The CrossCult Knowledge Base: a co-inhabitant of cultural heritage ontology and vocabulary classification

Andreas Vlachidis[1], Antonis Bikakis[1], Daphne Kyriaki-Manessi[2], Ioannis Triantafyllou [2], Angeliki Antoniou[3].

[1] Department of Information Studies, University College London
{a.vlachidis,a.bikakis}@ucl.ac.uk,
[2] Department of Library Science and Information systems,
Technological Educational Institute of Athens,
{dkmanessi,triantafi}teiath.gr,
[3] Department of Informatics and Telecommunications,
University of Peloponnese,
angelant@uop.gr

Abstract. CrossCult is an EU-funded research project aiming to spur a change in the way European citizens appraise History, fostering the re-interpretation of what they may have learnt in the light of cross-border interconnections among pieces of cultural heritage, other citizens' viewpoints and physical venues. Exploiting the expressive power, reasoning and interoperability capabilities of semantic technologies, the CrossCult Knowledge Base models and semantically links desperate pieces of Cultural Heritage information, contributing significantly to the aims of the project. This paper presents the structure, design rationale and development of the CrossCult Knowledge Base, aiming to inform researchers in Digital Heritage about the challenges and opportunities of semantically modelling Cultural Heritage data.

Keywords. Cultural Heritage, Ontology, Digital Humanities, Semantic Web, Vocabulary Classification, CIDOC-CRM

1. Introduction

Without any doubt the era of digital distribution has introduced new exciting avenues for producing, accessing and consuming information. Within this realm, access to cultural heritage information has been significantly benefited by digital technologies, facilitating new ways of engaging with heritage and broadening public participation. Such advances not only enable an interactive engagement with heritage, but also reinstitute what we mean by heritage and how it can be accessed [1].

The CrossCult Project[1] realising the advances of digital technologies, particularly focused on the aspects of interactivity, recollection, and reflection, aims to spur a change in the way European citizens appraise History. By facilitating interconnections among pieces of cultural heritage information, public view points and physical venues, the project aims to foster the re-interpretation of history as we know it, which goes beyond the conventional siloed presentation of historical data, and focuses on aspects that are cross-cultural, cross-border, cross-religion, and cross-gender qualities.

A key contribution to this endeavour is the creation of a semantic knowledge base capable of interrelating a wide set of (existing and future) disparate digital cultural heritage resources. This paper discusses the scope of the CrossCult Knowledge Base, the design choices leading to the definition of its underlying Upper-level ontology, and the data-modelling outcome of a data sample. The Upper-level ontology delivers formalisms that describe the "world" of CrossCult, accommodating common conceptual arrangements, enabling augmentation, semantic-based reasoning and retrieval across disparate data resources.

Section 2 outlines relevant projects and the role of standard conceptual models for mediating semantic interoperability. Section 3 discusses the aims and design choices leading to the definition of the CrossCult Upper-level ontology. Section 4 presents the results of a data modelling exercise aimed at applying the conceptual arrangements and definitions of the CrossCult Upper-level ontology to a range of cultural heritage data resources. The discussion of a particular data modelling follows in Section 5, providing an insight the opportunities and limitations of the adopted modelling method. The last two sections highlight the most important lessons learned while defining and using the CrossCult Upper-level ontology and present the future steps towards finalising the semantic modelling endeavour.

2. Background

A fundamental problem area in dealing with Cultural Heritage data is to make the content mutually interoperable, so that it can be searched, linked, and presented in a harmonised way across the boundaries of the datasets and data silos [2]. In the sphere of contemporary information science, there is abundance of instruments for managing and modelling any kind of information including cultural heritage data. The Dublin Core (DC) Metadata Elements and DC Terms[2], the Simple Knowledge Organization System (SKOS[3]), the Functional Requirements for Bibliographic Record (FRBR[4]), the Europeana Data Model (EDM[5]), the CIDOC-CRM[6], the MIDAS Heritage standard[7], the Lightweight Information Describing Objects (LIDO[8]) and the VRA Core[9] to name

[1] http://www.crosscult.eu

[2] http://dublincore.org/documents/dcmi-terms/

[3] https://www.w3.org/2004/02/skos/

[4] https://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[5] http://pro.europeana.eu/page/edm-documentation

[6] http://www.cidoc-crm.org/

[7] https://historicengland.org.uk/images-books/publications/midas-heritage/

[8] www.lido-schema.org

[9] https://www.loc.gov/standards/vracore/

but a few, have been employed by numerous projects to harmonise access to content across disparate datasets [3]. Each model contains merits and limitations determined by its scope and origin. Some models are defined as nationally accepted standards, whereas others enjoy an international consent. Some models are domain independent and lightweight, others are more closely related to particular domain, some are described as harvesting metadata models and others present integrated manifestations.

