeE: http://www.ke.tu-darmstadt.de/resources/ontology-matching/wesee-match/

Each of these tools suggests matches according to different criteria, and some may be suitable for certain sets of data rather than others. As such it is important to test the suitability of the tools against a sample of the data within each of the two topic areas. There are no particular restrictions on the type of interactive software that is suitable. For example, it does not matter whether the software is hosted on the desktop or on a server, as all that is required is it is accessible at one computer.

### 4.6.2 User instances/alignment

It is also important that researchers making use of the CENDARI infrastructure are able to contribute additional instances to the domain ontologies, and suggest alignments and relationships between instances. This will require a different interface to the large scale alignment carried out by those working on the creation of the CENDARI ontologies, focusing more on the addition of single instances, and the relating of one instance to another. The adding of additional instances is unlikely to be done in isolation, but rather in connection with particular documents and sources.

**Tagging a document**

The minimum type of knowledge organization system functionality that CENDARI users would expect to be able to make use of would be tagging, the assignment of keywords to aid with description, organization, and retrieval (Macgregor & McCulloch, 2006). This assignment may be the association of an existing concept or entity with a document, or the creation of a new concept.

It should be simply possible to associate a document with one or more of the six contextual types in the EDM, following the broad relationship **edm:isRelatedTo that is used within the EDM to link cultural heritage objects with concepts and other resources.**

**Applying a tag to an object will create a concept in a personal graph (either public, private, or shared). This concept can then be linked to the thing it stands for (agent, time-span, place, event, or physical thing) through the use of focus:focus (**http://xmlns.com/foaf/spec/#term_focus**). The VRE should suggest existing items in the CENDARI graph, with the opportunity to create a thing based on a simple form if they wish. Where an existing thing is not complete (or a user wishes to include alternative details) it should be possible to create new triples stored within the user's personal graph. It is important to recognize that many of these personalized concepts will have no associated thing, but rather will be for personal organizational purposes (e.g., toPrint, returnToLater).**

**Relating multiple things/concepts**

As well as associating things and concepts with documents, it may also be the case that a CENDARI user wishes to build relationships between multiple things or concepts:

- Build a relationship between two existing concepts or things.
- Create a new thing or concept, and relate it to an existing concept/thing.

It is expected that this will revolve around the VRE/annotation tool and either the documents or notes that the user is making. In such cases where one or more concepts/things have been associated with the document or note, it should be possible to make a connection between the different concepts and things, and allow the user to select one of the possible relationships defined in the medieval/WW1 EDM extensions. These relationships will be stored in a user's private, public, or shared graph.

The possibility for the addition of new things, concepts, and relationships, not associated with any particular document will also be investigated. Such a tool could provide a visual interface to a user expanding the CENDARI domain ontologies, which nonetheless restricts them to the schemas of the EDM medieval/WW1 extensions. These relationships will be stored in a user's private, public, or shared graph.

**Extending the CENDARI Element Set Extensions**

In future developments it may be desirable to allow the users to extend the CENDARI element set extensions, with finer distinctions within the EDM for the particular domains. Such functionality, however, is not likely to be widely used, and can probably be reconsidered in future iterations.

### 4.7 Integration with existing ontologies

As well as the building of connections between the things within the different ontologies, it is also necessary to integrate with the existing metadata schemas.

Explicit mappings from the appropriate EAG/EAD/CCS/MODs elements to EAD and the First World War and the medieval extensions will be created.

### 4.8 Implementation

The process for implementing the ontology is shown in **Error! Reference source not found.** above.

### 4.9 Evaluation

Evaluation of the domain ontologies, and the ontology process, will not only consider their ability to support CENDARI researchers but the ability to interact successfully with other CENDARI semantic services.

From a technical perspective the ontologies will be evaluated in terms of their ability to successfully interact with other services. Especially with regards to the more formal ontological requirements of the Pineapple service.

From an user-centric perspective the ontologies will be evaluated not only in terms of how well they meet users stated coverage requirements, but also their actual requirements as the first users query the system.

It is important that the ontologies that are created appropriately balance the needs of the researchers and the potential of the technologies, otherwise the archival divide between archivists and historians is merely going to be replaced by a technological divide between computer scientists and historians.

### 4.10 Documentation

Documentation is an often overlooked part of publishing ontologies, which is particularly problematic as the final ontologies are heavily influenced by the technologies and process that have been used to create them. As well as the publishing of the domain ontologies, and the Europeana Data Model extensions, a final version of the ontology building guidelines will also be publish.

## 5. Conclusion

These guidelines for ontology building for CENDARI reflect the state of understanding as of the end of April 2014. The process itself is iterative, and the scale of CENDARI ontology creation means that the process for their creation is heavily reliant on many of the technology decisions made in conjunction with other work packages.