In spite of the abundance of models and standards, the nature of cultural heritage data is such that does not simply lend to a straightforward cataloguing of information in the same way as warehouse data, administrational information or even library catalogues [4]. Influenced by different scholarly disciplines and perspectives, the cultural heritage data contain an inherited variability that is reflected by a range of different types of historical objects with their different characteristics. Hence, it is crucially important semantic interpretation of cultural heritage data to be driven by real world concepts and events modelling data based on the relationships between empirically surfaced arrangements rather than artificial generalisations and fixed field schemas [5].

During the past decade, the CIDOC-CRM, a core ontology for cultural heritage data, has matured and gained a growing popularity among projects aimed at providing data aggregation and semantic harmonisation of cultural heritage information. Standing for Conceptual Reference Model (CRM) of the International Council of Museums (ICOM) – International Committee for Documentation (CIDOC), CIDOC-CRM is a well-established ISO standard (ISO 21127:2006) in the modelling of cultural heritage information [6]. It provides an extensible semantic framework that any cultural heritage information can be mapped to.

The applicability of the CIDOC-CRM in information systems of the broader cultural heritage domain is evident in the literature by numerous large-scale projects such as, the Oxford University CLAROS[10] project, the British Museum ResearchSpace[11] and the EU FP7 Ariadne Infrastructure[12]. The above projects integrate vast datasets of classical antiquity, museum exhibits and archaeological research respectively, providing semantic interoperability and access to data based on the ontological and conceptual definitions of CIDOC-CRM. Specialisation of CRM instances to a terminological level is achieved by linking to external vocabulary sources, thesauri and classification schemes.

The CRM ontology provides a general mechanism for linking to terminological specialisations via the implementation of the E55 Type class, which enables connection to categorical knowledge commonly found in cultural documentation. A common implementation approach is to link CRM instances to thesauri concepts expressed as SKOS concepts. Simple Knowledge Organization System (SKOS) is a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, or any other type of structured controlled vocabulary [7]. It builds upon RDF and RDFS, and its main objective is to enable easy publication and use of such vocabularies as linked. SKOS structures can be linked to CIDOC-CRM instances to provide a specialised vocabulary.

[10] http://www.clarosnet.org/

[11] http://www.researchspace.org/

[12] http://www.ariadne-infrastructure.eu

3. Upper-level Ontology – Definition and Requirements

The CrossCult Upper-level ontology is defined as a generic upper-level conceptual structure that captures common concepts and relationships across a diverse range of cultural heritage data. As such, the ontology delivers formalisms that describe the “world” of CrossCult; it accommodates common conceptual arrangements and enables augmentation, linking, semantic-based reasoning and retrieval across disparate data resources. In order to achieve its semantic interoperability aims the Upper-level ontology adopts a single and generic upper-level design, based on a robust ontological definition, enabling efficient semantic-based reasoning and retrieval, while being scalable to be extended formally to specialised conceptual needs when required.

Specified as a knowledge representation resource benefiting from maximum reuse of established semantic web resources and standards, the Upper-level ontology adopts the standard ontology for modelling cultural heritage data, CIDOC-CRM. The use of CIDOC-CRM guarantees integration under well-defined and interoperable semantics that support the generic aims of the upper-level structure whilst providing specialisations that can benefit the individual needs of pilots. On the other hand, CIDOC-CRM as a formal and generic structure of concepts and relationships is not tied to any particular vocabulary of types, terms and individuals. This level of abstraction, albeit useful for the semantics of the broader cultural heritage domain, does not cover the need for a finer definition of types, terms and appellations. The need for an additional level of vocabulary semantics is addressed by the use of thesauri and glossary supplementing the CIDOC-CRM with specialised terms.

3.1 Rationale and Design Choices

In the process of defining the ontological arrangements, the project reviewed the pilots' datasets and engaged in a series of meetings before concluding to a set of requirements and shared semantics across the four pilot’s scenarios and data. The results led to the definition of the CC Upper-level ontology, which reuses terminology and maintains full compatibility with the widely-used standard in cultural heritage documentation CIDOC-CRM (ISO 21127:2006). The version of the upper-level ontology is a subset of CIDOC-CRM enhanced with additional semantics from the SKOS and FOAF [8] ontology.

The Upper-level ontology accommodates the range of shared semantics of the following commonly identified concepts across the four pilots; a) Physical items, as is any museum artefact, painting, venue item or landmark, b) Digital (audio-visual) content relating to one or more Physical Items, c) Places of spatial focus, which could refer to the location of an object, a place of an event or a depicted place on a painting, d)Time related definitions such as dates and periods, e) Actor as a person or organisation related to a physical item by properties of ownership creation and illustration and f) Reflective Topics carrying the semantics of subjects and topics of interest that drive the reflection and reinterpretation qualities of the application.

The Crosscult specific class *Reflective Topic*, acts as collection of primarily physical items (i.e E22_Man-made Objects) which are aggregated under a common theme that enables interaction with the content, based on predefined reflection and reinterpretation threads. Instances of the class (threads) can be topics such as Immigration, Women in Society, Healing, Painting Style, etc. Each instance contains links to relevant subjects from the CCCS vocabulary enabling retrieval and cross-reference, narratives describing the topic, associations to reflection modules (e.g. quiz games and ratings), while it realises standard CIDOC-CRM relationships across individual physical items in terms of their location, material, date of production etc. For example, the individual CC2279 (Figure 1), is a tombstone of the Middle Antonine period located at the Museum of Tripolis (Greece), and participating (cc.reflects) in the Reflective Topic *Woman Appearance* . Associations between individual physical items can be made through the use of a common reflection topic whereas other types relationships can be explored via the standard CIDOC-CRM properties.

3.2 Vocabulary Requirements and Semantics

The upper-level ontology incorporates the SKOS semantics, specifically the SKOS Concept and Concept Scheme classes and their associated properties, to provide access to specialised vocabularies. In CrossCult this need is met by a custom built Classification scheme (CCCS) aiming at enhancing the concept representation of the reflective topics developed by the four pilots. This is supplemented by domain dependent vocabularies of geographical and chronological terms.

The CCCS supplements the CC ontology by providing an additional layer of semantics through a controlled vocabulary of concepts providing a concise representation of reflection themes and their interrelationships and guiding the reflective process through these interrelations. The vocabulary incorporated into CCCS accommodates the reflective topics and the relevant social and cultural terminology. In this sense the Classification Scheme can be used as a means for modelling vocabularies contributing to the cultural heritage domain.

The scheme aggregates terminology from standard thesauri resources such as, the Arts and Architecture Thesaurus of Getty (AAT), the EUROVOC, the UNESCO Thesaurus and the Library of Congress Subject Authorities (LC) vocabulary, whereas it incorporates a limited number of CrossCult specific terminology designed to accommodate specialized needs of the reflective process deriving from the pilots' scenarios and narratives. The vocabulary is organised and defined in a hierarchical order of broader-narrower term relationships, whilst CCCS terms can be employed both as "types" (instances of the E55 Type class) and as "propositional objects" (instances of E89) to describe the subjects related to individuals of the CC Upper-level ontology.

The reuse of standardised resources ensures the validity of the CCCS structure and the consistency in the use of its terms. To a lesser extent, project specific terminology has been incorporated into the CCCS and has been inter-weaved within its structure. To ensure the comprehensiveness of CCCS and to maintain the project specific focus of the terminology, the contributing terms are derived from the scenario descriptions of the four pilots and the descriptions of relevant cultural heritage objects, including their meaning, symbolism, materials, cultural context and creative techniques. The definition of the CCCS involved the following steps: a) Identification of relevant vocabulary based on reviewing pilot scenarios and items involved for the building of the scenarios. This section relied heavily on cooperation with the historians, museum and venue curators and social scientists participating to the project as field experts; b) Verification of vocabulary against authority thesauri and incorporation of authority terms as preferred terms when applicable; c) Integration of mapped terms into the CCCS structure considering both original and CCCS hierarchies; d) Further enhancement of CCCS vocabulary with related terms, suggested by the mappings with authority thesauri; e) reviewing of CCCS structure and supplementing hierarchies as needed

4. Data Modelling

Data modelling in the context of this paper refers to the specific process of applying the conceptual arrangements and definitions of the CrossCult Upper-level ontology to a range of disparate cultural heritage data resources. The origin of the data as well their coverage and granularity vary significantly.