## References

Aim 24 (2012) Trenches to Triples. http://data.aim25.ac.uk/about_t3.php. Accessed 2 December 2013

Bechhofer S, Miles A (2008) Using SKOS and OWL. W3C. http://www.w3.org/2006/07/SWD/SKOS/skos-and-owl/master.html. Accessed 3 December 2013

Europeana (2013) Definition of the Europeana Data Model v5.2.3. http://pro.europeana.eu/documents/900548/0d0f6ec3-1905-4c4f-96c8-1d817c03123c

Fernández-López M, Gómez-Pérez A, Juristo N (1997) METHONTOLOGY: From Ontological Art Towards Ontological Engineering. Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series, 33-44. http://oa.upm.es/5484/1/METHONTOLOGY_.pdf. Accessed 2 December 2013

Gruber TR (1993) A translation approach to portable ontology specifications. Knowl Acquis 5(2):199-220.

Grüninger M, Fox MS (1995) Methodology for the Design and Evaluation of Ontologies. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.44.8723&rep=rep1&type=pdf. Accessed 2 December 2013

Hepp, M., de Bruijn, J. (2007) GenTaz: A generic methodology for deriving OWL and RDF-S Ontologies from Hierarchical Classifications, Thesauri, and inconsistent taxonomies. Proceedings of the 4th European Semantic Web Conference (ESWC 2007), June 3-7,

Innsbruck, Austria, Springer LNCS Vol. 4519, Springer 2007, pp.129-144. http://www.heppnetz.de/files/hepp-de-bruijn-ESWC2007-gentax-CRC.pdf

Library Linked Data Incubator Group (2011) Datasets, value vocabularies, and metadata element sets. http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

Macgregor, G., & McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review,* 55(5), 291-300.

Ngo, D., & Beelahsene, Z. (2012) YAM++: (not) Yet Another Matcher for Ontology Matching Task. *BDA 2012. http://hal-lirmm.ccsd.cnrs.fr/docs/00/72/06/48/PDF/YAM%2B%2B_not_Yet_Another_Matcher_for_Ontology_Matching_Task.pdf*

Noy NF, McGuinness DL (2001) Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. http://www.w3.org/TR/owl2-overview/. Accessed 2 December 2013

Pan-Canadian Documentary Heritage Network (2012). Out of the Trenches: Linked Open Data of the First World War: Final Report. http://www.canadiana.ca/sites/pub.canadiana.ca/files/PCDHN%20Proof-of-concept_Final-Report-ENG_0.pdf Accessed 2 December 2013

Paulheim, H. Hertling, S. & Ritze, D. (2013). Towards evaluating interactive ontology matching tools. In: Cimiano, P., Corcho, O., Presutti, C., Hollink, L., Rudolph, S. (eds). The Semantic Web: Semantics and Big Data. Springer: Berlin. pp. 31-45.

Törnroos J, Mäkelä E, Lindquist T, Hyvönen E (2013) World War 1 as Linked Open Data. Semantic Web J. In press. http://www.semantic-web-journal.net/content/world-war-1-linked-open-data. Accessed 2 December 2013

Uschold M, King M (1995) Towards a Methodology for Building Ontologies. Presented at "Workshop on Basic Ontological Issues in Knowledge Sharing". http://www1.cs.unicam.it/insegnamenti/reti_2008/Readings/Uschold95.pdf. Accessed 2 December 2013

W3C (2012) OWL 2 Web Ontology Language – Document Overview. http://www.w3.org/TR/owl2-overview/. Accessed 2 December 2013