Four distinct pilots contribute data to the CrossCult project covering a unique range of cultural heritage venues across Europe. From the large venue of National Gallery in London to the considerably smaller venue of the Archaeological Museum in Tripolis (Greece) and from the archaeological site of thermal springs in Montegrotto (Italy) to the historical points of interest in the cities of Luxembourg and Malta. Each pilot contributed data from about 25-30 unique items. The data sample describes museum exhibits, gallery items, archaeological sites and points of interest in terms of their unique identifier, associated descriptions, multimedia elements, and relevant keywords describing their content, use and/or symbolism.

The project ingests a wide range of diverse data associated to cultural heritage objects, events and subjects that span from antiquity to modern times and have a geographic span that runs across Europe. Hence, data is inherited to a wide array of formats, technologies, management and classification approaches relevant to each data provider or resource. The data modelling exercise relied on a rigorous set of Upper-level ontology definitions in order to express a diverse range of cultural heritage data on the same level of semantics and with the same degree of granularity.

Overall, the data modelling exercise delivered 80 uniquely identified items that are composed of 102 Physical Man Made Objects and 17 Physical Man Made Things. This translates to 3440 ontology (OWL) statements of named individual declaration and property assertion.

4.1 Method

The data modelling method addresses issues relating to the diversity of content types, data formats, and level of data detail. The process is abstracted into three main stages: i) selecting and curating the source data for each pilot; ii) data cleansing and normalisation, followed by data mapping to the Upper-level ontology; and iii) automatic data assignment to CC ontology ensuring compliance with the model.

The **Manual Data Extraction** stage was dedicated to impose a data structure across a range of unstructured sample data available in text format. The volume of the data was not such to justify the development of a Natural Language Processing application for the automatic extraction of information from textual snippets. The task identified textual instances of relevant types (i.e. type of exhibit and related material), temporal and spatial information, dimensions, and other features of interest such as inscriptions or visual representations.

The **Semi-Automatic Database Construction** stage aimed at populating a set of relational database tables with structured data, from spreadsheets originating directly from the pilots or from the previous Manual Data Extraction stage. The relational database acted as a mediating layer between the semi-structured data files and the final OWL output feeding the routines of the Automatic Statements Generation stage with structured data. The database introduced a series of tables that stored the different types of CSV data, such as temporal, spatial, dimension, features, and other information associated to the cultural heritage data and conforming to CIDOC CRM structure of the ontology.

The **Final Automatic OWL Generation** stage**,** ingested the structured data of the relational database into the CC Upper-level ontology. The process employed a series of PHP routines driven by SQL queries for retrieving selected database records and declaring them as ontology individuals using OWL class and property assertions. The routines cater for the automatic generation of statements with respect to individual(s) declaration, class assertion, object property assertion, and data property assertion. String cleansing techniques were also applied for the generation of URI friendly values whereas in many cases complex SQL Join statements were used for retrieving record relationships across the database tables.

4.2 Data Modelling Example

The data modelling exercise delivered a representative example of pilot data with respect to the semantics of the Upper-level ontology. It managed to harmonise diverse data under a common semantic layer enriching their structure and enabling inference and retrieval. A leading modelling choice is the adoption of the specialised CIDOC CRM classes; E22.physical Man Made Object and E24.Physical Man Made Thing, which provide a unified semantic view to a range of items of interest across the four pilots. This is augmented by the SKOS Concept and Concept Scheme classes drawing in the concepts incorporated in the CCCS. Hence, the range of artefacts, paintings, museum exhibits, monuments, and points of interest is modelled as instances of the aforementioned specialised classes.

The following example presents the modelling arrangements of museum exhibit 2279 from the Archaeological Museum of Tripolis (Greece). The museum contributes approximately 25 museums exhibits containing rich descriptions and information about their temporal, geometrical, spatial and contextual characteristics as seen below.

The example presents some specific requirements with respect to the modelling of the provenance of exhibits. The provenance information of the exhibit is accommodated by an E5.Event of type 'excavation' that took place in Kynouria (Greece). Figure 1 captures the semantics of the tombstone with respect to dimension, date of production, material, and location. The model accommodates relationships to conceptual characteristics that describe the artefact in terms of its reflective topic and subject keywords, these being the notion of death, funerary rites and funerary art through the ages, etc. It should be noted here that concepts through the structure of the CCCS can be enhanced at the direct terminology level, i.e. “tombstones” are part of “cemeteries” and are linked to “funerary sculpture”.

> ***2279****: Marble pediment tombstone with a representation of a family (enface). The female figure bears a chiton and a cloak. The male figure and the boy bear a short chiton. On the architrave there is the inscription ΑΝΤΙΟΧΙC ΦΟΡΤΟΥΝΑΤΟΥ ΘΥΓΑΤΗΡ ΚΑΛΛΙΣΤΗ. Found in Herod Atticus villa in Loukou, Kynouria. Roman era work (middle Antonine era, 161 A.D - 180 A.D.). Dimensions: Height 1.60m, Width 0.82m. Location: Room 15, 1st floor*

In addition, the inscription of the tombstone is modelled with precise semantics available from the upper-level ontology where the specialized property P128.carries, enables the relationship between the actual artefact and the carried inscription to be fully expressed. It is a different semantic relationship than the P62.depicts that is used for connecting an artefact with a depicted visual item. It is a fine distinction between depiction and carried inscription, demonstrating the flexibility and breadth of the ontology to deal with precise semantics when required.

The notion of women’s dresses is given both as a reflective topic and a concept, as these coincide. The CCCS can lead the user of the app further to the “dress” as a “culture” element and what this expresses for “women’s appearance”.

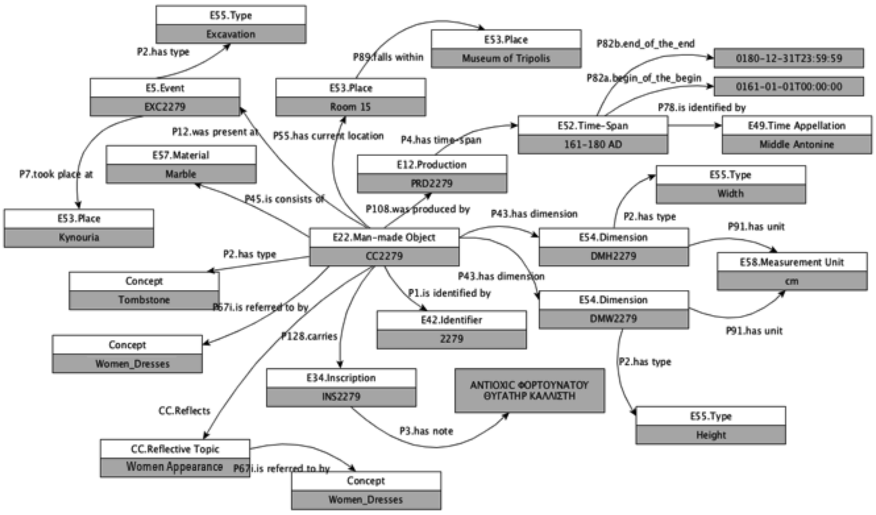

Figure 1: Data model of museum exhibit 2279 (Archaeological Museum of Tripolis, Greece)

5. Discussion

The design and development of the CrossCult Knowledge Base is not a straightforward data modelling exercise, but comes with some interesting research and practical challenges. The first challenge was the selection of the underlying ontology. Despite its growing popularity in the Cultural Heritage domain and its rich expressive capabilities, CIDOC-CRM was not an easy selection. Researchers with an Information Science background preferred solutions based on taxonomies or classification systems (e.g. Dublin Core), while software developers found CIDOC-CRM unnecessary complicated and verbose for the needs of the platform and mobile apps they develop. Considering the importance of modelling the relationships between the different cultural heritage resources used in the project, as well as the need for semantically linking such resources with external vocabularies and ontologies, we finally decided to adopt CIDOC-CRM.