# Appendix 1 – Trenches to Triples:Concept to the Europeana Data Model

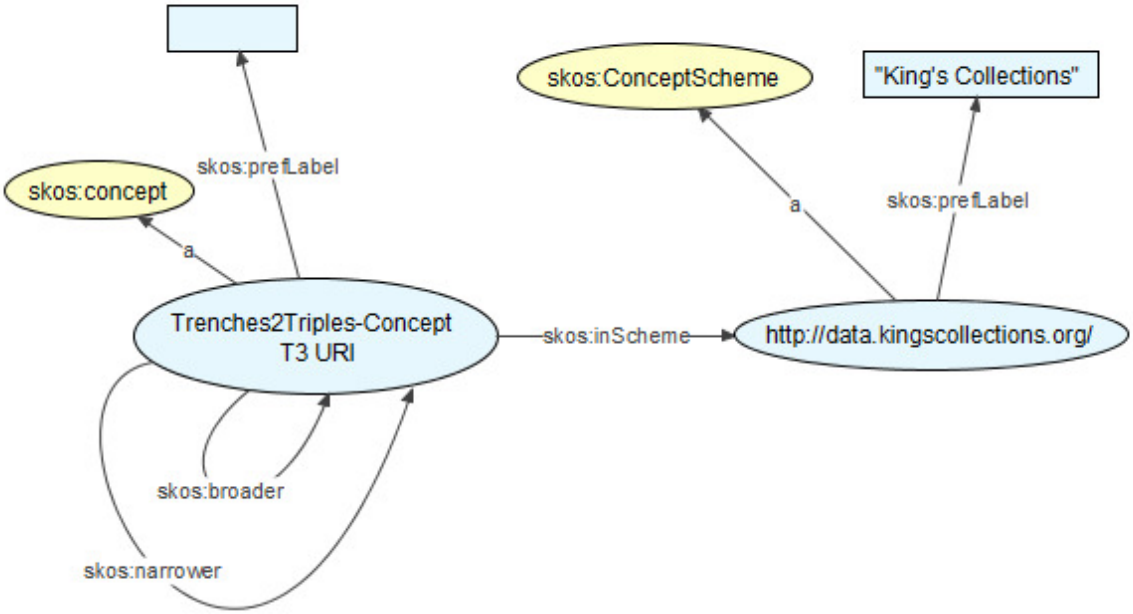


Figure 0-1 Trenches to Triples – concept data model (based on analysis of a number of concept records)

Trenches to triples already makes use of the SKOS vocabulary. These is, however, some repetition both within the record and across multiple records. There is also a need for CENDARI URIs.

**ORIGINAL:**

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:skos="http://www.w3.org/2004/02/skos/core#"
   xmlns:madsrdf="http://www.loc.gov/mads/rdf/v1#"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   xmlns:msg0="http://data.archiveshub.ac.uk/def/">
   <madsrdf:Topic
      rdf:about="http://data.aim25.ac.uk/id/concept/1917offensive=28april-may=29">
      <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
      <skos:Concept
         rdf:about="http://data.aim25.ac.uk/id/concept/1917offensive=28april-may=29">
         <skos:prefLabel xml:lang="en">1917 Offensive (April-May)</skos:prefLabel>
         <skos:inScheme>
            <rdf:Description rdf:about="http://data.kingscollections.org/">
               <skos:prefLabel xml:lang="en">Kings Collections</skos:prefLabel>
               <rdf:type resource="http://www.w3.org/2004/02/skos/core#ConceptShemes"/>
            </rdf:Description>
         </skos:inScheme>
```

```
         <skos:narrower>
            <rdf:Description
               rdf:about="http://data.aim25.ac.uk/id/concept/battleofdoiran,1917=2824-
25apriland8-10may=29">
               <skos:prefLabel xml:lang="en">
                      Battle of Doiran, 1917 (24-25 April and 8-10 May)
               </skos:prefLabel>
               <rdf:type resource="http://www.w3.org/2004/02/skos/core#Concept"></rdf:type>
            </rdf:Description>
         </skos:narrower>
         <skos:broader>
            <rdf:Description rdf:about="http://data.aim25.ac.uk/id/concept/macedonia">
               <skos:prefLabel xml:lang="en">Macedonia</skos:prefLabel>
               <rdf:type resource="http://www.w3.org/2004/02/skos/core#Concept"></rdf:type>
            </rdf:Description>
         </skos:broader>
      </skos:Concept>
   </madsrdf:Topic>
</rdf:RDF>
```

## TRANSFORMED RDF:

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf=”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
   xmlns:skos=”http://www.w3.org/2004/02/skos/core#”
   xmlns:owl="http://www.w3.org/2002/07/owl#">
   <skos:Concept
      rdf:about="http://data.aim25.ac.uk/id/concept/1917offensive=28april-may=29">
      <SAMEAS>
      <skos:prefLabel xml:lang="en">1917 Offensive (April-May)</skos:prefLabel>
      <skos:inScheme>
         <rdf:Description rdf:about="http://data.kingscollections.org/">
            <skos:prefLabel xml:lang="en">Kings Collections</skos:prefLabel>
               <rdf:type resource="http://www.w3.org/2004/02/skos/core#ConceptShemes"/>
         </rdf:Description>
         </skos:inScheme>
         <skos:narrower>
            <rdf:Description
               rdf:about="http://data.aim25.ac.uk/id/concept/battleofdoiran,1917=2824-
25apriland8-10may=29"/>
         </skos:narrower>
         <skos:broader>
            <rdf:Description rdf:about="http://data.aim25.ac.uk/id/concept/macedonia" />
         </skos:broader>
      </skos:Concept>
   </madsrdf:Topic>
</rdf:RDF>
```