Another critical challenge is related to the population of the ontology with appropriate individuals and statements describing the available cultural heritage resources. We presented the process of converting the available unstructured or semi-structured data into instances of the Upper-level ontology classes and statements using properties of the ontology. However, the mapping between the terms used by historians in the original descriptions of the resources and the elements of the ontology was not in many cases straightforward. Reaching a common understanding of the precise meaning of the original descriptions, and determining their mappings to the ontology required extensive communication between the ontology experts and the historians. By focusing on a representative sample from the four project pilots, we developed semi-automatic processes, which could then be re-used for all the pilot data.

The different backgrounds of the people who were involved in the development of the CrossCult Classification Scheme (information scientists, historians and museum experts) brought two more challenges to the project: how to determine the scope of the vocabulary, and how to come up with a commonly agreed structure. Two decisions that helped us address such challenges were: (i) to rely as much as possible to standard external vocabularies such as AAT; (ii) to setup and use an online environment for collaborative development and management of vocabularies, thesauri and taxonomies. Among others, the environment enables discussions on the terms and structure of the ontology, linking the vocabulary to external terms and creating RDF descriptions of the vocabulary.

6. Conclusion and Future Steps

The paper presented the main design decisions, tasks and challenges associated to the development of the CrossCult Knowledge Base. Apart from serving the specific aims of the project, the research we present in this paper, has three more general contributions to the Digital Heritage domain: (i) it demonstrates the use and deployment of standard cultural heritage ontologies, which have so far been used mainly for research purposes, in the context of user-oriented applications; (ii) it develops a vocabulary for historical reflection and integrates it into standard cultural heritage ontologies; (iii) it harmonizes datasets describing disparate cultural heritage resources, from museum exhibits and archaeological sites, to Points of Interest in urban settings.

We also presented a data modelling example, which demonstrated the semantic description of project pilots' data with respect to the semantics of the Upper-level ontology, which underlies the CrossCult Knowledge Base. The next stages will focus: (i) augmenting the data with media content and narratives that enhance their reflection and re-reinterpretation qualities; (ii) semantically enriching the resource descriptions with links to external standardised semantic web resources; (iii) further refining the scope and structure of Reflective Topics and their relation to keywords, narratives and other reflection proposals; (iv) extending the ontology to accommodate other project-related concepts, such as the pilots' venues and the users of the pilot apps.

Acknowledgements

This work has been funded by CrossCult: "Empowering reuse of digital cultural heritage in context-aware crosscuts of European history", funded by the European Union's Horizon 2020 research and innovation program, We would like to thank Louis Deladiennee (LIST), Kalliopi Kontiza (National Gallery), Yannick Naudet (LIST), Joseph Padfield (National Gallery) and Evgenia Vasilakaki (TEI-A) for their valuable comments and contributions during the course of this research.

References

[1] Adair, B., Filene, B. and Koloski, L. eds., (2011). *Letting Go?: Sharing Historical Authority in a User-Generated World*. Left Coast Press.

[2] Hyvönen, E., (2012). Publishing and using cultural heritage linked data on the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 2(1), pp.1-159..

[3] Ronzino, P., Amico, N. and Niccolucci, F., (2011). Assessment and comparison of metadata schemas for architectural heritage. *Proc. of CIPA, Sep 12*.

[4] Oldman, D., Doerr, M., de Jong, G. and Norton, B., (2014). Realizing lessons of the last 20 years: A manifesto for data provisioning & aggregation services for the digital humanities (a position paper*). D-lib magazine*, 20(7/8).

[5] King, L., Stark, J.F. and Cooke, P., 2016. Experiencing the Digital World: The Cultural Value of Digital Engagement with Heritage. *Heritage & Society*, 9(1), pp.76-101.

[6] Doerr, M. (2003). The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, vol. 24, no. 3, pp. 75–92.

[7] A. Miles and S. Bechhofer, SKOS simple knowledge organization system reference, available at http://www.w3.org/TR/skos-reference/, 2009.

[8] D. Brickley and L. Miller, FOAF Vocabulary Specification 0.99, available at http://xmlns.com/foaf/spec/, 2014